ICIAM 2019 SEMA SIMAI Springer Series 1

Tomás Chacón Rebollo
Rosa Donat
Inmaculada Higueras *Editors*

Recent Advances in Industrial and Applied Mathematics

ICIAM
2019
VALENCIA

OPEN ACCESS

Springer

SEMA SIMAI Springer Series

ICIAM 2019 SEMA SIMAI Springer Series

Volume 1

This sub-series of the SEMA SIMAI Springer Series aims to publish some of the most relevant results presented at the ICIAM 2019 conference held in Valencia in July 2019.

The sub-series is managed by an independent Editorial Board, and will include peer-reviewed content only, including the Invited Speakers volume as well as books resulting from mini-symposia and collateral workshops.

The series is aimed at providing useful reference material to academic and researchers at an international level.

More information about this subseries at https://link.springer.com/bookseries/16499

Tomás Chacón Rebollo · Rosa Donat ·
Inmaculada Higueras
Editors

Recent Advances in Industrial and Applied Mathematics

Springer

Editors
Tomás Chacón Rebollo
Departamento de Ecuaciones Diferenciales
y Análisis Numérico & Instituto de
Matemáticas (IMUS)
Universidad de Sevilla
Facultad de Matemáticas
Sevilla, Spain

Inmaculada Higueras
Departamento de Estadística, Informática y
Matemáticas, Edificio Los Pinos, Campus
Arrosadia
Universidad Pública de Navarra
Pamplona (Navarra), Spain

Rosa Donat
Departament de Matemàtiques
Facultat de Matemàtiques
Universitat de València
Burjassot (Valencia), Spain

ISSN 2199-3041 ISSN 2199-305X (electronic)
SEMA SIMAI Springer Series
ISSN 2662-7183 ISSN 2662-7191 (electronic)
ICIAM 2019 SEMA SIMAI Springer Series
ISBN 978-3-030-86235-0 ISBN 978-3-030-86236-7 (eBook)
https://doi.org/10.1007/978-3-030-86236-7

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

Foreword

During the second week of July, the ICIAM 2019 Congress took place in Valencia with almost 4,000 participants, with 50 plenary talks, more than 300 mini-symposia, 550 contributed talks and 250 posters. A wide representation of world applied mathematics met in Valencia to present and discuss how mathematics was applied to the most diverse disciplines, such as applied mathematics for industry and engineering, biology, medicine and other natural sciences, control and systems theory, dynamical systems and nonlinear analysis, finance and management science, industrial mathematics, mathematics and computer science, numerical analysis, partial differential equations and simulation and modeling, to name some of them.

Within the organizing committee, the idea arose that these presentations and discussions should be reflected in some way for the future. And the offer from Springer came up to launch a series of volumes that would record the most notable advances that took place in it.

This offer crystallized in the *ICIAM 2019 SEMA SIMAI Springer Series*, which includes the present volume, dedicated to the conferences of the invited speakers, which occupies a very central and special place, since it is offered in open access mode, thanks to the support of Sociedad Española de Matemática Aplicada (SeMA).

The selection of the 336 mini symposia of the ICIAM 2019 was made by its academic committee. In a very direct relationship with it, the editorial committee of this series was formed by F. Arándiga Llaudes, M. Gómez Mármol, F. Guillén-González, F. Ortegón Gallego, C. Parés, P. Quintela, C. Vázquez-Cendón, S. Xambó-Descamps and myself. The members of this committee were in charge of selecting the proposals, many of them derived from mini-congress symposia, and also to act as the editors in charge for some of the 14 volumes that make up this series:

1. *Recent Advances in Industrial and Applied Mathematics*, edited by Tomás Chacón Rebollo, Rosa Donat and Inmaculada Higueras.
2. *Stabilization of Distributed Parameter Systems: Design Methods and Applications*, edited by Grigory Sklyar and Alexander Zuyev.
3. *Cartesian CFD Methods for Complex Applications*, edited by Ralf Deiterding, Margarete Oliveira and Kai Schneider.

4. *Applications of Wavelet Multiresolution Analysis*, edited by Juan Pablo Muszkats, Silvia Alejandra Seminara and María Inés Troparevsky.
5. *Progress in Industrial Mathematics: Success Stories*, edited by Manuel Cruz, Carlos Parés and Peregrina Quintela.
6. *Applied Mathematics for Environmental Problems*, edited by María Isabel Asensio, Albert Oliver and José Sarrate.
7. *Improving Applied Mathematics Education*, edited by Ron Buckmire and Jessica M. Libertini.
8. *Fractals in Engineering: Theoretical Aspects and Numerical Approximations*, edited by Maria Rosaria Lancia and Anna Rozanova-Pierrat.
9. *Recent Advances in Differential Equations and Control Theory*, edited by Concepción Muriel and Carmen Pérez-Martínez.
10. *Emerging Problems in the Homogenization of Partial Differential Equations*, edited by Patrizia Donato and Manuel Luna-Laynez.
11. *Multidisciplinary Mathematical Modeling*, edited by Francesc Font and Tim Myers.
12. *Mathematical Descriptions of Traffic Flow: Micro, Macro and Kinetic Models*, edited by Gabriella Puppo and Andrea Tosin.
13. *Systems, Patterns and Data Engineering with Geometric Calculi*, edited by Sebastià Xambó-Descamps.
14. *Modeling, Simulation and Optimization in the Health and Energy Sector*, edited by Rene Pinnau, Nicolas R. Gauger and Axel Klar.

As can be easily seen, the application of mathematics spreads through the most diverse areas, such as industry, health and energy, engineering data science, environmental problems, geometric calculi, numerical approximation, traffic flow, education, etc.

Now is the time for the reader to delve into the volumes of this series and learn, reflect, incorporate new ideas and generally enjoy their content, hoping that the volumes of this series can serve as a reference for even more innovative applications of mathematics in the future.

Finally, it is time of acknowledgements. Starting with the ICIAM 2019 Congress, especially its executive committee led by Tomás Chacón and Rosa Donat as living forces of the event, as well as the scientific committee led by Alfio Quarterioni and the multiple organizers of mini-symposia, speakers and attendees. Continuing with Francesca Bonadei as the promotor within Springer of the need for the existence of this series, and with the members of the editorial board of this series, and ending with the editors in charge and authors of each volume, which with its excellent work, are the real creators of the message of this series.

Barcelona, Spain Amadeu Delshams

Preface

The papers appearing in this volume are authored by some of the invited speakers of the **9th International Congress of Industrial and Applied Mathematics**, held in València from July 15 to 19, 2019. This volume is part of a series dedicated to ICIAM 2019-Valencia.

The congress, hosted by the Spanish Society for Applied Mathematics (SeMA), was organized at the Universitat de València (Spain), on behalf of the International Council for Industrial and Applied Mathematics (ICIAM). With 3983 participants from 99 different countries, more than 3400 lectures delivered and nearly 250 poster presentations, ICIAM 2019 has been a great success. These data represent a net increase in participation, with respect to an already rising trend in previous editions of this series of events, which can be considered a sound proof of the growing interest of the applied and industrial mathematics community in ICIAM congresses.

The industrial aspect of the congress was further enriched by organizing a specific mathematical technology transfer oriented activity: '*The Industry Day*'. Fourteen speakers, selected from a broad representation of different sectors, presented the results of ongoing collaborations with academy and the benefits derived from it, such as better products and services, optimization of processes, organization and accounting, and growth and innovation. In addition, 19 industrial mini-symposia were scheduled during the congress, and 48 'industry-related' posters were on display during '*The Industry Day.*'

Thirty-five satellite events took place during 2018 and 2019 covering a broad range of topics within industrial and applied mathematics. These events included two CIMPA schools (Kenitra, Morocco and Tunis, Tunisia, 2019), devoted to initiate young students from developing countries into research. Also, several Spanish towns/regions were appointed sub-venues of ICIAM-2019-Valencia (Bilbao, Galicia, Málaga, Seville and Zaragoza) and, as such, organized 12 satellite events. We are deeply thankful to the organizers of all satellite events.

The preparation of the candidacy in 2012 started the long process involved in the planning of this complex event. Our deepest gratitude and heartiest thanks go to all the people who helped with their abilities to create ICIAM 2019-Valencia. A list of all the committees and people involved in this task is given in this book.

The congress could not have been possible without the support of a large set of sponsors. A special mention is due to our main sponsors: Banco Santander, who financed over 70% of the Grant Program of the congress, and the Universitat de València, for its generous offer to make available their facilities to hold the conference. Thanks are also due to the Spanish universities that contributed to fund over 20% of the Grant Program and to the individual donors who contributed to the remaining 5%.

A thankful recognition is also due to our four institutional sponsors: Ministry of Science, Innovation and Universities, Generalitat Valenciana, Diputació de València and Ajuntament de València.

On behalf of ICIAM 2019, we would like to express our most sincere gratitude to the invited speakers that have contributed to this volume for taking the time to provide their valuable contributions, helping us to make this the reference publication of the congress.

Sevilla, Spain	Tomás Chacón Rebollo
Valencia, Spain	Rosa Donat
Pamplona, Spain	Inmaculada Higueras

ICIAM Congresses

1987	Paris
1991	Washington, D.C.
1995	Hamburg
1999	Edinburgh
2003	Sydney
2007	Zurich
2011	Vancouver
2015	Beijing

2019-Valencia

Opening ceremony

ICIAM
2019
VALENCIA
9th International Congress on
Industrial and Applied Mathematics
July 15-19
Valencia · Spain

Traditional valencian dances and Muixeranga (human towers)

The Onsager Conjecture. Deterministic version of Komogorov law
Any weak solution which belongs to the space $C^{0,\alpha}$ with $\alpha > \frac{1}{3}$ conserves the energy.
Complete mathematical proofs were given in (1994) by Eyink then Constantin , E, Titi. in 1994 and several other extensions A Cheskidov, Constantin, Friedlander, and Shvydkoy. (2008).
national gress on trial and Applied ematics
ly 15-19 ia · Spain
ICIA 20 VALEN
SēMA
ICIAM

Plenary talk

Closing ceremony and transfer of ICIAM flag

ICIAM Prize Winners

ICIAM Collatz Prize

1999	Stefan Müller, Max Planck Institut für Mathematik Leipzig, Germany.
2003	Weinan E, Princeton University, USA.
2007	Felix Otto, Universität Bonn, Germany.
2011	Emmanuel J. Candès, Stanford University and CALTECH, USA
2015	Annalisa Buffa, CNR-IMATI, Italy.
2019	Siddharta Mishra, ETH Zürich, Switzerland.

ICIAM Lagrange Prize

1999	Jacques-L. Lions, Collège de France and Académie des Sciences de Paris, France.
2003	Enrico Magenes, Università di Pavia, Italy.
2007	Joseph Keller, Stanford University, USA.
2011	Alexandre J. Chorin, U.C. Berkeley and LBNL, USA.
2015	Andrew J. Majda, New York University, USA.
2019	George Papanicolaou, Stanford University, USA.

ICIAM Maxwell Prize

1999	Grigory I. Barenblatt, UC Berkeley, USA, and University of Cambridge, UK.
2003	Martin D. Kruskal, Rutgers University, USA.
2007	Peter Deuflhard, Zuse Institute Berlin, Germany.
2011	Vladimir Rokhlin, Yale University, USA.

2015 Jean-Michel Coron, Université Pierre et Marie Curie, France.
2019 Claude Bardos, Université Paris Diderot (Paris VII), France.

ICIAM Pioneer Prize

1999 Ronald R. Coifman, Yale University, USA,
Helmut Neunzert, University of Kaiserslautern, Germany.
2003 Stanley Osher, University of California, Los Angeles, USA.
2007 Ingrid Daubechies, Princeton University, USA,
Heinz Engl, Johannes Kepler University, and Austrian Academy of Sciences, Austria.
2011 James Albert Sethian, UC Berkeley and LBNL, USA.
2015 Björn Engquist, The University of Texas at Austin, USA.
2019 Yvon Maday, Sorbonne University, Paris, France.

ICIAM Su Buchin Prize

2007 Gilbert Strang, Massachusetts Institute of Technology, USA.
2011 Edward Lungu, University of Botswana, Botswana.
2015 Li Tatsien, Fudan University, P.R. China.
2019 Giulia Di Nunno, University of Oslo, Norway.

2019 ICIAM Prize Ceremony. From left to right: Joan Ribó (Major of Valencia), Ximo Puig (President of the Generalitat Valenciana), Y. Maday, G. Di Nunno, His Majesty Felipe VI, G. Papanicolaou, C. Bardos, S. Mishra, M. Esteban and Pedro Duque (Minister for Science, Innovation and Universities)

Organization of ICIAM 2019-Valencia

Congress Director: Tomás Chacón, University of Seville, Spain

Honorary Committee

President: His Majesty King Felipe VI of Spain

Members

Mr. Pedro Duque, Minister for Science, Innovation and Universities, Spain
Mr. Ximo Puig, President of Generalitat Valenciana, Spain
Mr. Vicent Marzà, Conseller d'Educació, Investigació, Cultura i Sport of Generalitat Valenciana, Spain
Prof. Josefina Bueno, Directora General d'Universitats of Generalitat Valenciana, Spain
Prof. Julio Abalde, Rector, University of A Coruña, Spain
Prof. José Ángel Narváez, Rector, University of Malaga, Spain
Prof. Nekane Balluerca, Rector, University of the Basque Country, Spain
Prof. José Mora, Rector, Universitat Politècnica de València, Spain
Prof. Antonio López, Rector, University of Santiago de Compostela, Spain
Prof. Miguel Ángel Castro, Rector, University of Seville, Spain
Prof. M. Vicenta Mestre, Rector, Universitat de València, Spain
Prof. Manuel Joaquín, Reigosa, Rector, Universidade de Vigo, Spain
Prof. José Antonio Mayoral, Rector, University of Zaragoza, Spain
Mrs. Ana Botín, President of Banco Santander

Scientific Program Committee

Chair

Alfio Quarteroni, EPFL, Lausanne, Switzerland, and Politecnico di Milano, Italy

Members

Tony F. Chan, Hong Kong, China
Manuel Doblaré Castellano, Seville, Spain
Qiang Du, New York, USA
Enrique Fernández Cara, Seville, Spain
Irene Fonseca, Pittsburgh, USA
Irene Gamba, Austin, USA
Markus Hegland, Canberra, Australia
Ilse Ipsen, Raleigh, USA
Ravi Kannan, Bangalore, India
Claudia Kluppelberg, Munich, Germany
Karl Kunisch, Graz, Austria
Yasumasa Nishiura, Sendai, Japan
Benoit Perthame, Paris, France
Daya Reddy, Rondebosch, South Africa
Claudia Sagastizabal, Rio de Janeiro, Brazil
Jeffrey Saltzman, Waltham, USA
Wil Schilders, Eindhoven, Netherlands
Endre Suli, Oxford, UK
Eric Vanden Eijnden, New York, USA
Pingwen Zhang, Beijing, China

Executive Committee

Chair

Tomás Chacón (US)

Co-Chairs

Rosa Donat (UV)
Luis Vega (UPV/EHU)

Members

María Paz Calvo (UVA)
Eduardo Casas (UC)
Amadeu Delshams (UPC)
Henar Herrero (UCLM)

Inmaculada Higueras (UPNA)
Juan Ignacio Montijano (UNIZAR)
Peregrina Quintela (USDC)
Carlos Vázquez-Cendón (UDC)
Elena Vázquez-Cendón (USC)

Thematic Committees

Academic

Chair: Amadeu Delshams (UPC)

Lino Álvarez-Vázquez (UVIGO)
Rafael Bru (UV)
Fernando Casas (UJI)
Eduardo Casas (UC)
Enrique Fernández-Nieto (US)
Javier de Frutos (UVA)
Dolores Gómez-Pedreira (USC)
Jesús López-Fidalgo (UNAV)
Pep Mulet (UV)
Francisco Ortegón-Gallego (UCA)
Francisco Padial (UPM)
Carlos Vázquez-Cendón (UDC)

Finance

Chair: Eduardo Casas (UC)

Antonio Baeza (UV)
Luis Alberto Fernández (UC)
Julio Moro (UC3M)
Carlos Vázquez-Cendón (UDC)

Fundraising

Carlos Vázquez-Cendón (UDC)
Jesús Sanz-Serna (UC3M)

Industrial Advisory

Chair: Peregrina Quintela (USC)

Emilio Carrizosa (US)
David Pardo (UPV/EHU)
Antonio Huerta (UPC)
Carlos Parés (UMA)
Wenceslao González-Manteiga (USC)

Communication and Outreach

Chair: Henar Herrero (UCLM)

Sergio Blanes (UPV)
Fernando Casas (UJI)
Bartomeu Coll (UIB)
Inmaculada Higueras (UPNA)
Juan Ignacio Montijano (UNIZAR)
Alfred Peris (UPV)
Francisco Ortegón-Gallego (UCA)
Francisco Pla (UCLM)
Joan Solá-Morales (UPC)
Sebastià Xambó-Descamps (UPC)

Publications and Promotions

Chair: Inmaculada Higueras (UPNA)

Rafael Bru (UPV)
María Paz Calvo (UVA)
Domingo Hernández-Abreu (ULL)
Henar Herrero (UCLM)
Mariano Mateos (UNIOVI)
Julio Moro (UC3M)

Satellite and Embedded Meetings

Chair: María Paz Calvo (UVA)

Francisco Guillén-González (US)
Carlos Parés (UMA)
Luis Rández (UNIZAR)
Carlos Vázquez-Cendón (UDC)
Luis Vega (UPV/EHU)

Travel Support Committee

Chair: Elena Vázquez-Cendón (USC)

Macarena Gómez-Mármol (US)
José Manuel González-Vida (UMA)
Pep Mulet (UV)
Francisco Javier Sayas (U. of Delaware)
Rodrigo Trujillo-González (ULL)

Local Arrangements

Chair: Rosa Donat (UV)

José María Amigó (UMH)
Francesc Aràndiga (UV)

Ana María Arnal (UJI)
Antonio Baeza (UV)
Sergio Blanes (UPV)
Rafael Bru (UPV)
Fernando Casas (UJI)
Cristina Chiralt (UJI)
Rafael Cantó (UPV)
José Alberto Conejero (UPV)
Isabel Cordero-Carrión (UV)
Cristina Corral (UPV)
Juan Carlos Cortés (UPV)
María Teresa Gassó (UPV)
Olga Gil-Medrano (UV)
Alicia Herrero (UPV)
Leila Lebtahi (UV)
María del Carmen Martí (UV)
Vicente Martínez (UJI)
José Mas (UPV)
José Salvador Moll (UV)
Francisco Gabriel Morillas-Jurado (UV)
Pep Mulet (UV)
Mari Carmen Perea (UMH)
Rosa Peris (UV)
Alfred Peris (UPV)
Sergio Segura de León (UV)
Ana María Urbano (UPV)
Pura Vindel (UJI)

Acronyms of Spanish Universities

UC3M	Carlos III University, Madrid
UDC	University of A Coruña
UA	University of Alicante
UCA	University of Cadiz
UC	University of Cantabria
UCLM	University of Castilla La Mancha
ULL	University of La Laguna
UMA	University of Malaga
UNAV	University of Navarra
UNIOVI	University of Oviedo
USC	University of Santiago de Compostela
US	University of Seville
UVA	University of Valladolid

UVIGO	University of Vigo
UNIZAR	University of Zaragoza
UPV/EHU	University of the Basque Country
UMH	Miguel Hernández University
UPM	Technical University of Madrid
UPNA	Public University of Navarra
UIB	University of the Balearic Islands
UV	Universitat de València
UJI	Universitat Jaume I
UPC	Universitat Politècnica de Catalunya
UPV	Universitat Politècnica de València

Collaborators at the Universitat de València

M. Vicenta Mestre, Rector

Rector’s Cabinet

Justo Herrera, Vice-rector
Juan Vte. Climent, Manager
Beatriz Gómez, Vice-manager
José Ramírez, Vice-manager
Joan Enric Úbeda, Director
Carmen Fayos, Head of Staff

Facultat de Psicologia

M. Dolores Sancerni, Dean
Juan M. Rausell, Administrator
Juan J. Cancio, Coordinator
Concierges of the building

Facultat de Filosofía i Ciències de l´ Educació

Rosa M. Bo, Dean
Francisco J. Moreno, Administrator
Esther Bolinches, Coordinator
Concierges of the Building

Health, Safety and the Environment Service

M. José Vidal, Head of Staff
Miguel A. Toledo, Technician
Verónica Saiz, Technician
Vicente Caballer, Technician

Computer Service

Fuensanta Doménech, Head of Staff

Faustino Fernández, IT Infrastructure
Magdalena Ros, Quality Control

Blasco Ibáñez Campus Management Unit

Carmen Tejedo, Administrator
Dolores Cano, Head of Staff
M. Ángeles Llorens, Head of Staff
Inmaculada Yuste, Administrative
M. José Ballester, Services Coordinator
Maria Luisa Jordán, Concierge
Concierges of Aularios I, III y VI

Facultat de Medicina i Odontologia

Francisco J. Chorro, Dean
M. Vicenta Alandi, Administrator.
Guillermo Pérez, Coordinator
Concierges of the Building

Facultat de Filologia, Traducció i Comunicació

Amparo Ricós, Dean
Francisca Sánchez, Administrator
Josep M. Valldecabres, Coordinator
Concierges of the Building

Facultat de Geografia i Història

Josep Montesinos, Dean
Joaquín V. Lacasta, Administrator
Josep Vicó, Coordinator
Concierges of the Building

UVSports Service

Vicent Añó, Director
M. Paz Molina, Administrator
Francisco Vicent, Coordinator
Francisco Barceló, Concierge

Technical and Maintenance Service

Rosa M. Mochales, Head
Rafael Antón, Technician
M. Dolores Yagüe, Technician
Jorge Vila, Technician
Ramón Doménech, Maintenance
Modesto Ramírez, Maintenance

Carles Aguado, Maintenance
Diego Cantero, Maintenance

UVdisability Service

M. Celeste Asensi, Director
Restituto Vaño, Accessibility

Technical Unit

Luis Juaristi, Head of Staff
Vicente Tarazona, Technician
José M. Zapata, Technician

Collaborators at the University of Seville

Teresa Ayuga

Staff from External Partners

Jesús Ibáñez, Security director UV
Luis Briz and Security Staff from *Clece Security*
Concierges from *UTE Blasco Ibáñez*
Staff from *Grupo Fissa, Cleaning Company.* Amparo Cuadrado, Coordinator
Antonio Gonzalbo and Maintenance Staff from *Ferrovial*
Alexandre Andrés and Carlos J. Soler from Valnu
Ana Mª Gómez and Gardening Staff from *Special Employment Center IVASS*

Opening Ceremony

Tomás Chacón Rebollo, Congress Director

Your Majesty, President of the Region of Valencia, Minister of Science, Innovation and Universities, Major of Valencia, President of ICIAM, respected guests and delegates, on behalf of the Spanish Society for Applied Mathematics and the organizing committee, it is for me a pleasure to convey you our warmest welcome to ICIAM-2019-Valencia Congress.

Mathematics is silently shaping the present technological world. It provides a deep insight in numberless processes and systems, thereby advancing scientific knowledge. It also generates added value in virtually all economic sectors. On top of that, the last years have witnessed a change in paradigm, as mathematics directly provide the technological basis of emerging sectors related with data analysis.

The research and transfer in mathematics have experienced a fast development in Spain; besides all sciences, since the last decades of the twentieth century, Spain occupies today the 7th world position in mathematical research by citations. The mathematics play a relevant role in the Spanish economy; in fact, 10% of the national gross income and 6% of the employment are directly due to its use in the economic activity.

ICIAM 2019 Congress features 27 invited talks, the 5 ICIAM prices, the Olga Taussky-Todd Lecture and the Public Lecture. It will count on nearly 2000 talks as well as 250 posters. It also includes three special panels of great interest to understand the social framework in which our job as mathematicians takes place. This is industry talking about mathematics, instead of mathematicians talking about their collaborations with industry. ICIAM 2019 also includes an Industry Day, where 14 technological companies have agreed to present how mathematics helps to improve their production processes.

Thanks to four different funding programs, we have been able to offer over 230 scholarships to young researchers as well as to researchers coming from developing countries. In addition, we have implemented a volunteers program with over 170 young students that will greatly help the organization of the congress.

All this has been possible thanks to the collaborative work of the scientific panel committee, chaired by Prof. Alfio Quarteroni, and an enthusiastic organizing committee. I convey my deepest thanks to all of them. Special thanks are addressed to the Spanish Society for Applied Mathematics, and its president, Prof. Rosa Donat, who also chairs the local organizing committee. Let me also acknowledge the role of our families, for their support all along the organization of the congress.

We are indebted to ICIAM for trusting us to organize this congress and especially to her past and present presidents, Profs. Barbara Keyfitz and Maria Esteban, for their help and advice in the organization process. We also address our deepest thanks to the many organizations that have sponsored the congress: the Spanish Government, the Region of Valencia, the Diputació de València, the City Council and the University of Valencia, Spanish centers, departments and institutes of mathematics, Springer Publishing House, Santander Bank and the many individual donors. We are also indebted to SIAM for embedding their annual meeting in this ICIAM Congress and also to all you for organizing and participating in the many activities that take place within it.

You find yourself at the perfect time and place to learn about new mathematical tools, exchange ideas and move ahead in the thrilling challenge of shaping the world with mathematics.

Welcome to ICIAM 2019-Valencia Congress!!

Maria J. Esteban, President of ICIAM

ICIAM
2019
VALENCIA
9th International Congress on
Industrial and Applied Mathematics
July 15-19
Valencia · Spain

His Majesty the King, President of the Generalitat of Valencia, Major of Valencia, Minister of Science, Innovation and Universities, Congress Director, ladies and gentlemen, dear colleagues,

It is my great honor and pleasure, to welcome you all to ICIAM 2019, the ninth International Congress on Industrial and Applied Mathematics.

The ICIAM congresses are the main event organized by our international organization, a network of more than 50 learned societies. The global ICIAM community covers many countries and all topics that are related to the applications of mathematics to the real world, to industry, to health, to economy, to climate, to artificial intelligence and so on. Mathematics is unavoidable in the development of new technologies and in the advancement of our societies. As the recent report on the impact of mathematics on the Spanish economy shows, investing in mathematics is a very good idea, because the economic returns are high. This was also apparent in similar impact studies carried out previously in the UK, the Netherlands and France.

This congress is the occasion when worldwide applied and industrial mathematicians show to each other what they have done in the past years and what they plan to do next. During these days, we will prepare the future.

Spain was chosen six years ago to organize this big congress, the main event in our community, taking place only every four years. In 2015, we were in Beijing, and in 2023, we will be in Tokyo. Here today in the beautiful city of Valencia, we host more than 4000 mathematicians from all over the world, junior, senior, students, professors, researchers and engineers. During these six years, our Spanish colleagues have worked nonstop to make this congress a big success. In the name of the whole ICIAM community, let me thank the organizers for their huge effort. Thank you very much to the Spanish Society of Applied Mathematics (SEMA) and to the whole Spanish applied mathematics community. Thanks also to all official Spanish institutions that have offered their support.

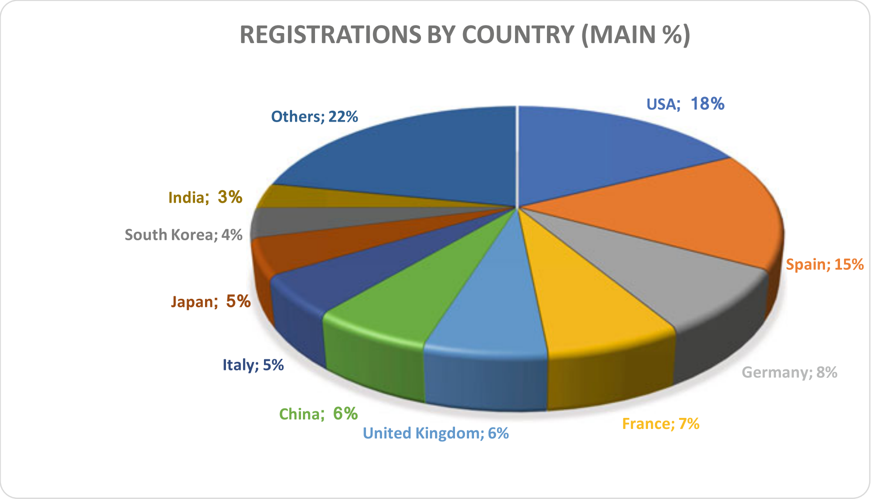
And now, to all of you who are eager to see how the congress will develop, I wish you a productive week. Just be patient and courageous, because the program of the congress is very heavy, but this is the only way to show the whole span of our community's work in only five days. I thank you all for being here, and I wish you a great congress and a very pleasant week!

ICIAM 2019 in Numbers

Scope

- 37 Distinguished Lectures (27 Invited, 5 ICIAM Prize, 3 SIAM Prize, Olga Taussky-Todd, Public Lecture)
- Industry Day (14 invited, industry-driven lectures)
- 2148 scientific contributions from participants (336 mini-symposia: 1344 talks, 555 contributed talks, 249 posters displayed). 90 simultaneous sessions.
- 234 scholarships awarded, 176 volunteers.
- 35 satellite meetings
- Outreach activities: Math-Work for children. Arquímedes: A Planetarium Opera.
- Public lecture by Víctor M. Pérez-García (UCLM) (Can mathematics help in the war against cancer?)

3983 Registered Delegates (Geographical Distribution)



Percentage of participants per country

Number of Talks and Posters by Topic

Number of mini-symposia talks, contributed talks and posters by topic

Topics	Mini-symposia talks	Contributed	Talks posters
1. Applied Mathematics for Industry and Engineering	148	64	35
2. Astronomy, Astrophysics and Geophysics	4	8	2
3. Biology, Medicine and other natural sciences	68	41	15
4. Chemistry, Chemical Engineering	0	2	2
5. Computational Geometry	4	3	2
6. Computer Science	8	3	3
7. Control and Systems Theory	40	29	5
8. Discrete Mathematics	4	3	3
9. Dynamical Systems and Nonlinear Analysis	72	30	15

(continued)

(continued)

Topics	Mini-symposia talks	Contributed	Talks posters
10. Education	16	4	3
11. Finance and Management Science	32	7	3
12. Fluids Physics and Statistical Mechanics	32	19	6
13. Information, Communication, Signals	12	2	5
14. Linear Algebra and Geometry	16	19	1
15. Materials Science and Solid Mechanics	28	11	2
16. Mathematics and Computer Science	92	15	13
17. Numerical Analysis	316	102	39
18. Optimization and Operations Research	28	23	14
19. Ordinary Differential Equations	0	9	4
20. Partial Differential Equations	144	57	22
21. Probability and Statistics	20	12	2
22. Real and Complex Analysis	0	2	2
23. Simulation and Modeling	144	46	27
24. Social Science	0	1	0
25. Other Mathematical Topics and their Applications	40	16	11
26. General	0	1	0
MIA.1. Industrial mathematics success stories	40	11	3
MIA.2. Industrial mathematics case studies	12	8	7
MIA.3. Industrial mathematics education	4	2	3
MIA.4. Industrial mathematics infrastructures to promote industry—academia collaborations	20	5	0
Total	**1344**	**555**	**249**

Satellite Meetings

Bilbao

- 5thWorkshop on Quantitative Biomedicine for Health and Disease (QBIO2019), February 13–14, 2019.
- 150 European Study Group with Industry (150 ESGi), October 21–25, 2019,
- FCPNLO 2018—6th Workshop on Fractional Calculus, Probability and Nonlocal Operators: Applications and Recent Development, September 26–28, 2018.
- BiDAS 3—Third Bilbao Data Science Workshop, November 8–9, 2018.

Galicia

- 147 European Study Group With Industry (147 ESGI), April 8–12, 2019, Santiago de Compostela.
- ICCF 2019—International Conference on Computational Finance, July 8–12, 2019, A Coruña.
- I Conference on Transfer between Mathematics and Industry, July 22–24, 2019, Santiago de Compostela.

Sevilla

- Workshop on PDEs for Biology Systems, April 8–10, 2019.
- ECMI Postgraduate Modeling Week/VI Iberian Modeling Week, July 8–13, 2019.

Málaga

- NumHyp 2019—Numerical Approximation of Hyperbolic Systems with Source Terms and Applications, June 17–21, 2019.

Zaragoza

- Tecnologías en la Divulgación Matemática, DI-MA, May 10–11, 2018.
- Fifteenth International Conference Zaragoza-Pau on Mathematics and its Applications, September 10–12, 2018.
- International Conference on Adaptive Modeling and Simulation 2019 (ADMOS 2019), May 27–29, 2019, Alicante, Spain.
- CSAI—ECCOMAS Thematic Conference Computational Sciences and AI in Industry: new digital technologies for solving future societal and economical challenges, June 12–14, 2019, Jyväskylä, Finland.
- Patterns in Life and Social Sciences, June 13–20, 2019, Granada, Spain.
- 5th ALAMA Workshop: Numerical Linear Algebra (NLA 2019), June 17–18, 2019, Valencia, Spain.
- 17th School on Interactions between Dynamical Systems and Partial Differential Equations (JISD2019), June 17–21, 2019, Barcelona, Spain.
- MEGA 2019—Effective Methods in Algebraic Geometry, June 17–21, 2019, Madrid, Spain.
- ICASQF 2019—Third International Congress on Actuarial Science and Quantitative Finance, June 19–22, 2019, Manizales, Colombia.
- CIMPA Research School 'Data Science for Engineering and Technology,' June 25–July, 2019, Tunis, Tunisia.
- 30 Years of SIMAI: status and perspectives of applied and industrial mathematics in Italy and in Europe (SIMAI30th), July 1–2, 2019, Milan, Italy.
- École de Recherche CIMPA 'Modélisation, analyze mathématique et calcul scientifique dans la gestion des déchets ménagers,' July 3–13, 2019, Kenitra, Marocco.
- AIP 2019—Applied Inverse Problems, July 8–12, 2019, Grenoble, France.

- Tutorial Workshop within the program ‘Geometry, compatibility and structure preservation in computational differential equations,’ July 8, 2019, Cambridge, UK.
- SIAM Conference on Applied Algebraic Geometry 2019, July 9–13, 2019, Bern, Switzerland.
- Mathematical Modeling in Engineering and Human Behavior 2019, July 10–12, 2019, Valencia, Spain.
- CCMA2019 Madrid—International Conference Challenges in Mathematical Architecture: theory, modeling and applications, July 11–13, 2019, Madrid, Spain.
- SciCADE 2019—International Conference on Scientific Computation and Differential Equations, July 22–26, 2019, Innsbruck, Austria.
- Summer School and Workshop Wisla 19, August 19–29, 2019, Wisla, Poland.
- 8th International Conference on Matrix Analysis and its Applications (MAT TRIAD 2019), September 8–13, 2019, Liblice, Czech Republic.
- 2019 Rouen Probability Meeting: Stochastic geometry, Analysis of algorithms, Particle systems, September 23–27, 2019, Rouen, France.
- EHF2018—XVIII Jacques-Louis Lions Spanish-French School on Numerical Simulation in Physics and Engineering, June 25–29, 2018, Las Palmas de Gran Canaria, Spain.
- AGACSE 2018—7th Conference on Applied Geometric Algebras in Computer Science and Engineering, July 23–27, 2018, Campinas, Brazil.
- Second Spain–Brazil Joint Meeting of Mathematical Societies, December 11–14, 2018, Cádiz, Spain.

Other Satellite Meetings

- European Workshop on High Order Numerical Methods for Evolutionary PDEs: Theory and Applications (HONOM 2019), April 1–5, 2019, Madrid, Spain.

Contents

Invited Lectures

Editors and Contributors

About the Editors

Tomás Chacón Rebollo is a full professor at the Department of Differential Equations and Numerical Analysis and the Institute of Mathematics of the University of Seville (IMUS). He is Ph.D. in Mathematics and in Numerical Analysis at the universities of Seville and Paris 6, respectively. His scientific interests are numerical and reduced-order modeling in fluid mechanics and their applications to real-world problems. He is interested in the promotion of mathematical research and transfer. He was the director of BCAM (2012–2013) and IMUS (2015–2019) and chairman of ICIAM 2019 Congress.

Rosa Donat is a professor at the Department of Mathematics of the University of Valencia. She has worked on numerical methods for hyperbolic conservation laws and systems and on multiresolution and subdivision frameworks that incorporate nonlinear approximation techniques. She was actively involved in the Spanish Society for Applied Mathematics (SeMA), as a president in the 2016–1020 period.

Inmaculada Higueras is a full professor at the Department of Statistics, Computer Science and Mathematics of the Public University of Navarre (Pamplona, Spain). Her research focuses on time stepping methods for differential problems (ODEs, DAEs and PDEs). She has worked on numerical stability, numerical preservation of qualitative properties (positivity, monotonicity, contractivity, etc.) and on the design and implementation of robust and efficient numerical schemes.

Contributors

Marsha Berger Courant Institute, New York University, New York, NY, USA

Alfredo Bermúdez Departamento de Matemática Aplicada, Instituto de Matemáticas, Universidade de Santiago de Compostela, Santiago de Compostela, Spain;
Instituto Tecnológico de Matemática Industrial (ITMATI), Santiago de Compostela, Spain

Zhenning Cai Department of Mathematics, National University of Singapore, Singapore, Singapore

Huangxin Chen School of Mathematical Sciences and Fujian Provincial Key Laboratory on Mathematical Modeling and High Performance Scientific Computing, Xiamen University, Fujian, China

Albert Cohen Laboratoire Jacques-Louis Lions, Sorbonne Université, Paris, France

Carlos Conca Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Centro de Modelamiento Matemático UMR 2071 CNRS-UChile, Centro de Biotecnología y Bioingeniería, Universidad de Chile, Santiago, Chile

Wolfgang Dahmen Mathematics Department, University of South Carolina, Columbia, SC, USA

Ron DeVore Department of Mathematics, Texas A & M University, College Station, TX, USA

Leah Edelstein-Keshet University of British Columbia, Vancouver, BC, Canada

Yuwei Fan Department of Mathematics, Stanford University, Stanford, CA, USA

Maria Garzon Department of Applied Mathematics, University of Oviedo, Oviedo, Spain

Naohiro Horio Department of Cardiovascular Surgery, Okayama University Hospital, Okayama, Japan

Viet Q. H. Huynh Advanced Institute for Materials Research, Tohoku University, Aobaku, Sendai, Japan

Kristin Lauter Cryptography and Privacy Research, Microsoft Research, Redmond, USA

Claude Le Bris Ecole des Ponts and Inria, Paris, France

Haitao Leng School of Mathematical Sciences, South China Normal University, Guangzhou, Guangdong, China

Ruo Li CAPT, LMAM and School of Mathematical Sciences, Peking University, Beijing, People's Republic of China

Koki Otera Graduate School of Environmental and Life Sciences, Okayama University, Okayama, Japan

Kazue Sako Waseda University, Tokyo, Japan

Robert I. Saye Mathematics Group, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

James A. Sethian Department of Mathematics, University of California, Berkeley, California, USA

Hiroshi Suito Advanced Institute for Materials Research, Tohoku University, Aobaku, Sendai, Japan

Kenji Takizawa Faculty of Science and Engineering, Waseda University, Shinjuku City, Japan

Takuya Ueda Department of Diagnostic Radiology, Tohoku University Hospital, Sendai, Japan

Dong Wang School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong, China

Xiao-Ping Wang Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China

J. A. C. Weideman Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa

Invited Lectures

Asteroid-Generated Tsunamis: A Review

Marsha Berger

Abstract We study ocean waves caused by an asteroid airburst located over the ocean. The concern is that the waves would damage distant coastal cities. Simple qualitative analysis suggests that the wave energy is proportional to the ocean depth and the strength and speed of the blast. Computational simulations using GeoClaw and the shallow water equations show that explosions from realistic asteroids do not endanger distant cities. We explore the validity of the shallow water, Boussinesq, and linearized Euler equations to model these water waves.

1 Introduction

This talk will review some of the basics behind the simulation of asteroid-generated tsunamis, and how this piece of the Asteroid Threat Assessment Program (ATAP) got its start.

In 1994, the United States Congress asked NASA to identify 90% of asteroids larger than 1 km in diameter that could pose a threat to Earth. This led to the Near Earth Observing (NEO) program, which catalogued the objects and tried to determine their characteristics. In 2005, NASA's mission was expanded to track near Earth objects greater than 140 m in diameters. Obviously the largest dinosaur-killing asteroids are the most dangerous. However, the question arises, how small does an asteroid have to be before we don't have to worry about it? Little is known about asteroids smaller than 140 m in diameter, and whether they are safe to ignore. What if one exploded over an ocean. Could it generate a tsunami that would change it from a regional to a more global hazard that would threaten coastal populations far away?

As it turns out, in February, 2013 an approximately 20-m asteroid exploded about 15 miles above the ground over Chelyabinsk, Russia. This airburst provided an unprecedented opportunity for data collection. Teams of scientists visited, collected

With many thanks to my collaborators Michael Aftosmis, Jonathan Goodman and Randy LeVeque.

M. Berger (✉)
Courant Institute, New York University, 251 Mercer St., New York, NY 10012, USA
e-mail: berger@cs.nyu.edu

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_1

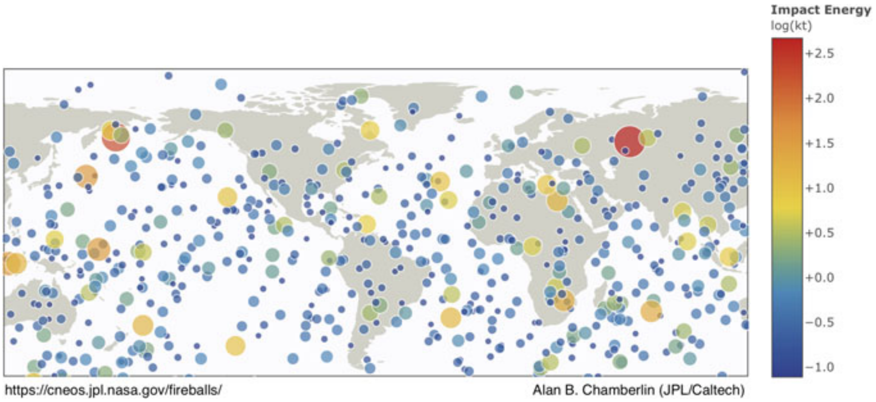

Impact Energy
log(kt)
+2.5
+2.0
+1.5
+1.0
+0.5
+0.0
−0.5
−1.0
https://cneos.jpl.nasa.gov/fireballs/
Alan B. Chamberlin (JPL/Caltech)

Fig. 1 Airbursts reports from April, 1988 to Dec, 2019. Figure taken from https://cneos.jpl.nasa.gov/fireballs

samples of the meteor to determine its composition, analyzed web cams from Russian cars to determine the trajectory and energy deposition, canvassed the region to see how far away windows broke (evidence of the blast overpressure), etc. [15]. In other words, data was collected that could be used for model validation. The ATAP project started shortly thereafter.

A reader might wonder how often such airbursts really occur. Figure 1 shows that in fact airbursts happens quite regularly. Since most of the world's surface is water, an investigation into airburst-generated tsunamis seems warranted.

In this talk I will focus only on simulations of smaller asteroids that explode before hitting the ground. There is very little literature on the effects of these airbursts. There is some literature on simulations of larger asteroids that do reach the ocean, and sometimes reach the ocean floor [4, 17, 18]. Impact simulations are generally performed using hydrocodes that simulate material deformation and failure, multimaterial phase changes (e.g. water turns into vapor and rises through the atmosphere), sediment excavation from the ocean floor, shock waves traveling through water, etc. A nice discussion can be found in the chapter by Gisler in [6]. These are very expensive calculations, so they tend to be axisymmetric to reduce cost, including the bathymetry.[1] Asteroid impact simulations is a dynamic area that is receiving a lot of recent attention [12, 13, 16].

In the next section we will present our simulations using the shallow water equations modeled with the GeoClaw software package, and describe how GeoClaw was adapted to model asteroid airbursts. We will review our analysis of a model problem that helps understand the simulations results. However, it turns out that airburst-generated tsunamis have smaller length scales that earthquake-generated tsunamis. Hence we will turn to the linearized Euler equations to bring in the effects of compressibility and dispersion. It will turn out that dispersion is a much more important

[1] Bathymetry is underwater topography.

factor at the length scales and pressures of interest, and luckily the shallow water equations seem to overestimate the effect. We will conclude that airburst-generated tsunamis do not pose a global threat. This was the conclusion reached by all participants in the joint NASA-NOAA tsunami workshop in 2016 using a variety of codes and test problems, summarized in [11].

2 Simulations of Airburst-Generated Tsunamis

2.1 Background

The simulations we first present use the open-source software package GeoClaw [9]. GeoClaw solves the depth-averaged shallow water equations on bathymetry. It uses a second order finite volume scheme with a robust Riemann solver to deal with wetting and drying [5]. Very important for trans-oceanic wave propagation where coastal inundation is also important is the use of adaptive mesh refinement. GeoClaw uses patch-based mesh refinement, allowing resolution in deep water with grid cells the size of kilometers, and on land on the order of meters. Other issues such as well-balancing (an ocean at rest on non-flat bathymetry stays at rest), and a well-balanced and conservative algorithm for adding and removing patches, are also part of GeoClaw. Desktop-level parallelism using OpenMP has also been implemented. There is no data from asteroid-generated tsunamis to use for benchmarking. We mention however that GeoClaw has had many benchmarking studies performed for earthquake-generated tsunamis, especially extensively in 2011 in [7]. This set of benchmarks was performed to allow GeoClaw to be used in hazard assessment work funded by the U.S. National Tsunami Hazard Mitigation Program.

The shallow water equations can be derived from the incompressible irrotational Euler equation using the long wavelength scaling, by assuming the ratio $\epsilon = h/L \ll 1$. Here, h is the depth of the water and L is the length scale of interest. This scaling leads to the conclusion that the velocity of the water in the z direction only enters at $O(\epsilon)$, and the horizontal velocities are constant in the vertical direction to $O(\epsilon^2)$. Eliminating the need to compute the vertical velocity reduces the three-dimensional simulation to a much more affordable calculation using only the horizontal velocities u and v.

Ordinarily the pressure only appears as a gradient in the shallow water equations, allowing the value for the pressure itself to be set arbitrarily. In our simulations however we will need to match the pressure at the top of the water column with the atmospheric pressure produced by the asteroid blast wave. Re-deriving the shallow water equations and retaining the pressure produces the following set of equations for simulation:

$$
\begin{aligned}
h_t + (hu)_x + (hu)_y &= 0 \\
(hu)_t + \left(hu^2 + \frac{1}{2}gh^2\right)_x + (huv)_y &= -\,ghB_x - \frac{h\,p_{ex}}{\rho_w} - Du \\
(hv)_t + (huv)_x + \left(hv^2 + \frac{1}{2}gh^2\right)_y &= -\,ghB_y - \frac{h\,p_{ey}}{\rho_w} - Dv
\end{aligned}
\tag{1}
$$

The other terms in (1) are g, gravity, p_e, the external atmospheric pressure at the water surface, and $\rho_w = 1025\,\text{kg/m}^3$ is the density of salt water. $B(x, y)$ is the bathymetry (underwater topography, or depth of the ocean floor). Note that the pressure forcing appears in a non-conservative form, as does the bathymetry. In these equations, a flat ocean would have $h(x, y) = -B(x, y)$. This is often described using the water elevation $\eta(x, y) = h + B$, where sealevel is $\eta(x, y) = 0$. In these equations we have neglected the Coriolis force (often considered unimportant for tsunami propagation). The term $D = \frac{gM^2\sqrt{(u^2+v^2)}}{h^{1/3}}$ is the drag, which is important in numerical simulations that include inundation. $M = 0.025$ is the Manning coefficient which we take to be constant.

To simulate the equation set (1), the external pressure must be known. This is obtained from detailed simulations of an asteroid entering the earth's atmosphere at a given speed, angle, and material composition, performed by others in the ATAP project [1]. The asteroid deposits its energy in the atmosphere, causing a blast wave. The simulations extract the ground pressure $p_e(x, y)$, and the width and amplitude of a Friedlander profile, an idealized blast wave profile, is fit to the data. This functional form is then used in the simulations for the pressure forcing. For simplicity we use a radially symmetric source term corresponding to a vertical entry angle for the asteroid. (In other simulations we have performed anisotropic simulations, with no change to our conclusions.) The blast wave in these simulations travels at 391.5 m/s, which we take to be constant. This is somewhat faster than the speed of sound in air.

Figure 2 shows a typical profile. A Friedlander profile has a characteristic width that describes the distance from the leading shock to the ensuing underpressure. Figure 2 is used in the simulations as follows: At a given time t in the simulation, each grid point needs to evaluate the atmospheric pressure. If the leading blast wave travels at speed $s = 391.5$ m/s, then at time t it has travelled a distance $d = 391.5 \times t$ meters. If the grid point is farther than d from the initial location of the blast wave there is no change to the ambient pressure. If it is less, the pressure profile is evaluated at that distance away and fed to the solver. The blue curve in Fig. 2 shows the profile at 50 s. The amplitude of the overpressure at that time is approximately 100% of ambient pressure. It is zero ahead of the blast, and decays as it gets closer to blast center. These values are used in Eq. (1).

The simulation in Fig. 2 resulted from a 250 MT asteroid. This roughly corresponds to a meteor with a 200 m diameter entering the atmosphere with a speed of 20 km/s. Note that the maximum overpressure of the airburst is approximately 450%. (Explosions are measured in terms of MT (megatons) of TNT, relating the equivalent destructive power to the uses of dynamite; this is also used to quantify nuclear

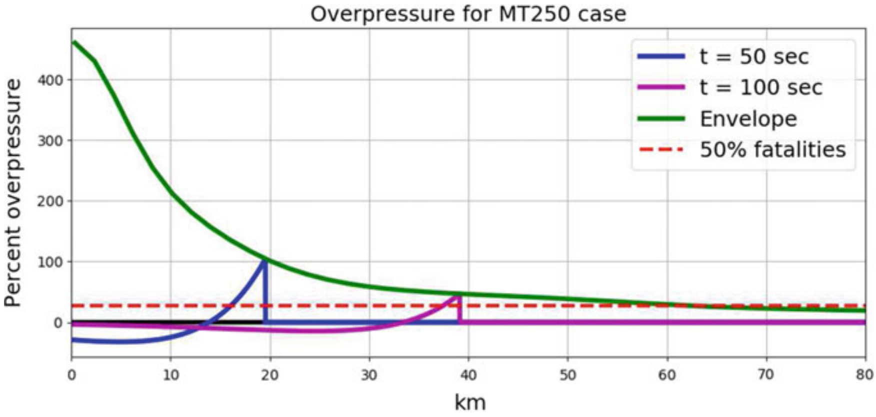


Fig. 2 A typical blast wave profile is drawn at two times. The amplitude is fit with a sum of decaying exponentials and the profile is scaled to get the pressure forcing at a given time. This functional form is then used in numerical simulations

bombs). For comparison, the explosion of Mount Saint Helens was estimated to be 25–35 MT. The largest volcanic explosion ever records was Mount Tamboura, which was approximately 10–20 Gt, and caused global climate change and mass destruction. The airburst over Chelyabinsk was approximately 520 KT. The Tunguska event, the largest airburst of the previous century, is now thought to be about 15–20 MT.

We point out that the length scale of the Friedlander profiles are significantly shorter than those of earthquake-generated tsunamis, which are typically on the order of 50–100 km. We will come back to this point in Sect. 3.

2.2 Analytical and Computational Results for Shallow Water Equations

In [2], we propose and analyze a one-dimensional model problem that helps describe the results seen in our simulations. The model problem first assumes that the pressure disturbance is a traveling wave and then builds on this to solve the problem where the pressure disturbance starts impulsively at time zero. Of course the actual pressure disturbance is a decaying function that will generate further waves as it changes amplitude, but the initial waves are the strongest and most important.

When the pressure pulse from the airburst hits the water, it causes two distinct waves with two different wave speeds. One will be related to the pressure pulse with speed s_b, and the other is the gravity wave, moving with speed s_g. What we call the *response wave* is an instantaneous disturbance of the sea surface that is in direct response to the amplitude of the moving pressure pulse and that propagates at the same speed, $s_b = 391.5\,\text{m/s}$ (this is called η above, but we change notation here to indicate it is a response to the pressure forcing).

Our analysis shows the following relationship between the response wave and the pressure disturbance p_e:

$$h_r = \frac{h_0 p_e}{\rho_w (s_b^2 - s_g^2)} \tag{2}$$

In (2), h_0 is the undisturbed height of the water (i.e. when $\eta = 0$). This shows that the response wave is stronger is deeper water, (almost linearly, since s_g depends on h_0 too). For 4.5 times atmospheric pressure, at a depth of 3 km, the response wave would have an initial height of approximately 10.8 m. This amplitude would decay rapidly with the strength of the blast wave. Note that this response wave has *positive* amplitude, since $p_e > 0$ and $s_b > s_g$. This is counterintuitive, since one would think that pushing on water would have lower its height. With hurricanes, the air pressure disturbance is negative, and hurricane travel slower than water waves, so again the water height increases, but this is more intuitive.

There are also *gravity waves* which move at the slower speed $s_g = \sqrt{gh}\,\text{m/s}$. When $h = 3000\,\text{m}$, this gravity wave moves at slightly less than 171 m/s, less than half the speed of the response wave. The initial gravity waves generated can also be estimated by linearizing the model problem and solving the homogeneous equation to get:

$$h(x,t) = h_r(x - s_b t) - \left(\frac{s_b}{s_g} + 1\right)\frac{h_r(x - s_g t)}{2} + \left(\frac{s_b}{s_g} - 1\right)\frac{h_r(x + s_g t)}{2} \tag{3}$$

The first term in (3) is the response wave traveling at blast wave speed s_b, and the next two are the gravity waves moving to the right and left with speed s_g. We see that their amplitude is also a function of the amplitude of the response wave.

We next show results from two simulations at different distances from shore and ocean depths. More details on these particular simulations are in [2]. The first set of simulations are located off the coast of Westport, Washington. This area has been well-studied because of its proximity to the earthquake-prone M9 Cascadia subduction zone. The blast was located 180 km from shore, about 30 km from the continental shelf, and the ocean was 2575 m deep underneath the blast. Figure 3 shows the region of interest.

Figure 4 shows 3 snapshots at intervals of 25 s after the blast wave. A black circle is drawn indicating the location of the blast, the red just inside the circle is the response wave, and further interior to the circle is the gravity waves. Note that the leading gravity is a depression (negative amplitude). Contours of the bathymetry from -1000 to -100 are drawn to show the location of the continental shelf. Although the colorbar scale is from -1 to 1, the response wave height near the blast is over 10 m.

Figure 5 shows a zoom of the waves approaching shore (2000 s), about to hit the peninsula (3000 s), and mostly reflecting (4000 s), with some smaller waves entering Grays harbor. Note that the landscape is better resolved as the waves approach,

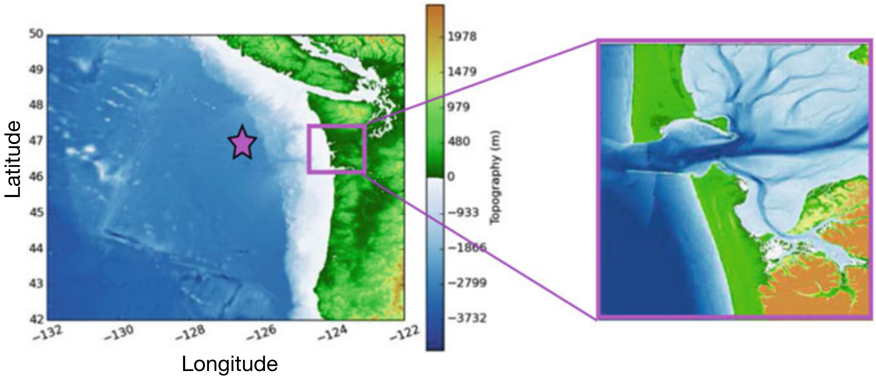


Fig. 3 The first set of simulations has the blast located 180 km offshore from Westport, in 2575 m deep water, indicated by the purple star. The zoom shows the region of interest studied for inundation

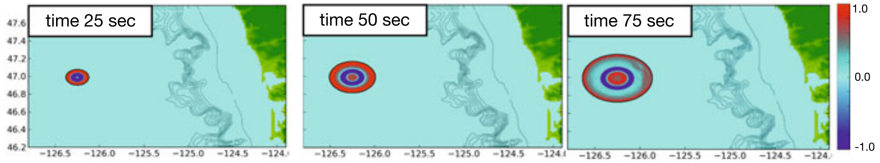


Fig. 4 Westport simulations at intervals of 25 s after the blast. The waves are spreading symmetrically around the blast center. The largest wave is over 10 m at the start

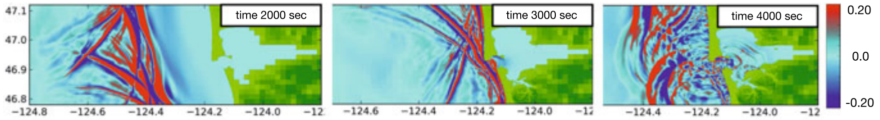


Fig. 5 Selected times as gravity waves approach Westport coastline. The zooms cover a changing region closer and closer to shore. No inundation is observed. Note the colorbar scale is a factor of 5 smaller than in the figure above

indicating that the refinement level has increased. The wave amplitudes have greatly decreased, and no inundation is observed. Note that the colorbar scale (in units of meters) has been reduced by a factor of 5 in these later plots.

Since the first set of results did not show any inundation despite such a large blast, the second set puts the blast much closer to shore. We locate the blast 30 km off the coast of Long Beach, California, an area with a lot of important infrastructure. Figure 6 shows the topography. The water at the center of the blast is 797 m deep.

Figure 7 shows 3 snapshots at intervals of 25 s after the blast wave. Several features are evident. The black circle, which indicates the location of the blast wave at that time, no longer coincides with the leading elevation of the response wave (the red contours). This is because the topography becomes more shallow at the blast wave approaches Catalina Island, so its instantaneous amplitude has decreased, as expected

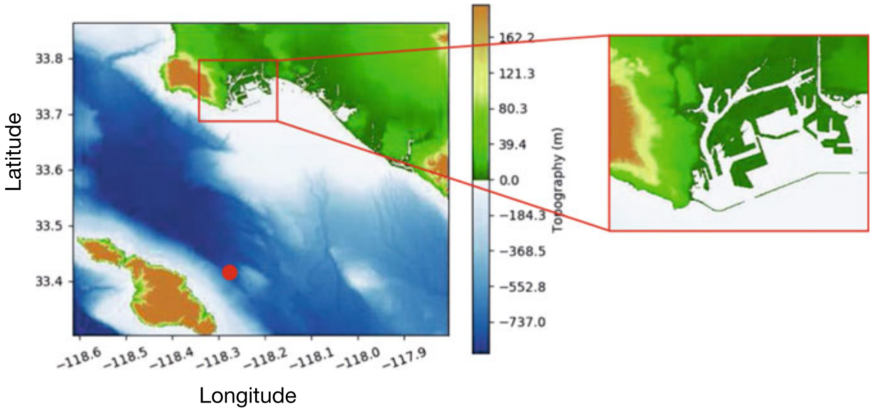


Fig. 6 The second set of simulations has the blast located 30 km from Long Beach, in 797 m deep water, indicated with the red dot. The zoom shows the region of interest studied for inundation

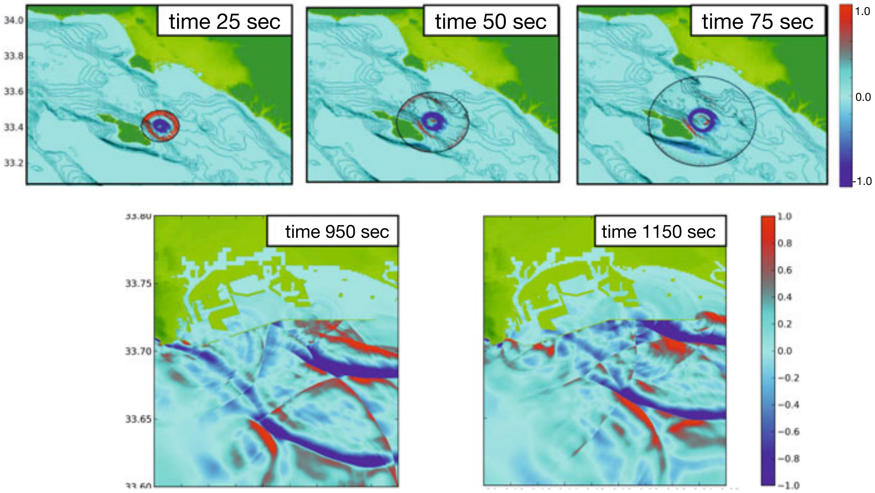


Fig. 7 First row shows computed solution for Long Beach simulation at intervals of 25 s after the blast. The black circle indicates the location of the blast wave in air. Bottom row shows zooms near shore at two later times

from Eq. (2). Also notice that that atmospheric blast wave in the atmosphere jumps over the island, and the response wave reappears when the blast is again over water. Once again we see that the gravity waves are mostly a depression.

With this proximity to shore, the blast wave has not greatly decayed before it hits shore. The blast wave will be the more important cause of casualties and damages, and not the ensuing tsunami. The zooms in Fig. 7 have more refinement than the early times. The breakwater is now resolved, and water only approaches shore through the breakwater gaps or around the edge. But since the port infrastructure is two meters high, there is still no flooding. A very tiny bit of flooding is seen along the river (not visible in these plots).

We performed a number of additional simulations in a variety of locations, bathymetries, and asteroid strengths, including one with one Gt of energy. We have not found any examples where airbursts have caused significant onshore inundation. However, in the next section we examine whether the shallow water equations is an appropriate model for airburst-generated tsunamis, and compare the previous results with similar analyses and computations using the linearized Euler equations.

3 The Linearized Euler Equations

As reviewed earlier, the shallow water equations are a long wavelength approximation to the full 3D equations. Since the length scales of the Friedlander profile are on the order of 10 km, the ratio of water depth to length scale is not that small in a 4 km ocean. Closer to shore the shallow water equations may be more appropriate. The length scales are also important in determining the effect of dispersion, which is not present in the shallow water equations.

To examine this more closely, we compare the results from the previous section using the shallow water equations with those from the linearized Euler equations. This brings in the effects of both compressibility and dispersion. The latter equations have the advantage that the free surface boundary condition of the full Euler equations becomes a simple boundary condition when linearized, so the free water surface and the atmosphere do not have to be tracked or computed. Unfortunately it does require that the vertical direction be discretized along with the two horizontal directions, and so is much more expensive than a depth-averaged equation set.

3.1 Analytical and Computational Results for Linearized Euler

Again, we first review the results from [2] for our model traveling wave problem but for the linearized Euler equations (which are also derived there). Unlike the shallow water equations, which do not have any dependence on wave length, there is such a dependence in the Euler equations. We first present results for a single frequency k, where the length scale $L = 2\pi/k$. We then apply our results to a function with many frequencies. Finally we show some preliminary results of radially symmetric simulations confirming the model problem conclusions.

If we denote the external pressure forcing $p_e(m) = A_k e^{ikm}$, where $m = x - s_b t$ is the traveling wave variable in our model problem, we can compute the response coefficients as a function of wave number, i.e. $h_r(m) = \widehat{h_r} e^{ikm}$ and amplitude A_k, and similarly for the velocities u and now the vertical velocity w too. The traveling wave

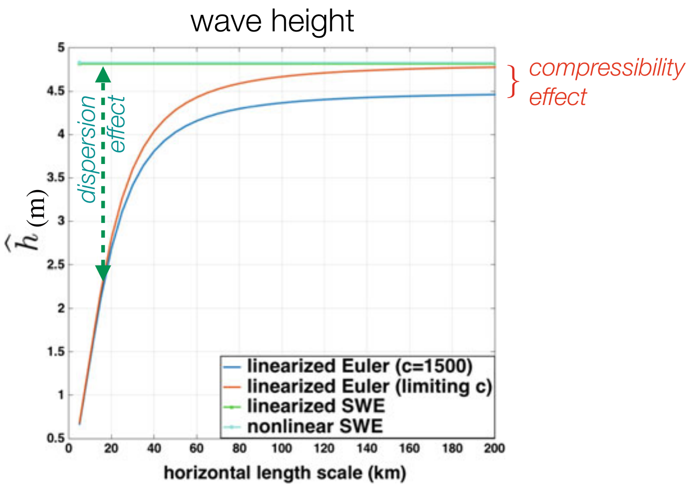


Fig. 8 Comparison of response wave amplitudes as a function of length scale for the shallow water and linearized Euler equations. These were evaluated for a 4 km deep ocean, and 1 atm overpressure. At smaller length scales the dominant difference is due to dispersion, not to compressibility

problem can no longer be solved exactly, but can be evaluated numerically. In Fig. 8, we evaluate the solution to the model problem using an ocean depth of 4 km, and an amplitude of 1 atmosphere for the overpressure. We take the speed of sound in water $c_w = 1500$ m/s, and density $\rho_w = 1025$ kg/m^3. Figure 8 also evaluates the results for an artificially faster speed $c_w = 10^8$, in order to approach the incompressible limit.

The green curve in Fig. 8 is the shallow water amplitude of the response wave. It is constant, since as expected there is no dependence on wave number. We can also compute the nonlinear response, which is done in [2], and overlays the linearized response. The blue curve is the linearized Euler result using the real sound speed of water. This does not appear to approach the shallow water curve. The red curve uses the artificially larger sound speed $c_w = 10^8$, which approaches the incompressible limit and does approach the shallow water curve, giving us more confidence in the results. The difference between the linearized Euler curve and the shallow water curve is roughly 10%. We are calling this the effect due to compressibility. However, at the length scale of interest for airburst-generated tsunamis, the difference between the curves is over a factor of 2. We conclude that dispersion is a much more important effect.

Figure 8 showed the amplitude response due to a single frequency pressure perturbation. In Fig. 9 we evaluate the response to a Gaussian pressure pulse $p_e(m) = \exp(-0.5(m/5)^2)$ that includes all frequencies. We take the Fourier transform, multiply each frequency by the Fourier multiplier shown in Fig. 8 and transform back, so this is still a *static* response. The left figure shows results in 4 km deep water, and the right in 1 km deep water. Again we see that compressibility accounts for a smaller portion of the height difference between shallow water and linearized Euler results

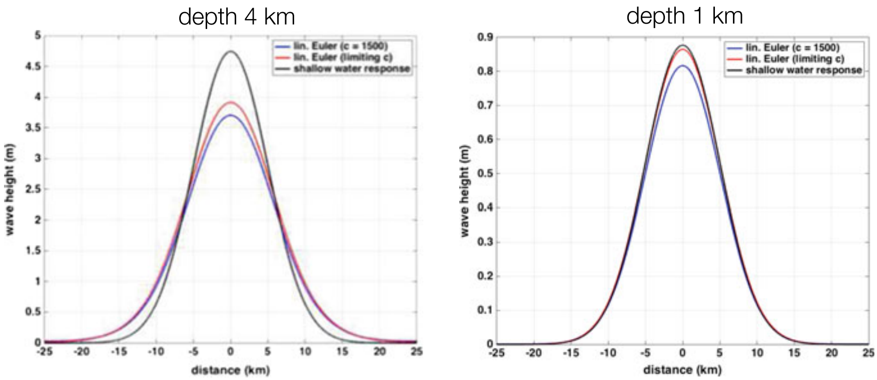


Fig. 9 Comparison of responses to a Gaussian pressure pulse in 4 km deep water (left) and 1 km deep water (right)

than dispersion. Note also that the Euler results have broadened, an indication of dispersion. The results in shallower water match better, as expected. Luckily, in all cases the shallow water results overestimate the response including compressibility and dispersion.

Finally, in Figs. 10 and 11 taken from [3] we show snapshots from time dependent simulations with the 250 Mt airburst and compare linearized Euler (denoted AG for acoustic with gravity in the legends), shallow water, and two different Boussinesq[2] models [8, 14]. We thank Popinet for the use of Basilisk in simulations using the Serre-Green-Naghdi (SGN) set of equations, and Jiwan Kim for the use of Bouss-Claw, which uses the Madsen Sørensen equation set [10].

We first show results in a 4 km deep flat ocean, then 1 km deep. Note that the scales are not the same in the two figures. Also, since the tsunami travels more slowly in shallower water, we only show those results every 100 s. Note that the leading shallow water gravity wave is a depression in both simulations. Also note that the two Boussinesq simulations agree with each other better than with the linearized Euler runs. The SGN simulation is in two space dimensions, and plotted as a function of radius, hence is much noisier than the other simulations which were one-dimensional radially symmetric computations. We point out that Boussinesq waves decay inversely proportional to distance traveled, whereas shallow water waves decay inversely to the square root of distance. Finally, all 4 codes show the same response wave behavior as an elevation in sealevel, albeit with different magnitudes.

We do not think that the depth-averaged equations are suitable for simulating the initiation of gravity waves, since there is significant variation in the vertical velocity. It does seem that depth-averaged equations can be used to propagate the waves, once

[2] Generally speaking, the Boussinesq equations keep the next term in the long-wavelength expansion for the shallow water equations. They are depth-averaged, but much more complicated than shallow water since they include dispersive terms with third order derivatives. We do not describe them further.

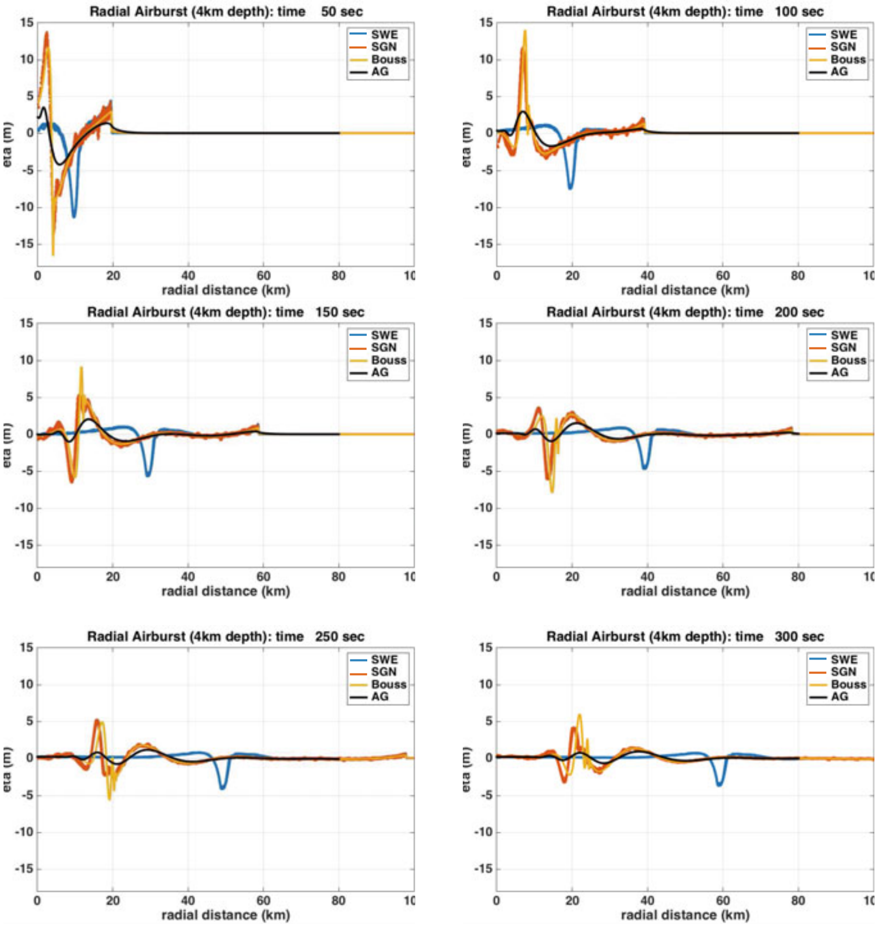


Fig. 10 Comparison of initial generation of airburst tsunami using all 4 models in a 4 km deep ocean. Selected frames every 50 s. After 300 s, the SGN and BoussClaw resuls match linearized Euler in the leading gravity wave, but not (yet) the rest. The SWE model does not generate gravity waves that match at any of the times

initiated by a higher fidelity simulation. This has been demonstrated in [3]. We do not yet know how this translates into shoreline inundation. Preliminary evidence indicates that the shallow water model provides an overestimate of run-in due to airbursts, as it did in predicting wave height for the response wave, but we need more evidence for this hypothesis.

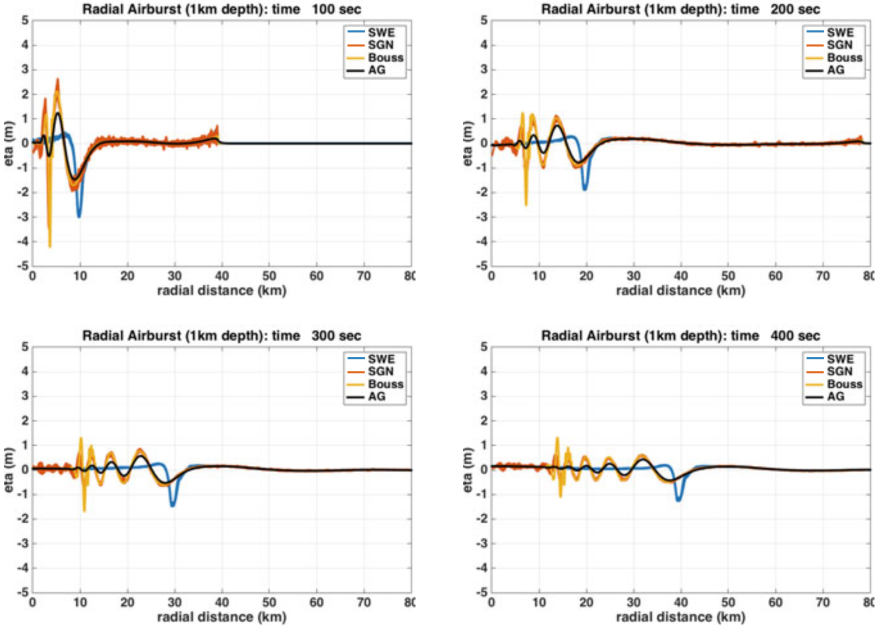


Fig. 11 Comparison of airburst generated tsunamis using all 4 models in a 1 km deep ocean. Selected frames every 100 s. After about 200 s, SGN and BoussClaw match the linearized Euler results in the leading gravity wave, and by 400 s, the next few waves are very similar, though the amplitude is not quite right. The shallow water model still has very different waves

4 Conclusions

We have presented several numerical simulations of the shallow water equations in response to a 250 Mt airburst. The results are further explained using a traveling wave model problem, for both the shallow water and linearized Euler equations. All results show that there is no significant water response (in either the response wave or the gravity wave) to the airburst. The most serious danger from an airburst would be from the blast itself if close enough to the blast center, rather than from water waves it generated.

We also found that because of the shorter wave-lengths of an airburst, the shallow water equations do not provide an accurate simulation of propagation for these waves, compared to simulations using Boussinesq or linearized Euler models. However it may be possible to use the shallow water equations to give an estimate of shoreline inundation. This is a matter for future study.

Acknowledgements Many thanks to my collaborators on this project: Michael Aftosmis, Jonathan Goodman and Randy LeVeque. This work was supported in part by a subcontract with Science and Technology Corporation (STC) under Contract NNA16BD60C, and by the Bay Area Environmental Research Institute (BAERI) under NASA Contract NNX16AO96A as part of the Asteroid Threat Assessment Project (ATAP).

References

1. Aftosmis, M.J., Mathias, D.L., Nemec, M., Berger. M.J.: Numerical simulation of bolide entry with ground footprint prediction. In: AIAA-2016-0998 (2016)
2. Berger, M.J., Goodman, J.: Airburst generated tsunamis. Pure Appl. Geophys. **175**(4) (2018)
3. Berger, M.J., Leveque, R.J.: Modeling issues in asteroid-generated tsunamis. NASA/CR 2018-219786. National Aeronautics and Space Administration, Apr 2018
4. Crawford, D.A., Mader, C.L.: Modeling asteroid impact and tsunami. Sci Tsunami Hazards **16**(1), 21–30 (1998)
5. George, D.L.: Augmented Riemann solvers for the shallow water equations over variable topography with steady states and inundation. J. Comput. Phys. **227**(6), 3089–3113 (2008)
6. Gisler, G.R.: Tsunami generation: other sources. In: E.N. Bernard, A.R. Robinson (eds.) The Sea: Tsunamis, vol. 15, pp. 179–200. Harvard University Press (2009)
7. González, F.I., LeVeque, R.J., Chamberlain, P., Hirai, B., Varkovitzky, J., George, D.L.: Geoclaw model. In: Proceedings and Results of the 2011 NTHMP Model Benchmarking Workshop, pp. 135–211. National Tsunami Hazard Mitigation Program, NOAA (2012)
8. Kim, J., Pedersen, G.K., Løvholt, F., LeVeque, R.J.: A Boussinesq type extension of the GeoClaw model—a study of wave breaking phenomena applying dispersive long wave models. Coast. Eng. **122**, 75–86 (2017)
9. LeVeque, R.J., George, D.L., Berger, M.J.: Tsunami modelling with adaptively refined finite volume methods. Acta Numer., 211–289 (2011)
10. Madsen, P.A., Sørensen, O.R.: A new form of the Boussinesq equations with improved linear dispersion characteristics. Part 2. A slowly-varying bathymetry. Coast. Eng. **18**(3–4):183–204 (1992)
11. Morrison D, Venkatapathy, E.: Asteroid generated tsunami: summary of NASA/NOAA workshop. Technical report NASA/TM-2194363. NASA Ames Research Center, Jan 2017
12. Report of the Near-Earth Object Science Definition Team: Study to determine the feasibility of extending the search for near-earth objects to smaller limiting diameters. Technical report. National Aeronautics and Space Administration. Prepared at the request of NASA Office of Space Science, Planetary Science Division, Aug 2003
13. Report of the Near-Earth Object Science Definition Team: Update to determine the feasibility of enhancing the search and characterization for neos. Technical report. National Aeronautics and Space Administration. Prepared at the request of NASA Office of Space Science, Planetary Science Division, Sept 2017
14. Popinet, S.: A quadtree-adaptive multigrid solver for the Serre-Green-Naghdi equations. J. Comput. Phys. **302**, 336–358 (2015)
15. Popova, O.P., Jenniskens, P., Emelyanenko, V., Kartashova, A., Biryukov, E., et al.: Chelyabinsk airburst, damage assessment, meteorite recovery, and characterization. Science **342**(6162), 1069–1073 (2013)
16. Robertson, D.K., Gisler, G.R.: Near and far-field hazards of asteroid impacts in oceans. Acta Astronaut. **156**, 262–277 (2019)

17. Ward, S.N., Asphaug, E.: Asteroid impact tsunami: a probabilistic hazard assessment. Icarus **145**, 64–78 (2000)
18. Weiss, R., Wünnemann, K., Bahlburg, H.: Numerical modelling of generation, propagation and run-up of tsunamis caused by oceanic impacts: model strategy and technical solutions. Geophys. J. Int. **167**, 77–88 (2006)

Some Case Studies in Environmental and Industrial Mathematics

Alfredo Bermúdez

Abstract This presentation deals with four case studies in environmental and industrial mathematics developed by the mathematical engineering research group (mat+i) from the University of Santiago de Compostela and the Technological Institute for Industrial Mathematics (ITMATI). The first case involves environmental fluid mechanics: optimizing the location of submarine outfalls on the coast. This work, related to shallow water equations with variable depth, led us to develop a theory for numerical treatment of source terms in nonlinear first order hyperbolic balance laws. More recently, these techniques have been applied to solve Euler equations with source terms arising from numerical simulation of gas transportation networks when topography via gravity force is considered in the model. The last two problems concerns electromagnetism. One of them is related to nondestructive testing of car parts by using magnetic nanoparticles (the so-called magnetic particle inspection, MPI): mathematical modelling of magnetic hysteresis to simulate demagnetization. Finally, we present a mathematical procedure to reduce the computing time needed to achieve the stationary state of an induction electric machine when using transient numerical simulation.

1 Introduction

Four case studies developed by the Research Group in Mathematical Engineering from the University of Santiago de Compostela (USC) and the Technological Institute for Industrial Mathematics (ITMATI) are considered. Two of them are related to fluid mechanics. The first one was developed in the framework of a contract with the Ministry of Public Works of Galicia and concerns shallow water flows in a domain with variable depth. The second one deals with gas flow in transport networks and has

A. Bermúdez (✉)
Departamento de Matemática Aplicada, Instituto de Matemáticas, Universidade de Santiago de Compostela, Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela, Spain
e-mail: alfredo.bermudez@usc.es

Instituto Tecnológico de Matemática Industrial (ITMATI), Rúa de Constantino Candeira, 15705 Santiago de Compostela, Spain

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_2

been done for the Reganosa company. From the mathematical point of view both are modelled with systems of nonlinear hyperbolic partial differential equations with source terms and the goal is to set up suitable finite volume discretization of the source terms.

The other two case studies concern electromagnetism. The goal of the first one, that has been financed by CIE Automotive company, is numerical simulation of magnetization and demagnetization processes in magnetic particle inspection procedures. Finally, the last case study is related to numerical solution of electric machines with optimal design in view. The underlying mathematical problems are, respectively, mathematical and numerical analysis of models for electromagnetic hysteresis, and methods to determine appropriate initial conditions for transient electromagnetic simulations, in order to attain the steady state as soon as possible.

2 Environmental Flows. The Shallow Water Equations

The technical goal of this work, commissioned by the Galician government to our research team in the eighties, was to determine the optimal location of submarine outfalls along the coast of the Galician *rias*. For this purpose several steps were done involving modelling, simulation and optimal control:

- To compute the velocity field due to tidal currents and wind which was done by using the shallow water equations
- To solve a mathematical model giving the evolution and dispersion of some pollution indicators as fecal coliforms or biochemical oxygen demand (BOD)
- To formulate and solve some constrained optimal control problems related to outfall position and management of wastewater treatment systems.

Regarding the first step, as the shallow water equation is a nonlinear system of hyperbolic partial differential equations, numerical methods developed in the eighties of the last century for Euler equations can be applied to its numerical solution. We mean finite volume methods combined with approximate Riemann solvers. The unexpected problem we found was related to the discretization of the source term which is present in the shallow water equations when the bottom is not flat. In order to give some insight we refer to Fig. 1: we have solved the shallow water equations by using a finite volume scheme with the Van Leer Q-scheme as approximate Riemann solver for flux term upwinding, and a *centred* scheme to discretize the source term arising from non-flat bottom. We have considered a static configuration in a closed channel, more precisely, the initial condition (and then the solution along the time) corresponds to water at rest. In the left plot one can see the computed water level which is a quite good approximation. However, the right plot shows the computed velocity which varies between around -60 and 80 m/s while the exact velocity is null.

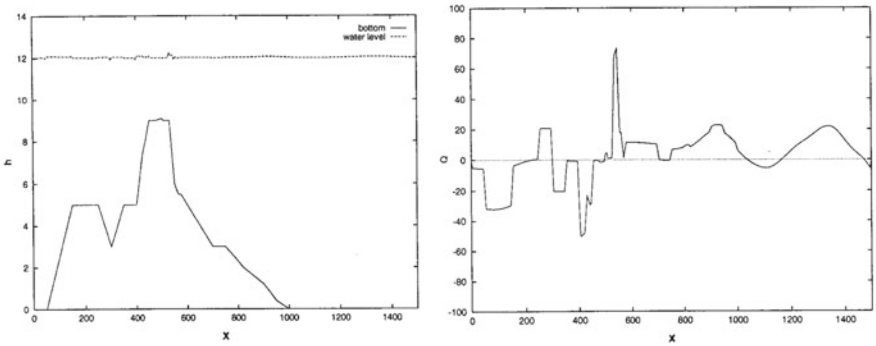


Fig. 1 Shallow water. Centred discretization of the source term. Computed water level (left) and computed velocity (right). Notice that the zero line is the result of a numerical simulation using [10]

Motivated by this problem, in the old paper [10] we developed a general methodology to discretize source terms in nonlinear systems of first order hyperbolic partial differential equations. In particular, our methods solve exactly the previous static problem. This paper is considered a seminal work in the theory of well-balanced schemes for numerical solution of conservation laws with source terms, an active field of research during the last years. Moreover, thirty years later, this methodology was applied by our research group to a different problem: Euler equations with gravity, more specifically, to numerical simulation of gas transportation networks on non-flat topography.

3 Gas Network Simulation

This industrial demand from the Reganosa company consisted in writing a software code for transient numerical simulation of a gas transport network. In Fig. 2 the high-pressure Spanish gas network is shown. Besides the great number of pipes, it includes entry (emission) and exit (consumption) points, underground storages and, more importantly, compression stations. The latter are needed to compensate the pressure drop along the network due to viscous friction of the gas on the pipe walls.

3.1 Mathematical Modelling: Homogeneous Gas Flow in a Pipe

The mathematical model for gas flow in a pipe consists of Navier-Stokes equations for compressible flows. More precisely, it involves the mass, momentum and energy conservation laws and some additional equations: the state equations for real gases and the Darcy-Weisbach law for turbulent friction between gas and pipe walls com-

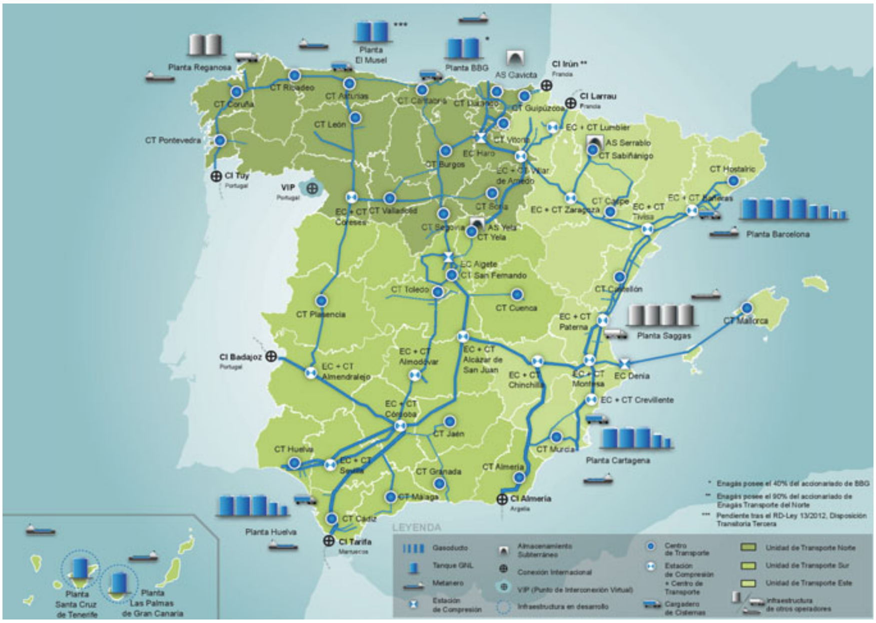


Fig. 2 Spanish gas transport network

bined with Colebrook equation to compute the friction factor. As the pipe length is much larger than the area of its cross-section we can use a 1D model:

$$\begin{aligned}
\frac{\partial \rho}{\partial t}(x,t) + \frac{\partial(\rho v)}{\partial x}(x,t) &= 0, \\
\frac{\partial(\rho v)}{\partial t}(x,t) + \frac{\partial(\rho v^2 + p)}{\partial x}(x,t) &= \underbrace{-\frac{\lambda\rho(x,t)}{2D}|v(x,t)|v(x,t)}_{\text{friction}}\, \underbrace{-g\rho(x,t)h'(x)}_{\text{gravity force}}, \\
\frac{\partial(\rho E)}{\partial t}(x,t) + \frac{\partial((\rho E + p)v)}{\partial x}(x,t) &= \underbrace{-g\rho(x,t)v(x,t)h'(x)}_{\text{power of gravity force}} \\
&\quad + \underbrace{\alpha\frac{4}{D}(\theta_{ext}(x,t) - \theta(x,t))}_{\text{heat exchange}}.
\end{aligned}$$

$$\text{Thermodynamic equation of state: } p = Z(\theta, p)\rho R\theta$$

$$\text{Caloric equation of state: } e = E - \frac{1}{2}|v|^2 \quad \text{with}$$

$$e = \hat{e}(\theta) = \hat{e}(\theta_0) + \int_{\theta_0}^{\theta} c_v(s)\,ds$$

- θ is absolute temperature (K)
- p is pressure (Pa)
- $Z(\theta, p)$ is the compressibility factor (dimensionless)
- E is the specific total energy (J/kg)
- e is the specific internal energy (J/kg)
- θ_0 is a reference temperature (K)
- $c_v(\theta)$ is the specific heat at constant volume (J/(kg K)).

3.2 Numerical Solution: One Single Pipe with Homogeneous Gas

Numerical methodology for solving the compressible Euler equations for homogeneous mixtures of perfect gases without sources has been well established since the eighties of the last century. For instance, one can use a simple first-order method consisting of Euler explicit for time discretization, finite volume method for space discretization, and approximate Riemann solvers (e.g., van Leer's Q-Scheme) for upwind discretization of the flux term (see, for instance, [24]). However, when source terms are present (e.g., the gravity term with variable heigth), numerics is more difficult and similar to the shallow water equations the use of well-balanced schemes is mandatory. This means that the discretization of source terms also needs some upwinding. In the last years many papers devoted to numerical solution of Euler equations with gravity have been written. Let us mention, for instance, [13–15, 23, 25, 27].

In order to highlight the need of using an upwind discretization of the source terms, we consider the following very simple test problem: $h(x)$ in the gravity source term is an arbitrary function and we look for a static isothermal solution, i.e., satisfying $v(x) = 0, \quad \theta(x) = \theta_{ext}, \quad \forall x \in (0, L)$. It is easy to see that the exact solution is given by $v(x) = 0$, $\rho(x) = \rho_0 \exp\left(-\dfrac{g}{R\theta_{ext}}\big(h(x) - h_0\big)\right)$, and $p(x) = R\theta_{ext}\rho_0 \exp\big(-\dfrac{g}{R\theta_{ext}}\big(h(x) - h_0\big)\big)$. For the data given in Table 1, the computed mass flow rate is shown in Fig. 3 as well as the exact solution which is null. One can see that the former is very bad, oscillating between around -10 and 10.

By using the general methodology developed in [10], we have proposed a discretization of the gravity term in [7] leading to a well-balanced scheme that reproduces the null solution exactly.

Table 1 Data for static isothermal test

R (J/(kg K))	θ_{ext} (K)	h(x) (m)	L (m)
480	288.15	$1000 \sin\left(\frac{10\pi}{40{,}000}x\right)$	40,000

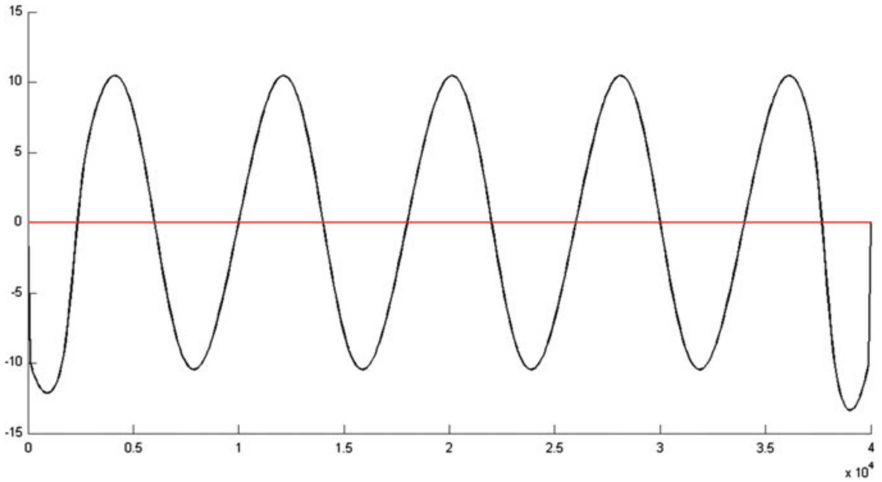


Fig. 3 Mass flux, (kg/(m^2 s)). Computed with centred discretization of source terms (black) and exact (red). The horizontal axis is the distance to the origin of the pipe

3.3 Network with Heterogeneous Gas

Simulation of heterogeneous gas flowing in a network is more difficult. New problems arise: junction modelling, gas quality simulation. These issues have been addressed in papers [8, 9].

3.4 Experimental Validation in a Real Small Network

The code has been used for a small gas network and the results have been compared to real measurements. The network can be seen in Fig. 4.

Topography is quite irregular as can be seen in Fig. 5. Results and measurements corresponding to mass flow rate and pressure for some particular nodes are shown in Figs. 6 and 7, respectively.

4 Non-destructive Testing: Magnetic Particle Inspection (MPI)

MPI is a non-destructive testing technique to detect near-surface defects in ferromagnetic pieces. The process is as follows: firstly, the workpiece is magnetized. Then, the presence of a surface discontinuity in the material allows the magnetic flux to leak, since air cannot support as much magnetic field per unit volume as metals. In order

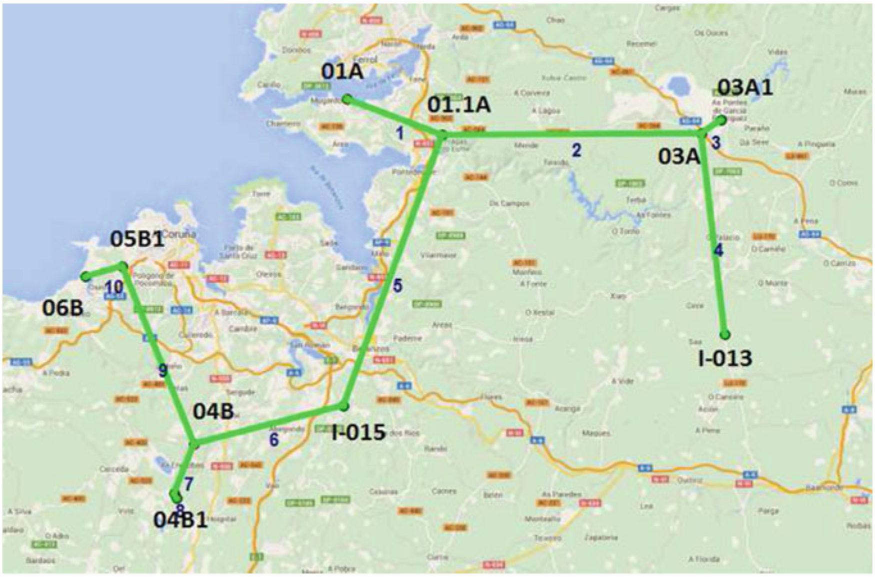


Fig. 4 The Reganosa network (Galicia. Spain)

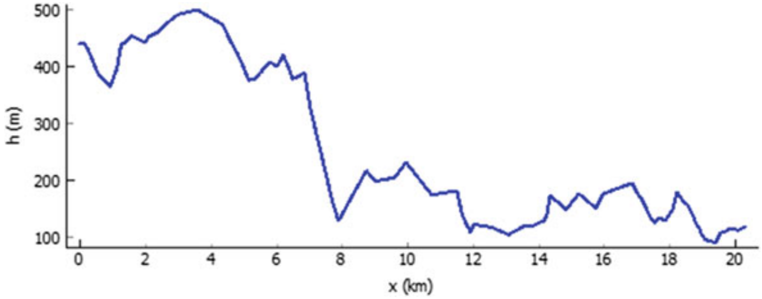


Fig. 5 Height function for edge #9

to identify a leak, ferrous particles, either dry or in a wet suspension, are applied to the workpiece. Then they are attracted to an area of flux leakage and form what is called an indication which is evaluated to determine its nature. Since cracks are more easily detected when they are perpendicular to the induced field, two magnetizations are made: circular and longitudinal. After inspection, a final demagnetization step is required for subsequent processing of the workpiece. In the next subsection we introduce an axisymmetric model for circular magnetization and present some numerical results (Figs. 8 and 9). Further details can be found in Refs. [2, 4–6].

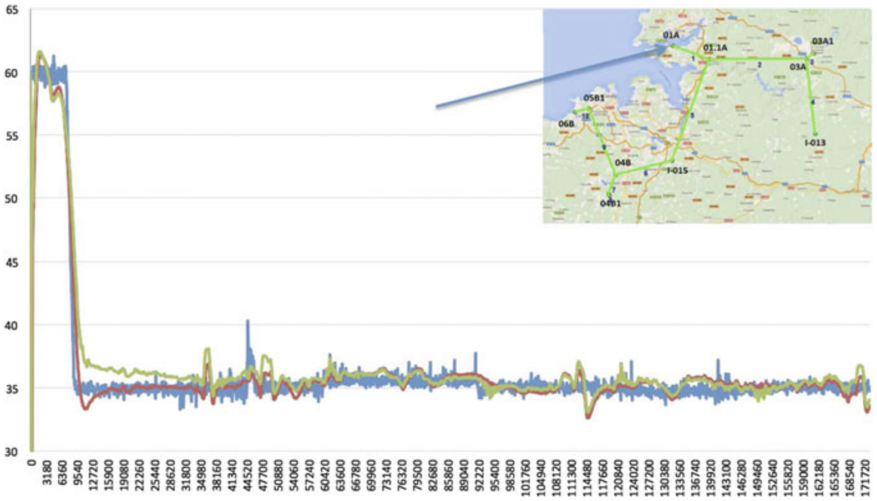


Fig. 6 Mass flow rate at node **01A**. Blue: real measurement. Red: computed with a homogeneous gas model. Green: computed with a heterogeneous gas model

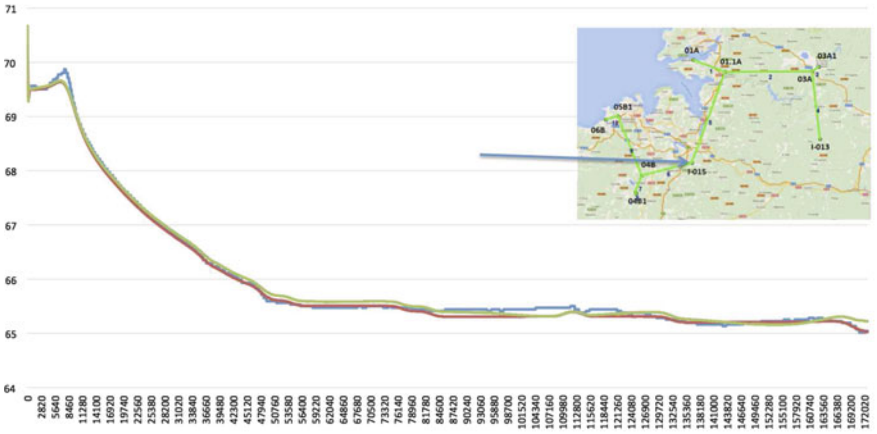


Fig. 7 Pressure at node **I-015**. Blue: real measurement. Red: computed with a homogeneous gas model. Green: computed with a heterogeneous gas model

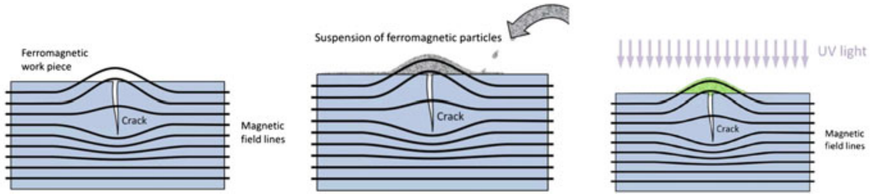


Fig. 8 Magnetic particle inspection

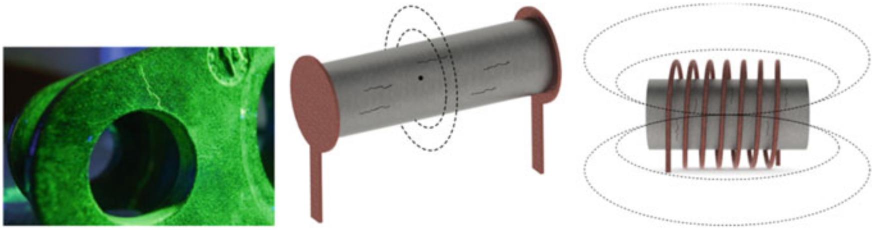
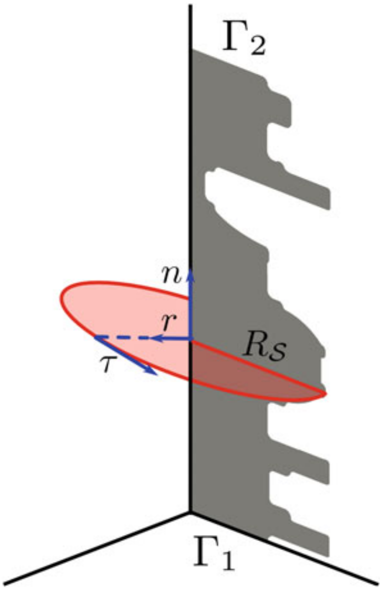


Fig. 9 Crack indication. Circular magnetization. Longitudinal magnetization

Fig. 10 Circular magnetization



4.1 Circular Magnetization. Axisymmetric Model

Let us introduce a mathematical model for circular magnetization. Thanks to axisymmetry, it can be written on a meridional section (see Fig. 10).

Given $I(t)$, the magnetizing or demagnetizing current, and an initial condition H_0, find H_θ in $\Omega \times (t_0, T]$ such that

$$
\begin{aligned}
\frac{\partial B_\theta}{\partial t}\boldsymbol{e}_\theta + \mathbf{curl}\left(\frac{1}{\sigma}\,\mathbf{curl}(H_\theta \boldsymbol{e}_\theta)\right) &= \mathbf{0} \quad \text{in } \varOmega \times (t_0, T],\\
H_\theta(0, z, t) &= 0 \quad \text{on } (0, L) \times (t_0, T],\\
H_\theta(R_{\mathcal{S}}(z), z, t) &= \frac{I(t)}{2\pi R_{\mathcal{S}}(z)} \quad \text{on } (0, L) \times (t_0, T],\\
\frac{\partial H_\theta}{\partial z}(\rho, z, t) &= 0 \quad \text{on } (\varGamma_1 \cup \varGamma_2) \times (t_0, T],\\
H_\theta(\rho, z, t_0) &= H_0(\rho, z) \quad \text{in } \varOmega.
\end{aligned}
$$

and

$$B_\theta(x, t) = \mathcal{B}(H_\theta(x, .), \xi(x))(t),$$

where $\mathcal{B}$ is a scalar *hysteresis operator* to be defined later.

4.2 Hysteresis Modelling

Mathematical modelling of hysteresis is now a well established subject (see, for instance, the reference books [11, 12, 17–19, 26]). Let us summarize the main issues of the theory. We consider a system whose state is characterized by two scalar variables, u and w, which are assumed to depend continuously on time t. In our case $u = H_\theta$ and $w = B_\theta$. The value of $w(t)$ is determined by $u(t)$ and by the values of $u(\tau)$ for $\tau < t$. Let us introduce some basic definitions and notations (Fig. 11).

At any instant t, $w(t)$ depends on the previous evolution of u, and on an initial state of the system to be called ξ. We can formalize this as follows:

$$w(t) = \mathcal{F}(u, \xi)(t) \quad \forall t \in [0, T].$$

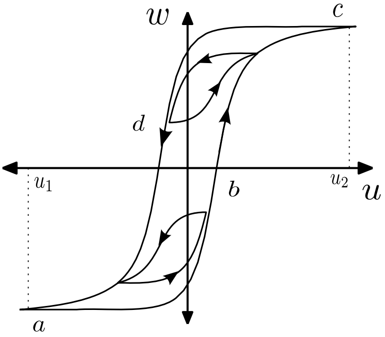


Fig. 11 Hysteresis major and minor loops

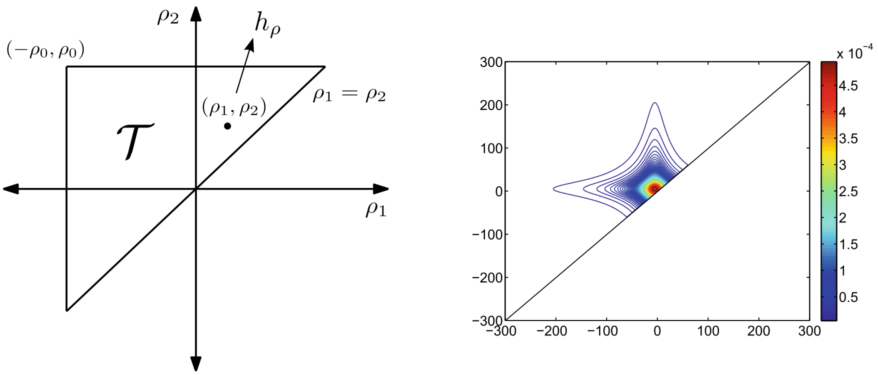


Fig. 12 Preisach triangle (left) and an example of Preisach function (right)

Here $\mathcal{F}(\cdot, \xi)$ represents an operator between suitable spaces of time-dependent functions. Notice that $\mathcal{F}$ is non-local in time. A particular example of hysteresis operator is the Preisach operator:

$$\mathcal{F} : C^0([0, T]) \times Y \longrightarrow C^0([0, T]),$$
$$[\mathcal{F}(u, \xi)](t) := \int_{\mathcal{T}} [h_\rho(u, \xi(\rho))](t)p(\rho)d\rho,$$

where $\mathcal{T}$ is the Preisach triangle, $0 < p \in L^1(\mathcal{T})$ is the Preisach function which is determined by physical experiments for each material (see Fig. 12), h_ρ is the relay function (see Fig. 13) and $\xi : \mathcal{T} \to \{-1, 1\}$ is a Borel measure representing the initial magnetic state.

The classical Preisach model is built with the so-called rate-independent relay: let us fix any pair $\rho := (\rho_1, \rho_2) \in \mathbb{R}^2$, $\rho_1 < \rho_2$. For any continuous function $u : [0, T] \to \mathbb{R}$ and any $\xi \in \{-1, 1\}$, we define $h_\rho(u, \xi)$ as follows.

Let $t_1 < \ldots < t_N < t$ be such that $u(t_i) \in \{\rho_1, \rho_2\}$. If $\{t_i\} = \emptyset$ or $t = 0$, then

$$h_\rho(u, \xi)(t) = \begin{cases} -1 \text{ if } u(t) \le \rho_1, \\ \ \ \xi \text{ if } \rho_1 < u(t) < \rho_2, \\ \ \ 1 \text{ if } u(t) \ge \rho_2, \end{cases}$$

else

$$h_\rho(u, \xi)(t) := \begin{cases} \ \ 1 \text{ if } u(t_N) = \rho_2, \\ -1 \text{ if } u(t_N) = \rho_1. \end{cases}$$

If we split $\mathcal{T} = S_u^+(t) \cup S_u^-(t)$, where

$$S_u^\pm(t) = \left\{(\rho_1, \rho_2) \in \mathcal{T} : [r_\rho(u, \xi)](t) = \pm 1\right\},$$

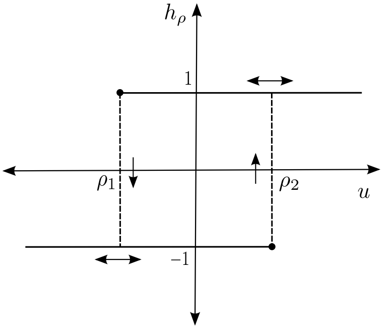


Fig. 13 Classical relay operator

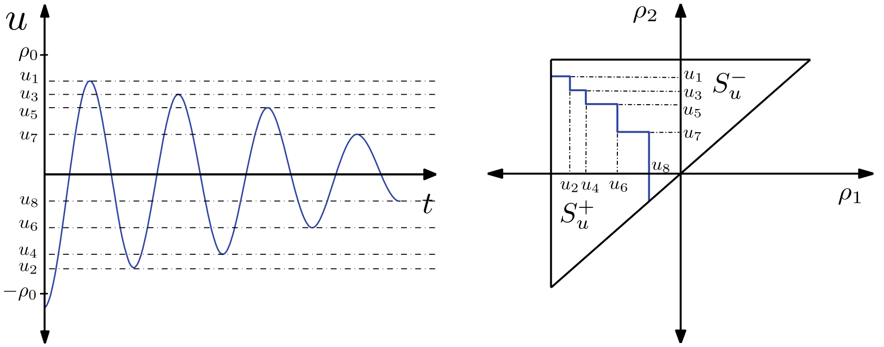


Fig. 14 Input $u(t)$ (left) and its corresponding splitting of Preisach triangle (right)

then

$$[\mathcal{F}(u,\xi)](t) := \int_{S_u^+(t)} p(\rho)d\rho - \int_{S_u^-(t)} p(\rho)d\rho.$$

We present some results obtained by solving the above model for a real crankshaft (see Fig. 14 for input data). Figure 15 shows the remanent magnetization after the circular magnetization process. In its turn, Fig. 16 shows the applied demagnetization current and the remanent magnetization after demagnetizing.

5 Accelerated Simulation of Electric Machines

In the design of electric machines (see Fig. 17), numerical simulation is an important tool. The engineer needs to know the behaviour of the machine in steady regime. In particular, he/she wants to know the torque. In order to get this steady state, finite element methods are used to solve a transient nonlinear system of PDEs derived

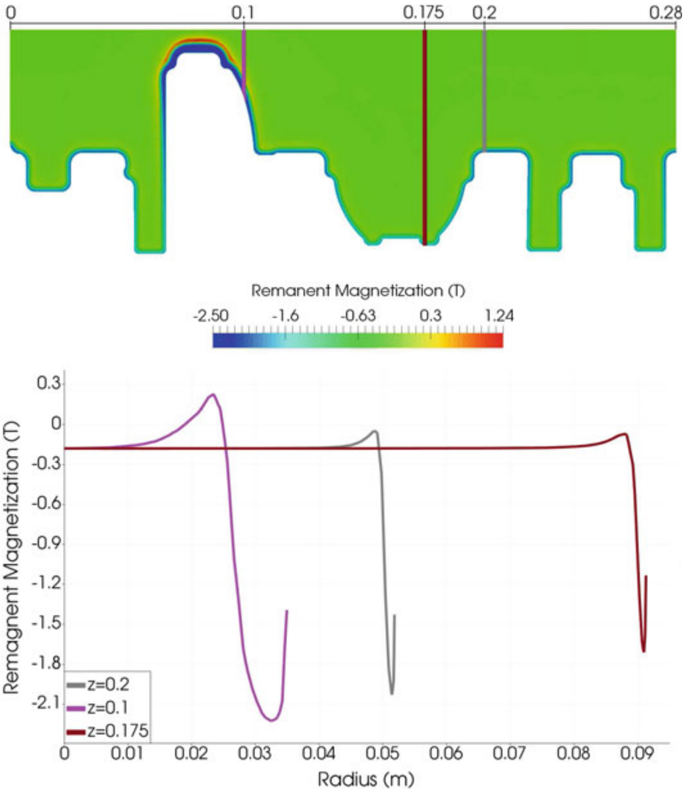


Fig. 15 Remanent magnetization

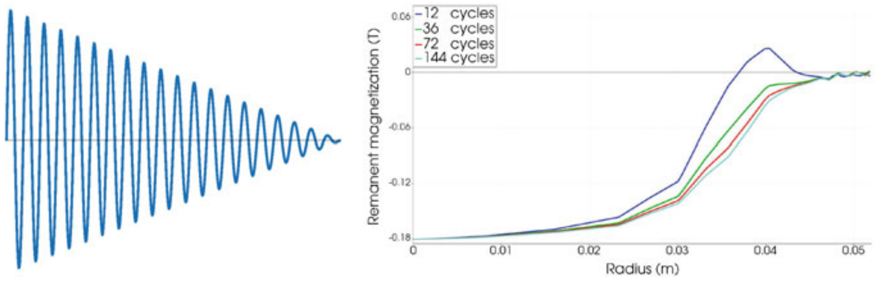


Fig. 16 Demagnetization current (left) and remanent magnetization after demagnetizing (right)

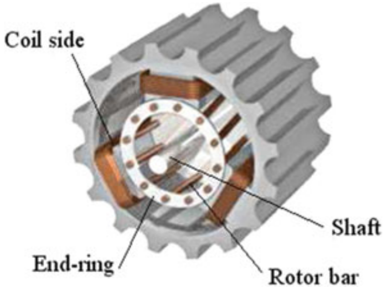


Fig. 17 Main parts integrating an induction motor. From Wikimedia Commons by Mtodorov 69 under license CC-BY-SA-3.0

from Maxwell equations, coupled with electrical circuit equations, starting from an (arbitrary) initial condition until the steady state is achieved. The time for this transient model to attain the steady state highly depends on the choice of the initial condition. When an unappropriate value is prescribed (for instance, when it is set to zero), a very long CPU time is needed to reach the steady state solution. Therefore, techniques leading to a suitable initial condition are in high demand and in the literature we can find several approaches to the problem. Let us mention, for instance, *time periodic finite element methods* [21], *time periodic-explicit error correction methods* [16], *time differential correction* [20], *parareal algorithms* [22]. A common drawback for these methods is the need of choosing a suitable time interval in which the solution is assumed to be periodic: the so-called *effective period*. Indeed, magnetic fields in rotor and stator oscillate at different frequencies and the common time at which both are periodic is generally quite large. However, the periodicity condition has to be defined in a short time interval for the method to be useful. Our methodology aims to compute a suitable initial condition and has the advantage of making use of periodicity property only in the rotor bars, so the above limitation does not apply. Moreover, the computational cost of our approach does not depend on the size of this period, and the number of unknowns is very small in comparison with the previously mentioned methods.

This work has been developed under contract with the company Robert Bosch GmbH from Stuggart (Stefan Kurz, Marcus Alexander). It has given rise to a Spanish patent. A detailed description of the methodology has been published in papers [1] and [3].

5.1 Description of the New Methodology

The main lines of the developed methodology can be described for a toy model. Let us consider a simple series circuit with an inductor and a resistor,

$$\mathrm{L}\dot{\mathrm{I}}(t) + \mathrm{R}\mathrm{I}(t) = \mathrm{E}(t),$$

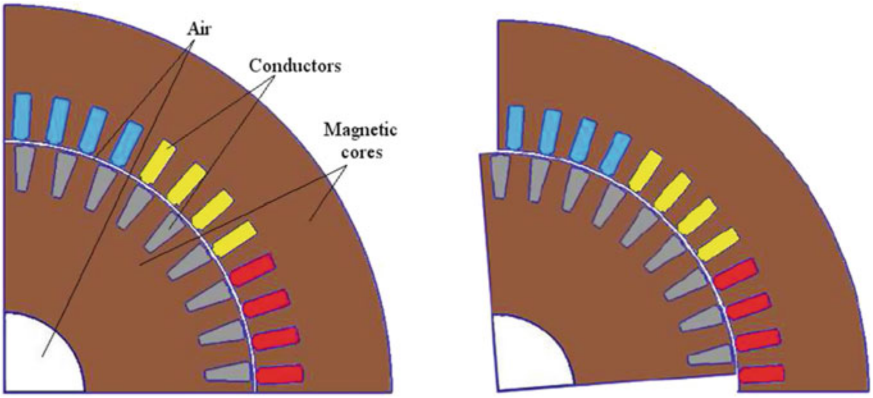


Fig. 18 A quarter of the geometric domain at time $t = 0$ (left) and $t > 0$ (right). Modification of a picture provided by Robert Bosch GmbH

with the electromotive force

$$\mathrm{E}(t) = \mathbb{E}\sin(\omega t)$$

The general solution is

$$\mathrm{I}(t) = \underbrace{Ae^{-\frac{\mathrm{R}}{\mathrm{L}}t}}_{\text{transient part}} + \underbrace{\frac{\mathbb{E}}{|\mathcal{Z}(\omega)|}\sin(\omega t - \varphi(\omega))}_{\text{steady solution}}$$

where $\mathcal{Z}(\omega) = \mathrm{R} + \omega\mathrm{L}i \in \mathbb{C}$ is the impedance of the circuit and $\varphi(\omega)$ its argument. We have two opposite extreme situations:

- If $\frac{\mathrm{R}T}{\mathrm{L}} \gg 1$, then the exponential vanishes quickly independently of the initial condition
- If $\frac{\mathrm{R}T}{\mathrm{L}} \ll 1$ then the transient part strongly depends on the initial condition. Moreover, in this case

$$\varphi(\omega) \approx \frac{\pi}{2} \text{ and } |\mathcal{Z}(\omega)| \approx \omega\mathrm{L}$$

and hence

$$\mathrm{I}(t) \approx Ae^{-\frac{\mathrm{R}}{\mathrm{L}}t} + \frac{\mathbb{E}}{\omega\mathrm{L}}\cos(\omega t).$$

If the equation is solved for $\mathrm{I}(0) = 0$, then the solution is approximately given by

$$\mathrm{I}(t) \approx -\frac{\mathbb{E}}{\omega\mathrm{L}}e^{-\frac{\mathrm{R}}{\mathrm{L}}t} + \frac{\mathbb{E}}{\omega\mathrm{L}}\cos\omega t,$$

so it includes a transient part. However, if the equation is solved for

$$\mathrm{I}(0) = \frac{\mathbb{E}}{\omega \mathrm{L}}.$$

then $A = 0$ and the transient part is close to zero from the beginning. The important remark is that, if $\frac{\mathrm{R}T}{\mathrm{L}} \ll 1$ then the above initial condition can be obtained without solving the ODE, as follows:

- Firstly, the term involving the resistor can be neglected
- Then, we integrate the equation twice: first between 0 and t and then between 0 and T. We get

$$\mathrm{L} \int_0^T \mathrm{I}(t)\, dt - \mathrm{L}T\mathrm{I}(0) = \mathbb{E} \int_0^T (T - s) \sin \omega s \, ds = \frac{\mathbb{E}T}{\omega}$$

- Moreover, since the steady solution is harmonic then $\int_0^T \mathrm{I}(t)\, dt = 0$ and from the above equation we deduce

$$\mathrm{I}(0) = \frac{1}{\mathrm{L}T} \frac{\mathbb{E}T}{\omega} = \frac{\mathbb{E}}{\omega \mathrm{L}}$$

which is the suitable initial condition previously obtained. The interesting feature of this method is that it can be used in more general settings; in particular, to the model of induction machines with squirrel cage. In this case, the problem to be solved is the following:

Given currents along the coil sides $I_n(t), n = N_b + 1, \ldots, N_c$, and initial currents along the bars y_n^0, $n = 1, \ldots, N_b$, find, for every $t \in [0, T]$, currents $y_n(t)$, $n = 1, \ldots, N_b$, along the bars such that $y_n(0) = y_n^0$, $n = 1, \ldots, N_b$, and

$$\mathcal{R}^b \frac{d}{dt} \boldsymbol{\mathcal{F}}\left(t, \mathbf{y}^b(t)\right) + \left(\mathcal{R}^b + \left(\mathcal{A}^b\right)^{\mathrm{T}} \mathcal{B}^{-1} \left(\mathcal{A}^b\right)\right) \mathbf{y}^b(t) + \lambda(t) \left(\mathcal{A}^b\right)^{\mathrm{T}} \begin{pmatrix} \mathbf{0} \\ \mathbf{e} \end{pmatrix} = \mathbf{0},$$

$$\mathcal{A}^b \mathbf{y}^b(t) \cdot \begin{pmatrix} \mathbf{0} \\ \mathbf{e} \end{pmatrix} = 0,$$

where $\boldsymbol{\mathcal{F}} : [0, T] \times \mathbb{R}^{N_b} \longrightarrow \mathbb{R}^{N_b}$ is the nonlinear operator defined as

$$\boldsymbol{\mathcal{F}}(t, \mathbf{w}) := \left(\int_{\Omega_1} \sigma A(x, y, t)\, dx\, dy, \ldots, \int_{\Omega_{N_b}} \sigma A(x, y, t)\, dx\, dy \right)^{\mathrm{T}} \in \mathbb{R}^{N_b},$$

for $t \in [0, T]$, $\mathbf{w} \in \mathbb{R}^{N_b}$, *with $A(x, y, t)$ the solution to the following nonlinear magnetostatic problem:*

Given a fixed $t \in [0, T]$, currents along the coil sides $I_n(t), n = N_b + 1, \ldots, N_c$, and $\mathbf{w} \in \mathbb{R}^{N_b}$, find a field *$A(x, y, t)$ such that*

$$
\begin{aligned}
-\operatorname{div}(\nu_0\,\mathbf{grad}A) &= 0 \quad \text{in } \Omega_0^{\text{rot}} \cup r_t\left(\Omega_0^{\text{sta}}\right),\\
-\operatorname{div}(\nu_0\,\mathbf{grad}A) &= \frac{w_n}{\operatorname{meas}(\Omega_n)} \quad \text{in } \Omega_n,\ n = 1,\dots,N_b,\\
-\operatorname{div}(\nu_0\,\mathbf{grad}A) &= \frac{I_n(t)}{\operatorname{meas}(\Omega_n)} \quad \text{in } r_t(\Omega_n),\ n = N_b+1,\dots,N_c,\\
-\operatorname{div}(\nu(\cdot,|\mathbf{grad}A|)\,\mathbf{grad}A) &= 0 \quad \text{in } \Omega_{\text{nl}}^{\text{rot}} \cup r_t\left(\Omega_{\text{nl}}^{\text{sta}}\right),
\end{aligned}
$$

with suitable transmission and boundary conditions.

5.2 Numerical Experiments with Real Electric Machines

We present the numerical results obtained for a particular induction machine with squirrel cage rotor. Firstly, we use our method to get a suitable initial condition. Next, we solve the transient model with this initial condition and compare the time needed to reach the steady-state with the one needed by taking null initial condition. The electric machine we have used for numerical experiments can be seen in Figs. 18 and 19. For confidentially issues it is a modification of a picture provided by Robert Bosch GmbH. Red, yellow and blue colors correspond to the three different phases. It is composed by 36 slots in the rotor and 48 slots in the stator. It is a three-phase machine having 2 pole pairs with 12 slots per pole. The source currents are characterized by an electrical frequency f_c and a RMS current I_c through each slot. The currents corresponding to each phase of the stator are defined as

$$
\begin{aligned}
I_A(t) &= \sqrt{2}\,I_c \cos\left(2\pi f_c t\right),\\
I_B(t) &= \sqrt{2}\,I_c \cos\left(2\pi f_c t + \frac{2\pi}{3}\right),\\
I_C(t) &= \sqrt{2}\,I_c \cos\left(2\pi f_c t - \frac{2\pi}{3}\right).
\end{aligned}
$$

We have considered four operating points corresponding to different electrical sources in the stator and different rotor velocities. They are described in Table 2. The physical time to reach the steady state for the different operating points can be seen in Table 3. Finally, in Fig. 20, the computed torque and current along the transient simulation are shown for operation point # 4.

Notes and Comments.

- We have presented four case studies in industrial mathematics, all related with numerical simulation by partial differential equations
- In addition to the industrial outcome, in all cases scientific papers related to the developed methods have been published
- This shows that industrial problems usually lead to new mathematical developments

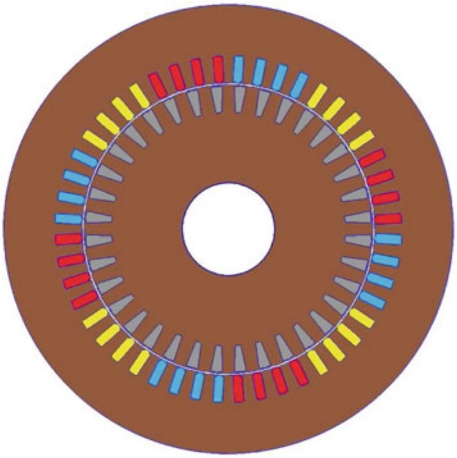


Fig. 19 Transversal section of an induction electric motor with squirrel cage

Table 2 Operation points for numerical tests

	f_c (Hz)	n_r (rpm)	I_c ($\mathrm{A_{RMS}}$)
Op. Point 1	42.1	1000	675
Op. Point 2	171.2	5000	314
Op. Point 3	417.5	12,000	675
Op. Point 4	632.0	18,000	531

Table 3 Time to get the steady state with null initial condition and with the one obtained by the new method

	Initial condition	T_{steady} (s)
Op. Point 1	$\mathbf{y}^b(0) = \mathbf{0}$	0.1200
	$\mathbf{y}^b(0) = Y\mathbf{u}$	0.0600
Op. Point 2	$\mathbf{y}^b(0) = \mathbf{0}$	0.0840
	$\mathbf{y}^b(0) = Y\mathbf{u}$	0.0120
Op. Point 3	$\mathbf{y}^b(0) = \mathbf{0}$	0.2100
	$\mathbf{y}^b(0) = Y\mathbf{u}$	0.0550
Op. Point 4	$\mathbf{y}^b(0) = \mathbf{0}$	0.3467
	$\mathbf{y}^b(0) = Y\mathbf{u}$	0.0133

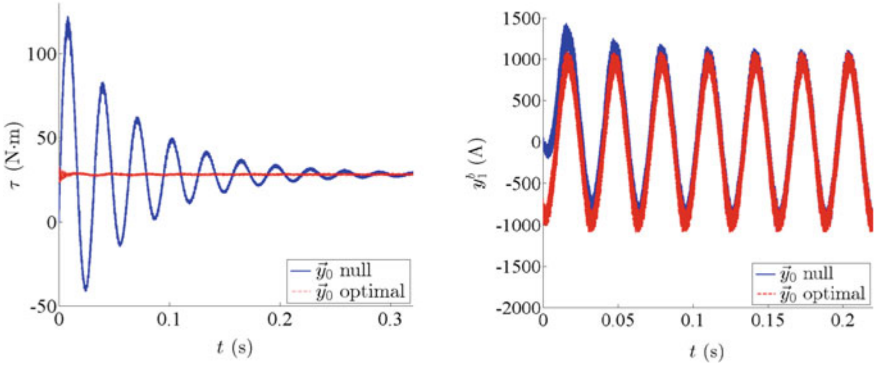


Fig. 20 Op. Point 4. Torque versus time (left). Current in bar 1 versus time (right)

- Industrial mathematics is a nice area with good opportunities for young mathematicians willing also to learn other scientific disciplines
- Postgraduate studies mixing applied mathematics and areas of application as physics, chemistry, biology, medicine, economy, etc. are a good initial step to develop a career in this promising area of increasing interest for companies and research institutions.

Acknowledgements This work has been partially supported by Robert Bosch GmbH under contract ITMATI-C31-2015, by FEDER and Xunta de Galicia (Spain) under grant GI-1563 ED431C 2017/60, by FEDER/Ministerio de Ciencia, Innovación y Universidades-Agencia Estatal de Investigación under the research project MTM2017-86459-R.

References

1. Bermúdez, A., Domínguez, O., Gómez, D., Salgado, P.: Finite element approximation of nonlinear transient magnetic problems involving periodic potential drop excitations. Comput. Math. Appl. **65**, 1200–1219 (2013)
2. Bermúdez, A., Dupré, L., Gómez, D., Venegas, P.: Electromagnetic computations with Preisach hysteresis model. Finite Elem. Anal. Des. **126**, 65–749 (2017)
3. Bermúdez, A., Gómez, D., Piñeiro, M., Salgado, P.: A novel numerical method for accelerating the computation of the steady-state in induction machines. Comput. Math. Appl. (2019)
4. Bermúdez, A., Gómez, D., Piñeiro, M., Salgado, P., Venegas, P.: Numerical simulation of magnetization and demagnetization processes. IEEE Trans. Magn. **53**(12) (2017)
5. Bermúdez, A., Gómez, D., Rodríguez, R., Venegas, P.: Mathematical analysis and numerical solution of axisymmetric eddy-current problems with Preisach hysteresis model. Rend. Semin. Mat. Univ. Politec. Torino **72**(1–2), 73–117 (2014)
6. Bermúdez, A., Gómez, D., Venegas, P.: Mathematical analysis and numerical solution of models with dynamic Preisach hysteresis. J. Comput. Appl. Math. **367** (2020). https://doi.org/10.1016/j.cam.2019.112452
7. Bermúdez, A., López, X., Vázquez-Cendón, M.E.: Numerical solution of non-isothermal non-adiabatic flow of real gases in pipelines. J. Comput. Phys. **323**, 126–148 (2016)

8. Bermúdez, A., López, X., Vázquez-Cendón, M.E.: Treating network junctions in finite volume solution of transient gas flow models. J. Comput. Phys. **344**, 187–209 (2017)
9. Bermúdez, A., López, X., Vázquez-Cendón, M.E.: Finite volume methods for multi-component Euler equations with source terms. Comput. Fluids **156**, 113–134 (2017)
10. Bermúdez, A., Vázquez-Cendón, M.E.: Upwind methods for hyperbolic conservation laws with source terms. Comput. Fluids **23**(8), 1049–1071 (1994)
11. Bertotti, G.: Hysteresis in Magnetism. Academic Press, New York (1998)
12. Brokate, M., Sprekels, J.: Hysteresis and Phase Transitions. Springer, Berlin (1996)
13. Chalons, C., Coquel, F., Godlewski, E., Raviart, P.A., Seguin, N.: Godunov-type schemes for hyperbolic systems with parameter-dependent source: the case of Euler system with friction. Math. Models Methods Appl. Sci. **20**, 2109–2166 (2010)
14. Chandrashekar, P., Klingenberg, C.: A second order well-balanced finite volume scheme for Euler equations with gravity. SIAM J. Sci. Comput. **37**(3), 382–402 (2015)
15. Käppeli, R., Mishra, S.: Well-balanced schemes for the Euler equations with gravitation. J. Comput. Phys. **259**, 199–219 (2014)
16. Katagiri, H., Kawase, Y., Yamaguchi, T., Tsuji, T., Shibayama, Y.: Improvement of convergence characteristics for steady-state analysis of motors with simplified singularity decomposition-explicit error correction method. IEEE Trans. Magn. **47**(6), 1786–1789 (2011)
17. Krejčí, P.: Hysteresis, Convexity and Dissipation in Hyperbolic Equations. Gakkōtosho Co. Ltd., Tokyo (1996)
18. Krasnosel'skiĭ, M.A., Pokrovskiĭ, A.V.: Systems with Hysteresis. Springer, Berlin (1989)
19. Mayergoyz, I.D.: Mathematical Models of Hysteresis. Springer, New York (1991)
20. Miyata, K.: Fast analysis method of time-periodic nonlinear fields. J. Math. Ind. **3**, 131–140 (2011)
21. Nakata, T., Takahashi, N., Fujiwara, K., Muramatsu, K., Ohashi, H., Zhu, H.L.: Practical analysis of 3-D dynamic nonlinear magnetic field using time-periodic finite element method. IEEE Trans. Magn. **31**(3), 1416–1419 (1995)
22. Schöps, S., Niyonzima, I., Clemens, M.: Parallel-in-time simulation of eddy current problems using parareal. IEEE Trans. Magn. **54**(3), 1–4 (2018)
23. Thomann, A., Zenk, M., Klingenberg, C.: A second order well-balanced finite volume scheme for Euler equations with gravity for arbitrary hydrostatic equilibria. Int. J. Numer. Methods Fluids **89**, 465–482 (2019)
24. Toro, E.: Riemann Solvers and Numerical Methods for Fluid Dynamics. Springer, New York (2009)
25. Varma, D., Chandrashekar, P.: A second-order, discretely well-balanced finite volume scheme for Euler equations with gravity. Comput. Fluids **181**, 292–313 (2019)
26. Visintin, A.: Differential Models of Hysteresis. Springer, Berlin (1994)
27. Xing, Y., Shu, C.-W.: High order well-balanced WENO scheme for the gas dynamics equations under gravitational fields. J. Sci. Comput. **54**, 645–662 (2013)

Modelling Our Sense of Smell

Carlos Conca

Abstract The first step in our sensing of smell is the conversion of chemical odorants into electrical signals. This happens when odorants stimulate ion channels along cilia, which are long thin cylindrical structures in our olfactory system. Determining how the ion channels are distributed along the length of a cilium is beyond current experimental methods. Here we describe how this can be approached as a mathematical inverse problem. Identification of specific functions of receptor neuron arrays is a major challenge today in both Mathematics and Biosciences. In this paper, two integral equations based mathematical models are studied for the inverse problem of determining the distribution of ion channels in cilia of olfactory neurons from experimental data.

1 Introduction

The first step in sensing smell is the transduction (or conversion) of chemical information into an electrical signal that goes to the brain. Pheromones and odorants, which are small molecules with the chemical characteristics of an odor are found all throughout our environment. The olfactory system (part of the sensory system we use to smell) performs the task of receiving these odorant molecules in the nasal mucosa, and triggering the physical-chemical processes that generates the electric current that travels to the brain. see Fig. 1 and Sect. 1.1.

What happens next is a mystery. Intuition tells us that the electrical wave generated gives rise to an emotion in the brain, which in turn affects our behavior. Of course, the workings of our other four senses is similarly a mystery. And so, we quickly come to perhaps one of the most fundamental questions in neurosciences for the future: How

C. Conca (✉)
Departamento de Ingeniería Matemática, Facultad de Ciencias Físicas y Matemáticas, Centro de Modelamiento Matemático UMR 2071 CNRS-UChile, Centro de Biotecnología y Bioingeniería, Universidad de Chile, Santiago, Chile
e-mail: cconca@dim.uchile.cl

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_3

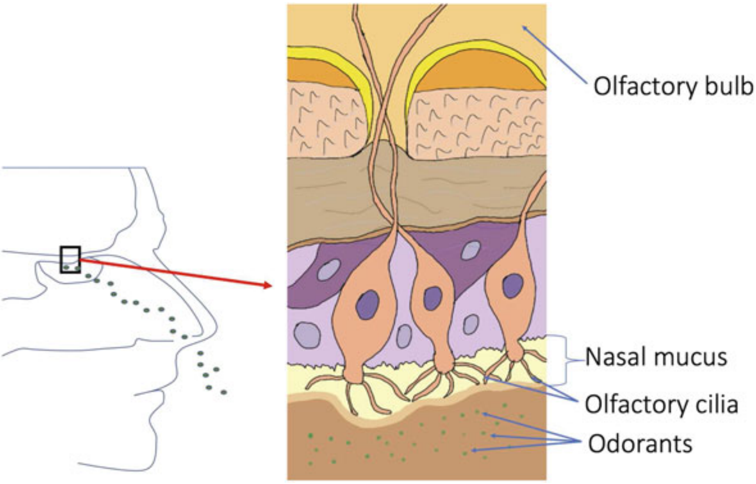


Fig. 1 Odorants reaching the nasal mucus (left) and structure of an olfactory receptor neuron (right)

does our consciousness processes external stimuli once reduced to electro-chemical waves and, over time, how does this mechanism lead us to become who we are?

How can we approach this problem with mathematics? Faced with these reflections, applied mathematicians take time to stop and wonder if it is possible to provide such far-reaching phenomena with a mathematical representation that allows us to understand and act. Biology is synonymous with "function", so the study of biological systems should start by understanding the corresponding underlying physiology. Consequently, to obtain a proper mathematical representation of the transduction of an odor into an electrical signal, and before any mathematical intervention, we must first detect which atomic populations are involved in the process and identify their respective functions.

1.1 Transduction of Olfactory Signals

The molecular machinery that carries out this work is in the olfactory cilia. Cilia are long, thin cylindrical structures that extend from an olfactory receptor neuron into the nasal mucus (Fig. 1).

The transduction of an odor begins with pheromones binding to specific receptors on the external membrane of cilia. When an odorant molecule binds to an olfactory receptor on a cilium membrane, it successively activates an enzyme, which increases the levels of a ligand or chemical messenger named cyclic adenosine monophosphate (cAMP) within the cilia. As a result of this, cAMP molecules diffuse through the interior of the cilia. Some of the cAMP molecules binds to cyclic nucleotide-gated (CNG) ion channels, causing them to open. This allows an influx of positively

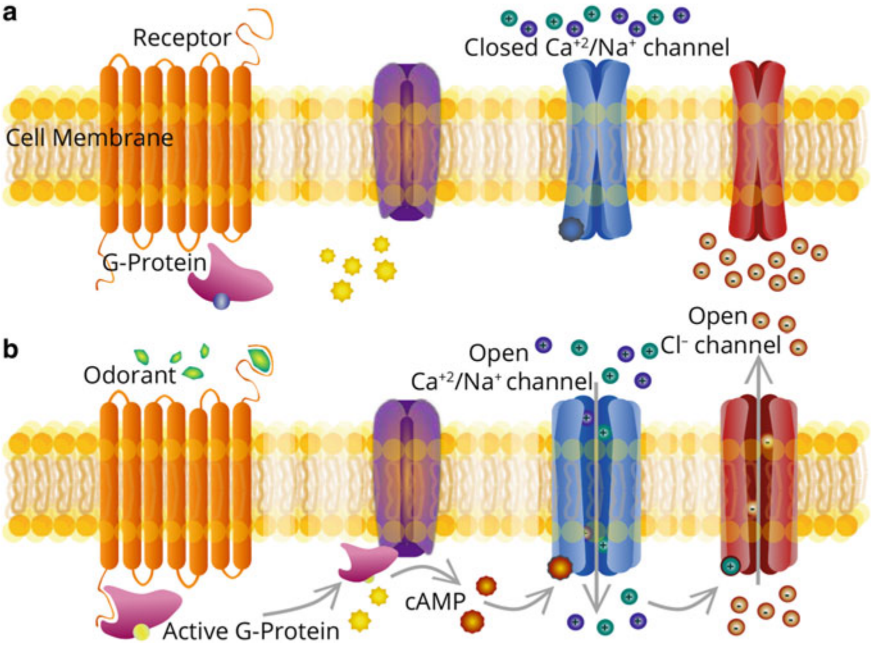


Fig. 2 Signal transduction mechanism for the olfactory system. **a** In the absence of stimulus channels are closed, system is at resting state. **b** Binding of odorants triggers cAMP synthesis and opening of CNG channels, leading to Ca^{2+} and Na^{+} transport and a Cl^{-} flux

charged ions into the cilium (mostly Ca^{2+} and Na^{+} as illustrated in Fig. 2), which causes the neuron to depolarize, generating an excitatory response. This response is characterized by a voltage difference on one side and another of the membrane, which in turn initiates the electrical current. This is the overall process that human beings share with all mammals and reptiles to smell and differentiate odors.

1.2 Kleene's Experimental Procedure

Experimental techniques for isolating a single cilium (from a grass frog) were developed by biochemist and neuroscientist Steven J. Kleene and his research team at the University of Cincinnati in the early 1990s [5, 6]. One olfactory cilium of a receptor neuron is detached at its base and stretched tight into a recording pipette. The cilium is immersed in a cAMP bath. As a result of the phenomenon previously described inside the cilium, the intensity of the current generated is recorded.

Although the properties of a single channel have been described successfully using these experimental techniques, the distribution of these channels along the cilia still remains unknown, and may well turn out to be crucial in determining the kinetics of the neuronal response. Ionic channels, in particular, CNG channels are called “micro-domains” in biochemistry, because of their practically imperceptible size. This makes their experimental description using the current technology very difficult.

1.3 An Integral Equation Model

Given the experimental difficulties, there is a clear opportunity for mathematics to inform biology. Determining ion channels distribution along the length of a cilium using measurements from experimental data on transmembrane current is usually categorized in physics and mathematics as an inverse problem. Around 2006, a multidisciplinary team (which brought together mathematicians with biochemists and neuroscientists, as well as a chemical engineer) developed and published a first mathematical model [4] to simulate Kleene’s experiments. The distribution of CNG channels along the cilium appears in it as the main unknown of a nonlinear integral equation model.

This model gave rise to a simple numerical method for obtaining estimates of the spatial distribution of CNG ion channels. However, specific computations revealed that the mathematical problem is poorly conditioned. This is a general difficulty in inverse problems, where the corresponding mathematical problem is usually ill-posed (in the sense of Hadamard, which requires the problem to have a solution that exists, is unique, and whose behavior changes continuously with the initial conditions), or else it is unstable with respect to the data. As a consequence, its numerical resolution often results in ill-conditioned approximations.

The essential nonlinearity in the previous model arises from the binding of the channel activating ligand (cAMP molecules) to the CNG ion channels as the ligand diffuses along the cilium. In 2007, mathematicians D. A. French and C. W. Groetsch introduced a simplified model, in which the binding mechanism is neglected, leading to a linear Fredholm integral equation of the first kind with a diffusive kernel. The inverse mathematical problem consists of determining a density function, say $\rho = \rho(x) \geq 0$ (representing the distribution of CNG channels), from measurements in time of the transmembrane electrical current, denoted $\mathrm{I}_0[\rho]$. This mathematical equation for ρ is the following integral equation: for all $t \geq 0$,

$$\mathrm{I}_0[\rho](t) = \int_0^L \rho(x)\ \mathbb{P}(c(t,x))\,\mathrm{d}x, \tag{1}$$

where $\mathbb{P}$ is known as the Hill function of exponent $n > 0$ (see Fig. 3). It is defined by:

$$\forall w \geq 0, \qquad \mathbb{P}(w) = \frac{w^n}{w^n + K_{1/2}^n}.$$

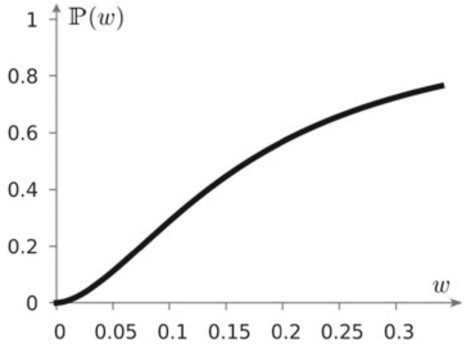


Fig. 3 The Hill function $\mathbb{P}$

In this definition, the exponent n is an experimentally determined parameter and $K_{1/2} > 0$ is a constant which represents the half-bulk (i.e., the ligand concentration for which half the binding sites are occupied); typical values for n in humans are $n \simeq 2$. Besides, in the linear integral equation above, $c(t, x)$ denotes the concentration of cAMP that diffuses along the cilium with a diffusivity constant that we denote as D; L denotes the length of the cilium, which for simplicity is assumed to be one-dimensional. Here, by concentration we mean the molar concentration, i.e., the amount of solute in the solvent in a unit volume; it is a nonnegative real number.

Hill-type functions are extensively used in biochemistry to model the fraction of ligand bound to a macromolecule as a function of the ligand concentration and, hence, the quantity $\mathbb{P}(c(t, x))$ models the probability of the opening of a CNG channel as a function of the cAMP concentration. The diffusion equation for the concentration of cAMP can be explicitly solved if the length of the cilium L is supposed to be infinite. It is given by:

$$c(t, x) = c_0 \text{erfc}\left(\frac{x}{2\sqrt{Dt}}\right),$$

where $c_0 > 0$ is the maintained concentration of cAMP with which the pipette comes into contact at the open end ($x = 0$) of the cilium (while $x = L$ is the closed end). Here, erfc is the standard complementary Gauss error function,

$$\text{erfc}(x) := 1 - \frac{2}{\sqrt{\pi}} \int_0^x e^{-\tau^2}\, \mathrm{d}\tau.$$

Accordingly, it is straightforward to check that c is decreasing in both its variables and that it remains bounded for all (t, x), $0 < c(t, x) \leq c_0$.

Despite its elegance (by virtue of the simplicity of its formulation), this new model does not overcome the difficulties encountered in its non-linear version. In fact the mathematical inverse problem associated to model (1) can be shown to be ill-posed. More precisely, since $\mathbb{P}(c(t, x))$ is a smooth mapping, the operator $\rho \mapsto \mathrm{I}_0[\rho]$ is compact from $\mathrm{L}^p(0, L)$ to $\mathrm{L}^p(0, T)$ for every $L, T > 0$, $1 < p < \infty$. Thus, even if the operator I_0 were injective, its inverse would not be continuous because, if so, then the identity map in $\mathrm{L}^p(0, L)$ would be compact, which is known to be false.

1.4 Non-diffusive Kernels

This last result certainly has a more general character. In fact, it is clear from its proof that any model based on a first-order integral equation with a diffusive *smooth* kernel necessarily results in the problem of recovering the density from measurements of the electrical current being ill-posed.

An initial, natural approach to tackling this anomaly in model (1) was developed in Conca et al. [3]. This exploited the fact that the Hill function converges pointwise to a single step function as the exponent n goes to $+\infty$, the strategy was to approximate $\mathbb{P}$ using a multiple step function.

Based on different assumptions of the spaces where the unknown ρ is sought, theoretical results of identifiability, stability and reconstruction were obtained for the corresponding inverse problem. However, numerical methods for generating estimates of the spatial distribution of ion channels revealed that this class of models is not satisfactory for practical purposes. The only feasible estimates for ρ are obtained for multiple step functions that are very close to a single-step function or, equivalently, for Hill functions with very large exponents, which imply the use of unrealistic models.

Another way to overcome the ill-posedness of the inverse problem in (1) consists of replacing the kernel of the integral equation with a non-smooth variant of the Hill function.

Specifically, let $a \in (0, c_0)$ be a given real parameter. A discontinuous version of $\mathbb{P}$ is obtained by forcing a saturation state for concentrations higher than a. By doing so, one is led to introduce the following disruptive variant of $\mathbb{P}$ (shown in Fig. 4):

$$\mathbb{H}(c) = \mathbb{P}(c)\,\mathbb{1}_{c\le a} + \mathbb{1}_{a<c\le c_0},$$

where $\mathbb{1}_J$ denotes the characteristic function of the interval J. The mathematical problem that recovers ρ from the electrical current data is therefore modelled by

$$\mathrm{I}_1[\rho](t) = \int_0^L \rho(x)\ \mathbb{H}(c(t,x))\,\mathrm{d}x, \tag{2}$$

where $c(t, x)$ is still defined as before. The introduction of this disruptive Hill function can be understood mathematically as follows: as $t \to \infty$, the factor $x/\sqrt{Dt}$ in the complementary error function defining the concentration tends to 0, and consequently $c(t, x)$ tends pointwise to c_0. An inverse mathematical problem and a direct problem are associated with both models (1) and (2). In the first, the electric current is measured

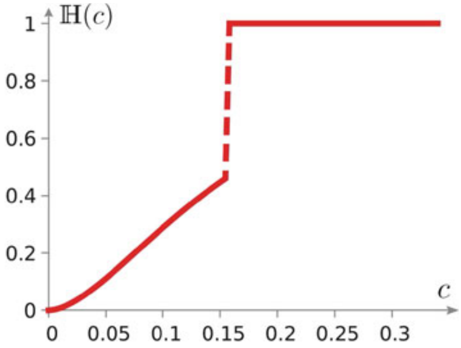


Fig. 4 A disruptive variant of $\mathbb{P}$ ($a = 0.157$)

and the unknown is the density ρ of ion channels, while in the direct problem the opposite is true. Since these are Fredholm equations of the first type, it is natural to tackle them using convolution. Once the variable ρ has been extended to $[0, \infty)$ by zero, the Mellin transform is revealed as being the most appropriate tool for carrying out this task (see the overview section "Mellin transform").

2 A General Convolution Equation

The Mellin transform is the appropriate tool to study model (2). It allows to reduce it in a convolution equation of the Mellin type. To do so, the key observation is the fact that $\mathbb{H}(c(t, x))$ can be written in terms of $\frac{\sqrt{t}}{x}$. Indeed, defining G as

$$G(z) = \mathbb{H}\left(c_0\text{erfc}\left(\frac{1}{2\sqrt{Dz}}\right)\right),$$

we have $\mathrm{I}_1[\rho](t) = \int_0^L \rho(x)G(\frac{\sqrt{t}}{x})\,\mathrm{d}x$. Thus, by extending ρ by zero to $[0, \infty)$, and rescaling time t in t^2, we obtain

$$\mathrm{I}_1[\rho](t^2) = \int_0^\infty x\rho(x)G\left(\frac{t}{x}\right)\frac{\mathrm{d}x}{x} = \Big(x\rho(x)\Big) * G$$

which is a convolution equation in $x\rho(x)$.

Taking Mellin transform on both sides and using its operational properties, we formally obtain

$$\frac{1}{2}\mathcal{M}\mathrm{I}_1[\rho](s/2) = \mathcal{M}G(s)\mathcal{M}\rho(s+1)$$

or equivalently,

$$\mathcal{M}\rho(s+1) = \frac{1}{2}\frac{\mathcal{M}\mathrm{I}_1[\rho]\,(s/2)}{\mathcal{M}G(s)}. \tag{3}$$

Mellin Transform

Austrian mathematician Robert Hjalmar Mellin (1854–1933) gave his name to the so-called Mellin transform, whose definition and properties are recalled below. The interested reader is referred to §2 of [1] or Lindelöf [7] for a summary of his work, and proof of the main results around this transform.

For $q \in \mathbb{R}$, $q + i\,\mathbb{R}$ will denote the vertical line $\{q + it, t \in \mathbb{R}\}$ of the complex plane having abscissa q, and for $p \in \mathbb{R}\,(p \geq 1)$, $\mathrm{L}^p\,([0,\infty), x^q)$, or simply L^p_q, will stand for the Lebesgue space with the weight x^q, i.e.,

$$\mathrm{L}^p_q = \left\{ f\colon [0,\infty) \to \mathbb{R} \mid \|f\|_{\mathrm{L}^p_q} < +\infty \right\},$$

where $\|f\|_{\mathrm{L}^p_q} = (\int\limits_0^\infty |f(x)|^p x^q \,\mathrm{d}x)^{1/p}$. L^p_q, endowed with this norm, is a Banach space.

Let f be in $\mathrm{L}^1\,([0,\infty), x^q)$. The Mellin transform of f is a complex-valued function defined on the vertical line $q + 1 + i\,\mathbb{R}$ by

$$\mathcal{M}f(s) = \int\limits_0^\infty x^s f(x) \frac{\mathrm{d}x}{x}$$

From its very definition, it is observed that the Mellin transform maps functions defined on $[0,\infty)$ into functions defined on $q + 1 + i\,\mathbb{R}$. Like in the Fourier transform, $\mathcal{M}f$ is continuous whenever f is in $\mathrm{L}^1\,([0,\infty), x^q)$. Specifically, we have

Theorem 1 (Riemann-Lebesgue) *The Mellin transform is a linear continuous map from* $\mathrm{L}^1\,([0,\infty), x^q)$ into $C^0(q+1+i\,\mathbb{R};\ \mathbb{C}) \hookrightarrow \mathrm{L}^\infty(q+1+i\,\mathbb{R};\ \mathbb{C})$; *its operator norm is* 1.

Proposition 1 *If f is in L^1_q for every real number q in (a,b) then its Mellin transform $\mathcal{M}f(\cdot)$ is holomorphic in the strip* $S = \{s \in \mathbb{C} \mid a + 1 < \mathrm{Re(s)} < \mathrm{b} + 1\}$.

The following table summarizes the main operational properties of the Mellin transform:

Function	Mellin transform
$f(at),\ a > 0$	$a^{-s}\mathcal{M}f(s)$
$f(t^a),\ a \neq 0$	$\lvert a\rvert^{-1}\mathcal{M}f(a^{-1}s)$
$f^{(k)}(t)$	$(-1)^k(s-k)_k\mathcal{M}f(s-k)$

where, $\forall x \in \mathbb{R}$ and $\forall k \geq 1$, $(x)_k$ stands for the so-called Pochhammer symbol, which is defined by

$$(x)_k = x \cdots (x-k+1) = \prod_{j=0}^{k-1}(x-j) \quad \text{if } k \geq 1,$$

and $(x)_0 = 1$, where x is in $\mathbb{R}$.

2.1 A Priori Estimates

Seeking continuity and observability inequalities for model (2) is then reduced to find lower and upper bounds for $\mathcal{M}G(\cdot)$ in suitable weighted Lebesgue's spaces. Doing so, one obtains

Theorem 2 (A Priori Estimates) *Let $k \in \mathbb{N} \cup \{0\}$ and $r \in \mathbb{R}$ be arbitrary. Assume that the Mellin transforms of ρ and $I_1[\rho]$ satisfy (3), then*

$$C_\ell^k \|\rho\|_{L^2_r} \leq \|(I_1[\rho])^{(k)}\|_{L^2_{2k+\frac{r-3}{2}}} \leq C_u^k \|\rho\|_{L^2_r},$$

where

$$C_\ell^k \overset{\text{(def)}}{=} \sqrt{2} \inf_{s \in \frac{r-1}{2}+i\,\mathbb{R}} \left|\left(\tfrac{s}{2}\right)_k \mathcal{M}G(s)\right|$$

$$C_u^k \overset{\text{(def)}}{=} \sqrt{2} \sup_{s \in \frac{r-1}{2}+i\,\mathbb{R}} \left|\left(\tfrac{s}{2}\right)_k \mathcal{M}G(s)\right|,$$

and $L_q^p = L^p\left([0,\infty), x^q\right)$ stands for the Lebesgue space with the weight x^q, $p \geq 1$, $q \in \mathbb{R}$.

Remark 1 It is worth noting that C_ℓ^k, C_u^k could *a priori* range from 0 to $+\infty$.

Proof Using the properties of the Mellin transform in Eq. (3), it follows that

$$(s-k)_k\ \mathcal{M}\mathrm{I}[\rho](s-k) = 2(s-k)_k\ \mathcal{M}G(2(s-k))\ \mathcal{M}\rho(2(s-k)+1) \tag{4}$$

Thanks to Parseval-Plancherel's isomorphism, for every s in $q + i\,\mathbb{R}$, we have

$$\begin{aligned} \left\| (\mathrm{I}[\rho])^{(k)} \right\|_{\mathrm{L}^2_{2q-1}} &= \frac{1}{\sqrt{(2\pi)}} \left\| (-1)^k (s-k)_k \, \mathcal{M}\mathrm{I}[\rho](s-k) \right\|_{\mathrm{L}^2(q+i\,\mathbb{R})} \\ &= \frac{2}{\sqrt{(2\pi)}} \left\| (s-k)_k \, \mathcal{M}G(2(s-k))\, \mathcal{M}\rho(2(s-k)+1) \right\|_{\mathrm{L}^2(q+i\,\mathbb{R})} \\ &= \frac{2}{\sqrt{(2\pi)}} \left\| (s)_k \, \mathcal{M}G(2s)\, \mathcal{M}\rho(2s+1) \right\|_{\mathrm{L}^2(q-k+i\,\mathbb{R})} \\ &= \frac{1}{\sqrt{\pi}} \left\| \left(\frac{s}{2}\right)_k \, \mathcal{M}G(s)\, \mathcal{M}\rho(s+1) \right\|_{\mathrm{L}^2(2(q-k)+i\,\mathbb{R})} \end{aligned} \tag{5}$$

As $\mathcal{M}$ is an isometry from $\mathrm{L}^2\,(2(q-k)+1+i\,\mathbb{R})$ on $\mathrm{L}^2_{4(q-k)+1}$,

$$\|\mathcal{M}\rho(s+1)\|_{\mathrm{L}^2(2(q-k)+i\,\mathbb{R})} = \|\mathcal{M}\rho(s)\|_{\mathrm{L}^2(2(q-k)+1+i\,\mathbb{R})} = \sqrt{2\pi}\, \|\rho\|_{\mathrm{L}^2_{4(q-k)+1}} \tag{6}$$

Thanks to (5) and (6) and the definitions of C_l^k, C_u^k, we get

$$C_l^k \, \|\rho\|_{\mathrm{L}^2_{4(q-k)+1}} \leq \left\| (\mathrm{I}[\rho])^{(k)} \right\|_{\mathrm{L}^2_{2q-1}} \leq C_u^k \, \|\rho\|_{\mathrm{L}^2_{4(q-k)+1}}$$

Taking $r = 4(q-k)+1$, that is $q = k + \frac{r-1}{4}$, provides the result.

Mellin Convolution

For two given functions f, g, the *multiplicative convolution* $f * g$ is defined as follows

$$(f * g)(x) = \int_0^\infty f(y)\, g\left(\frac{x}{y}\right) \frac{\mathrm{d}y}{y}$$

Theorem 3 (Mellin Transform of a Convolution) *Whenever this expression is well defined, we have*

$$\mathcal{M}(f * g)(s) = \mathcal{M}f(s)\, \mathcal{M}g(s)$$

Finally, the classical L^2-isometry has his Mellin counterpart.

Theorem 4 (Parseval-Plancherel's Isomorphism) *The Mellin transform can be extended in a unique manner to a linear isometry* (*up to the constant* $(2\pi)^{-1/2}$) *from* L^2_{2q-1} *onto the classical Lebesgue space* $\mathrm{L}^2(q + i\,\mathbb{R})$:

$$\mathcal{M} \in \mathcal{L}\left(\mathrm{L}^2_{2q-1}; \mathrm{L}^2(q + i\,\mathbb{R},\, \mathrm{d}x)\right)$$

3 Observability of CNG Channels

The *a priori* estimates in the theorem above also allow to determine a unique distribution of ion channels along the length of a cilium from measurements in time of the transmembrane electric current.

Theorem 5 (Existence and uniqueness of ρ) *Let* $a > 0$ *and* $r < 1$ *be given. If* $I_1 \in L^2\left([0,\infty), t^{\frac{r-3}{2}}\right)$, $I_1' \in L^2\left([0,\infty), t^{2+\frac{r-3}{2}}\right)$ *and* a *is small enough, then there exists a unique* $\rho \in L^2([0,\infty), x^r)$ *which satisfies the following stability condition:*

$$\|I_1\|_{L^2([0,\infty),t^{\frac{r-3}{2}})} + \|I_1'\|_{L^2\left([0,\infty),t^{2+\frac{r-3}{2}}\right)} \geq C\|\rho\|_{L^2_r},$$

where $C > 0$ *depends only on* a *and* r.

Proof The proof is based on the following technical lemmas and its corollaries.

Lemma 1 *Let* A *and* B *be two elements of* $[0,\infty]$, $k \in \cup\{0\}\mathbb{N}$ *be a nonnegative integer and* f *a function such that* $f^{(j)}$ *is in* $\mathrm{L}^1_j(A,B)$ *for every* $j = 0,\ldots,k$. *For every real number* t, *we have*

$$\int_A^B f(x)x^{it}\,\mathrm{d}x = \sum_{j=0}^{k-1}(-1)^j Q_j\left[x^{j+1}f^{(j)}(x)x^{it}\right]_A^B + (-1)^k Q_{k-1}\int_A^B x^k f^{(k)}(x)x^{it}\,\mathrm{d}x,$$

where $Q_j = Q_j(t) = \left(\prod_{l=0}^{j}(1 + l + it)\right)^{-1}$.

Proof We use induction on $k \in \mathbb{N}$. For $k = 0$, since $Q_{-1} = 1$, there is nothing to prove. We assume that the formula is true for an integer $k \in \mathbb{N}$. As $(k + 1 + it)Q_k = Q_{k-1}$, it remains to prove that

$$(k + 1 + it)\int_A^B x^k f^{(k)}(x)x^{it}\,\mathrm{d}x = \left[x^{k+1}f^{(k)}(x)x^{it}\right]_A^B - \int_A^B x^{k+1}f^{(k+1)}(x)x^{it}\,\mathrm{d}x$$

As $\frac{\mathrm{d}}{\mathrm{d}x}x^{it} = \frac{it}{x}x^{it}$, the previous relation follows by integration by parts. Indeed, we have

$$\begin{aligned} it\int_A^B x^k f^{(k)}(x)x^{it}\,\mathrm{d}x &= \int_A^B x^{k+1} f^{(k)}(x)(x^{it})'\,\mathrm{d}x \\ &= \left[x^{k+1} f^{(k)}(x)x^{it}\right]_A^B - (k+1)\int_A^B x^k f^{(k)}(x)x^{it}\,\mathrm{d}x \\ &\quad - \int_A^B x^{k+1} f^{(k+1)}(x)x^{it}\,\mathrm{d}x \end{aligned}$$

Corollary 1 *Let $f : [A, B] \to \mathbb{R}$ with $A, B \in [0, \infty]$ be a piecewise C^1 function. If f is non-negative, f' is non-positive, $f \in \mathrm{L}^1(A, B)$, $f' \in \mathrm{L}^1_1(A, B)$ and for all $t \in \mathbb{R}$: $[xf(x)x^{it}]_A^B = 0$, then*

$$\sqrt{1+t^2}\left|\int_A^B f(x)x^{it}\,\mathrm{d}x\right| \le \int_A^B f(x)\,\mathrm{d}x.$$

Proof From Lemma 1 with $k = 1$ one obtains

$$\forall t \in \mathbb{R}, \quad (1+it)\int_A^B f(x)x^{it}\,\mathrm{d}x = -\int_A^B xf'(x)x^{it}\,\mathrm{d}x$$

As $A, B \ge 0$ and $f' \le 0$, using this previous identity twice, for $t \ne 0$ and for $t = 0$, we get

$$\sqrt{1+t^2}\left|\int_A^B f(x)x^{it}\,\mathrm{d}x\right| \le \int_A^B \left|xf'(x)\right|\,\mathrm{d}x = \int_A^B f(x)\,\mathrm{d}x$$

Lemma 2 *Let $n, K > 0$, $q \in \mathbb{R}$ and $f = \frac{erfc^n}{erfc^n + K}$. There exists $x_q > 0$ such that the function $g_q : x \in [x_q, \infty) \mapsto f(x)\,x^{q-1}$ is decreasing. Let $\tilde{q} = \inf E_q$ where $E_q = \{c \ge 0 \mid g_q'(x) < 0\,\forall x \ge c\}$. The function $q \mapsto \tilde{q}$ is increasing and $\tilde{q} = (q/(2n))^{1/2} + o\left(q^{1/2}\right)$ as $q \to \infty$.*

Proof As $f > 0$, the inequality $g_q'(x) \le 0$ is equivalent to

$$\frac{f'(x)}{f(x)} \le -\frac{q-1}{x} \tag{7}$$

Let us compute $\frac{f'}{f}$. To do so, let $u = \mathrm{erfc}^n$, so that $f = \frac{u}{u+K}$. We have

$$\frac{f'}{f} = \frac{u'}{u}\frac{K}{u+K} = n\frac{\mathrm{erfc}'}{\mathrm{erfc}}\frac{K}{u+K} \tag{8}$$

Since $\mathrm{erfc}'(x) = -2\pi^{-1/2}e^{-x^2}$, for x large enough, $\mathrm{erfc}(x) = \pi^{-1/2}x^{-1}e^{-x^2} + o\left(x^{-1}e^{-x^2}\right)$, and so

$$\frac{f'(x)}{f(x)} = n\frac{\mathrm{erfc}'(x)}{\mathrm{erfc}(x)}(1 + o(1)) = -2nx + o(x) \tag{9}$$

This asymptotic expansion proves that the inequality (7) is satisfied for large enough values of x. As a consequence, for every q in $\mathbb{R}$, the set E_q is not empty, which justifies the definition of $\tilde{q}$. Note that the definition of $\tilde{q}$ implies $g_q'(\tilde{q}) = 0$, and hence, thanks to (7), $\frac{f'(\tilde{q})}{f(\tilde{q})} = -\frac{q-1}{\tilde{q}}$. Let $q_1 \geq q_2$ be two real numbers. In order to show that $\tilde{q_2} \leq \tilde{q_1}$, it is enough to prove that $g_{q_1}'(\tilde{q_2}) \geq 0$. This holds true because

$$\begin{aligned} g_{q_1}'(\tilde{q_2}) &= \tilde{q_2}^{q_1-2}(f'(\tilde{q_2})\tilde{q_2} + f(\tilde{q_2})(q_1-1)) \geq \tilde{q_2}^{q_1-2}(f'(\tilde{q_2})\tilde{q_2} + f(\tilde{q_2})(q_2-1)) \\ &= \tilde{q_2}^{q_1-q_2}g_{q_2}'(\tilde{q_2}) = 0 \end{aligned}$$

To find an expansion for $\tilde{q}$, let us recall the following classical lower bound on $\mathrm{erfc}(x)$ for $x \geq 0$,

$$\frac{1}{x + (x^2+2)^{1/2}} \leq \frac{1}{2}\pi^{1/2}\exp(x^2)\,\mathrm{erfc}(x)$$

As the function $u = \mathrm{erfc}^n$ takes its values in $(0, 1]$, $\frac{nK}{1+K} \leq \frac{nK}{u+K} \leq n$. Consequently, the identities (8) yield

$$-n\left(x + (x^2+2)^{1/2}\right) \leq \frac{f'(x)}{f(x)} \tag{10}$$

Let $q > 1$ and set $x_q = \frac{q-1}{(2n)^{1/2}(n+q-1)^{1/2}}$. The inequality $-\frac{q-1}{x} \leq -n\left(x + (x^2+2)^{1/2}\right)$ is equivalent to $x\left(x + (x^2+2)^{1/2}\right) \leq \frac{q-1}{n}$. A simple computation shows that this inequality is satisfied for $x = x_q$ (and becomes and equality). Thanks to (10), we conclude that x_q satisfies $\frac{f'(x_q)}{f(x_q)} \geq -\frac{q-1}{x_q}$, which leads to $\tilde{q} \geq x_q$, by definition of $\tilde{q}$ and by (7). This last inequality implies that $\tilde{q}$ tends to $+\infty$ as q tends to $+\infty$. Finally, from (9), we get the asymptotic for $\tilde{q}$, namely

$$-2n\tilde{q} + o(\tilde{q}) = \frac{f'(\tilde{q})}{f(\tilde{q})} = -\frac{q-1}{\tilde{q}}$$

This completes the proof of Lemma 2.

Proof of Theorem 5

We are now in a position to conclude the proof of Theorem 5. To do so, we begin by introducing

$$J(x) \overset{\text{(def)}}{=} \mathbb{H}(c_0 \text{erfc}(x)) = f(x)\, \mathbb{1}_{x \geq \alpha} + K\, \mathbb{1}_{0<x<\alpha},$$

where $f(x) = \frac{\text{erfc}(x)^n}{\text{erfc}(x)^n + c_0^{-n} K_{1/2}^n}$, $\alpha = \text{erfc}^{-1}\left(\frac{a}{c_0}\right)$, and $K = 1$. A brief calculation shows that G and J, and their corresponding Mellin transforms are related as follows

$$G(x) = J\left(\frac{1}{2\sqrt{D x}}\right) \quad \text{and} \quad \mathcal{M}G(s) = \frac{1}{2^s \sqrt{D^s}\mathcal{M}J(-s)} \tag{11}$$

Thus, in terms of J, Eq. (3) becomes

$$\mathcal{M}\rho(s+1) = \frac{2^{s-1}}{\sqrt{D^s}} \frac{\mathcal{M}\text{I}[\rho]\,(s/2)}{\mathcal{M}J(-s)} \tag{12}$$

From the estimate for erfc at $+\infty$, given in the proof of Lemma 2, the function J_1 is in L_k^1 for every $k > -1$. Thus $\mathcal{M}J_1$ is holomorphic on the right half-plane, see Proposition 1. Using Lemma 3.2 in [1] on the vertical line $\frac{1-r}{2} + i\,\mathbb{R}$ with $\frac{1-r}{2} > 0$, one deduces that bounds for $\mathcal{M}J(-s)$ amounts to estimate $|s\mathcal{M}J_1(s)|$, from above or from below, on the vertical lines $q + i\,\mathbb{R}$, for $q > 0$. The Mellin transform of J_1 at $s = q + it$ is given by

$$\mathcal{M}J_1(s) = K\int_0^{\alpha} x^{s-1}\,\text{d}x + c_0^n \int_{\alpha}^{+\infty} f(x)x^{s-1}\,\text{d}x = K\frac{\alpha^s}{s} + c_0^n \int_{\alpha}^{+\infty} f(x)x^{q-1}x^{it}\,\text{d}x$$

For any $a \geq 0$, $q > 0$ and $s \in q + i\,\mathbb{R}$ we have

$$|\mathcal{M}J_1(s)| \leq K\frac{\alpha^q}{q} + c_0^n \int_{\alpha}^{+\infty} f(x)x^{q-1}\,\text{d}x,$$

which is finite. Let $q > 0$. According to Lemma 2 the function $x \mapsto f(x)x^{q-1}$ is decreasing for $x \geq x_0$. Let $a < c_0\text{erfc}(x_0)$ so that $\alpha = \text{erfc}^{-1}(a/c_0) \geq x_0$. Let $g(x) = f(x)x^{q-1}\,\mathbb{1}_{x\geq\alpha}$. For every $t \in \mathbb{R}$, $\left[f(x)x^{it}\right]_{x_0}^{\infty} = 0$ because f vanishes for $x \leq \alpha$ and $x_0 \leq \alpha$, and $g(x) = \pi^{-n/2}x^{-n+q-1}e^{-nx^2} + o\left(x^{-n+q-1}e^{-nx^2}\right)$. Then Corollary 1 can be applied to the function g, with $A = \alpha$, $B = +\infty$, for $s \in q + i\,\mathbb{R}$, to give

$$\begin{aligned} |s\mathcal{M}J_1(s)| &\le K\left|\alpha^s\right| + c_0^n \frac{|s|}{\sqrt{1+t^2}}\sqrt{1+t^2}\left|\int_\alpha^\infty f(x)\, x^{s-1}\,\mathrm{d}x\right| \\ &\le K\alpha^q + c_0^n \max(1,q)\int_\alpha^\infty f(x)x^{q-1}\,\mathrm{d}x < \infty, \end{aligned}$$

because $\frac{|s|}{\sqrt{1+t^2}} \in [q,1] \cup [1,q]$, either $q \le 1$ or $q \ge 1$. For small values of a, the first term dominates the second one. The same calculation as above leads to

$$|s\mathcal{M}J_1(s)| \ge K\alpha^q - c_0^n \max(1,q)\int_\alpha^\infty f(x)x^{q-1}\,\mathrm{d}x$$

This latter expression is equivalent to $K\alpha^q$ as α tends to $+\infty$, therefore, it is positive for large values of α.

4 Unstable Identifiability, Non Existence of Observability Inequalities

Since the French-Groetsch model is also a Fredholm integral equation of the first kind, it is natural to apply a Mellin transform here too. This leads to interesting results: neither an observability inequality nor a proper numerical algorithm for recovering ρ can be established. However, an Identifiability result holds whenever the current is measured over an open time interval (see the Identifiability Theorem below).

Defining $\tilde{G}$ as

$$\tilde{G}(z) = \mathbb{P}\left(c_0 \mathrm{erfc}\left(\frac{1}{2\sqrt{Dz}}\right)\right),$$

and rescaling time t in t^2, we obtain a convolution equation very similar to (3):

$$\mathcal{M}\rho(s+1) = \frac{1}{2}\frac{\mathcal{M}\mathrm{I}_0[\rho]\,(s/2)}{\mathcal{M}\tilde{G}(s)} \tag{13}$$

A close study of the transform of $\tilde{G}(s)$ allows us to establish the following two theorems, which provide information about the behavior of the inverse problem associated with model (1). The proof of Theorems 3 and 4 requires to extend Mellin transform to functions in the Schwartz space and to prove that the Mellin transforms of such smooth and rapidly decreasing functions decay faster than polynomials on vertical lines.[1]

[1] The interested reader is referred to [1] for details on how this can be done and and for detailed proofs of Theorems 3 and 4.

Theorem 3 (Non Observability) *Let $r < 1$ be fixed. For every non-negative integer k there exists no constant $C_k > 0$ such that the observability inequality:*

$$\|(I_0[\rho])^{(k)}\|_{L^2\left([0,\infty),t^{2k+\frac{r-3}{2}}\right)} \geq C_k\|\rho\|_{L^2_r},$$

holds for every function $\rho \in L^2([0,\infty), x^r)$.

Note that this result shows that $\mathrm{I}_0 \in \mathcal{L}(L^2_r; L^2_{\frac{r-3}{2}})$, and that if the inverse problem were identificable (i.e., I_0 were injective), then I_0^{-1} could not be continuous.

Theorem 4 (Identifiability) *Let $r < 0$ and $\rho \in L^1([0,\infty), x^r)$ be arbitrary. If there exists a nonempty open subset $\mathcal{U}$ of $(0,\infty)$ such that for all $t \in \mathcal{U}$, $I_0[\rho](t) = 0$, then $\rho = 0$ almost everywhere on $(0,\infty)$.*

The interested reader is referred to [1, §4 and §5] for various numerical experiences associated with the different theoretical results of this paper. In particular, Theorems 5 and 6 are graphically illustrated in the quoted reference with data extracted from laboratory experiments carried out by Chen et al. [2] in the 1990s.

A Path Forward

The Mellin transform has been successful in mathematically analyzing models (1) and (2), allowing us to answer questions of existence (observability), uniqueness and identifiability of the distribution of ion channels along a cilium, as well as stability issues associated with both direct and inverse problems in these models. However, from a more holistic scientific point of view, not a purely mathematical one, the big question does not seem to be exactly this. Rather, it is about whether, by using and studying these models, Mathematics truly helps to improve our understanding of the olfactory system and, in general terms, the real world. In this sense, Kleene's experiments have been a great contribution, albeit insufficient. Much stronger validation of the models is required, which can only be achieved by forming multidisciplinary teams and designing ad-hoc experiments.

Acknowledgements C. C. is partially supported by PFBasal-001 and AFBasal170001 projects, and from the Regional Program STIC-AmSud Project NEMBICA-20-STIC-05.

References

1. Bourgeron, T., Conca, C., Lecaros, R.: Determining the distribution of ion channels from experimental data. Math. Mod. Numer. Anal. (ESAIM: M^{2}AN) **52**, 2083–2107 (2018)
2. Chen, C., Nakamura, T., Koutalos, Y.: Cyclic AMP diffusion coefficient in frog olfactory cilia. Biophys. J. **76**, 2861–2867 (1999)
3. Conca, C., Lecaros, R., Ortega, J.H., Rosier, L.: Determination of the calcium channel distribution in the olfactory system. J. Inverse Ill Posed Probl. **22**, 671–711 (2014)
4. French, D.A., Flannery, R.J., Groetsch, C.W., Krantz, W.B., Kleene, S.J.: Numerical approximation of solutions of a nonlinear inverse problem arising in olfaction experimentation. Math. Comput. Model. **43**, 945–956 (2006)

5. Kleene, S.J.: Origin of the chloride current in olfactory transduction. Neuron **11**, 123–132 (1993)
6. Kleene, S.J., Gesteland, R.C.: Transmembrane currents in frog olfactory cilia. J. Membr. Biol. **120**, 75–81 (1991)
7. Lindelöf, E.: Robert Hjalmar Mellin. Acta Math. **61**, i–vi (1933)

State Estimation—The Role of Reduced Models

Albert Cohen, Wolfgang Dahmen, and Ron DeVore

Abstract The exploration of complex physical or technological processes usually requires exploiting available information from different sources: (i) *physical laws* often represented as a family of *parameter dependent partial differential equations* and (ii) *data* provided by *measurement devices* or *sensors*. The amount of sensors is typically limited and data acquisition may be expensive and in some cases even harmful. This article reviews some recent developments for this "small-data" scenario where inversion is strongly aggravated by the typically large parametric dimensionality. The proposed concepts may be viewed as exploring alternatives to *Bayesian inversion* in favor of more deterministic accuracy quantification related to the required computational complexity. We discuss *optimality criteria* which delineate intrinsic information limits, and highlight the role of *reduced models* for developing efficient computational strategies. In particular, the need to *adapt* the reduced models—not to a specific (possibly noisy) data set but rather to the sensor system—is a central theme. This, in turn, is facilitated by exploiting *geometric* perspectives based on proper *stable variational formulations* of the continuous model.

1 Introduction

Modern sensor technology and data acquisition capabilities generate an ever increasing wealth of data about virtually every branch of science and social life. Machine learning offers novel techniques for extracting quantifiable information from such large data sets. While machine learning has already had a transformative impact on

A. Cohen
Laboratoire Jacques-Louis Lions, Sorbonne Université, 4, Place Jussieu, 75005 Paris, France
e-mail: cohen@ann.jussieu.fr

W. Dahmen (✉)
Mathematics Department, University of South Carolina, 1523 Greene Street, Columbia, SC 29208, USA
e-mail: dahmen@math.sc.edu

R. DeVore
Department of Mathematics, Texas A & M University, College Station, TX 77843-3368, USA

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_4

a diversity of application areas in the "big-data" regime, particularly in image classification and artificial intelligence, it is yet to have a similar impact in many other areas of science.

Utilizing data observations in the analysis of scientific processes differs from traditional learning in that one has the additional information that these processes are described by mathematical models—systems of partial differential equations (PDE) or integral equations—that encode the physical laws that govern the process. Such models, however, are often deficient, inaccurate, incomplete or need to be further calibrated by determining a large number of *parameters* in order to accurately represent an observed process. Typical guiding examples are Darcy's equation for the pressure in ground-water flow or electron impedance tomography. Both are based on second order elliptic equations as core models. The diffusion coefficients in these examples describe premeability or conductivity, respectively. The parametric representations of the coefficients could arise, for instance, from Karhunen-Loève expansions of a random field that represent "unresolvable" features to be captured by the model. In this case the number of parameters could actually be *infinite*.

The use of machine learning to describe complex states of interest or even the underlying laws, solely through data, seems to bear little hope. In fact, data acquisition is often expensive or even harmful as in applications involving radiation. Thus, a severe undersampling poses principal obstructions to *state* or *parameter estimation* by solely processing observational data through standard machine learning techniques. It is therefore more natural to try to effectively combine the data information with the knowledge of the underlying physical laws represented by parameter dependent families of PDEs.

Methods that fuse together data-driven and model-based approaches fall roughly into two categories. One prototype of a *data assimilation* scenario arises in meteorology where data are used to stabilize otherwise chaotic dynamical systems, typically with the aid of (stochastic) filtering techniques. A second setting, in line with the above examples, uses an underlying *stable* continuous model to *regularize* otherwise ill-posed estimation tasks in a "small-data" scenario. *Bayesian inversion* is a prominent way of regularizing such problems. It relaxes the estimation task to asking only for *posterior probabilities* of states or parameters to explain given observations.

The present article reviews some recent developments on data driven state and parameter estimation that can be viewed as seeking alternatives to Bayesian inversion by placing a stronger focus on deterministic uncertainty quantification and its relation to *computational complexity*. The emphasis is on foundational aspects such as the optimality of algorithms (formulated in an appropriate sense) when treating estimation tasks for "small-data" problems in *high-dimensional parameter* regimes. Central issues concern the role of *reduced modeling* and the exploitation of intrinsic problem metrics provided by the *variational formulation* of the underlying continuous family of PDEs. This is used by the so called *Parametrized Background Data-Weak* (PBDW) framework, introduced in [20] and further analyzed in [4], to identify a suitable trial (Hilbert) space $\mathbb{U}$ that accommodates the states and eventually also the data. An important point is to distinguish between the *data* and corresponding *sensors*—here linear functionals in the dual $\mathbb{U}'$ of $\mathbb{U}$—from which the data are generated. This will be seen to actually open a *geometric* perspective that sheds light on intrinsic estimation limits. Moreover, in the deterministic setting, a pivotal role is played by the so called *solution manifold*, which is the set of all states that can be attained when the parameters in the PDE traverse the whole parameter domain.

Even with full knowledge of a state in the solution manifold, to infer from it a corresponding parameter is a *nonlinear severely ill-posed* problem typically formulated as a *non-convex* optimization problem. On the other hand, state estimation from data is a *linear*, and hence a more benign inversion task mainly suffering under the current premises from a severe undersampling. We will, however, indicate how to reduce, under certain circumstances, the latter to the former problem so as to end up with a *convex* optimization problem. This motivates focusing in what follows mainly on state estimation. A central question then becomes how to best invoke knowledge on the solution manifold to regularize the estimation problem without introducing unnecessarily ambiguous bias. Our principal viewpoint is to recast state estimation as an *optimal recovery* problem which then naturally leads one to explore the role and potential of *reduced modeling*.

The layout of the paper is as follows. Section 2 describes the conceptual framework for *state estimation* as an *optimal recovery task*. This formulation allows the identification of lower bounds for the best achievable recovery accuracy.

Section 3 reviews recent developments concerning a certain *affine* recovery scheme and highlights the role of *reduced models* adapted to the recovery task. The overarching theme is to establish certified recovery bounds. When striving for optimality of such affine recovery maps, high parameter dimensionality is identified as a major challenge. We outline a recent remedy that avoids the *Curse of Dimensionality* by trading deterministic accuracy guarantees against analogs that hold with quantifiable high probability.

Even optimal affine reduced models can, in general, not be expected to realize the benchmarks identified in Sect. 2. To put the results in Sect. 3 in proper perspective, we comment in Sect. 4 on ongoing work that uses the results on affine reduced models and corresponding estimators as a central building block for nonlinear estimators. We also indicate briefly some ramifications on parameter estimation.

2 Models and Data

2.1 The Model

Technological design or simulating physical processes is often based on continuum models given by a family

$$\mathcal{R}(u, y) = 0, \quad y \in \mathcal{Y}, \tag{2.1}$$

of partial differential Equations (PDEs) that depend on parameters y ranging over a parameter domain $\mathcal{Y} \subset \mathbb{R}^{d_y}$. We will always assume *uniform well-posedness* of

(2.1): for each $y \in \mathcal{Y}$, there exists a unique solution $u = u(y)$ in some trial Hilbert space $\mathbb{U}$ which satisfies $\mathcal{R}(u(y), y) = 0$.

Specifically, we consider only linear problems of the form $\mathcal{B}_y u = f$, that is,

$$\mathcal{R}(u, y) = f - \mathcal{B}_y u. \tag{2.2}$$

Here f belongs to the dual $\mathbb{V}'$ of a suitable *test space* $\mathbb{V}$ and $\mathcal{B}_y$ is a linear operator acting from $\mathbb{U}$ to $\mathbb{V}'$ that depends on $y \in \mathcal{Y}$. Here, uniform well-posedness means then that $\mathcal{B}_y$ is boundedly invertible with bounds independent of y. By the Babuška-Banach-Nečas Theorem, this is equivalent to saying that the bilinear form

$$(u, v) \mapsto b_y(u, v) := (\mathcal{B}_y u)(v) \tag{2.3}$$

satisfies the following *continuity* and *inf-sup conditions*

$$\sup_{u \in \mathbb{U}} \sup_{v \in \mathbb{V}} \frac{b_y(u, v)}{\|u\|_{\mathbb{U}} \|v\|_{\mathbb{V}}} \le C_b \quad \text{and} \quad \inf_{u \in \mathbb{U}} \sup_{v \in \mathbb{V}} \frac{b_y(u, v)}{\|u\|_{\mathbb{U}} \|v\|_{\mathbb{V}}} \ge c_b > 0, \quad y \in \mathcal{Y}, \tag{2.4}$$

together with the property that $b_y(u, v) = 0,\ u \in \mathbb{U}$, implies $v = 0$ (injectivity of $\mathcal{B}_y^*$). The relevance of this stability notion lies in the entailed validity of the *error-residual relation*

$$C_b^{-1} \|f - \mathcal{B}_y v\|_{\mathbb{V}'} \le \|u(y) - v\|_{\mathbb{U}} \le c_b^{-1} \|f - \mathcal{B}_y v\|_{\mathbb{V}'}, \quad v \in \mathbb{U},\ y \in \mathcal{Y}, \tag{2.5}$$

where $\|g\|_{\mathbb{V}'} := \sup\{g(v) \ : \ \|v\|_{\mathbb{V}} = 1\}$. Thus, errors in the trial norm are equivalent to residuals in the dual test norm which will be exploited in what follows.

For a wide range of problems such as space-time variational formulations, e.g. of parabolic or convection-diffusion problems, indefinite or singularly perturbed problems, the identification of a suitable pair $\mathbb{U}, \mathbb{V}$ that guarantees stability in the above sense is not entirely straightforward. In particular, trial and test space may have to differ from each other, see e.g. [6, 11, 17, 23] for examples as well as some general principles.

The simplest example, used for illustration purposes, is the *elliptic* family

$$\mathcal{R}(u, y) = f + \text{ div } (a(y)\nabla u), \tag{2.6}$$

set in $\Omega \subset \mathbb{R}^{d_x}$ where $d_x \in \{1, 2, 3\}$, with boundary conditions $u|_{\partial\Omega} = 0$. Uniform well-posedness follows then for $\mathbb{U} = \mathbb{V} = H_0^1(\Omega)$ if we have for some fixed constants $0 < r \le R < \infty$ the bounds

$$r \le a(x, y) \le R, \quad (x, y) \in \Omega \times \mathcal{Y}, \tag{2.7}$$

readily implying (2.4).

Aside from well-posedness, a second important structural property of the model (2.1) is *affine parameter dependence*. By this we mean that

$$\mathcal{B}_y u = \mathcal{B}_0 u + \sum_{j=1}^{d_y} y_j \mathcal{B}_j u, \quad y = (y_j)_{j=1,\dots,d_y} \in \mathcal{Y}, \tag{2.8}$$

where the operators $\mathcal{B}_j : \mathbb{U} \to \mathbb{V}'$ are *independent* of y. In turn, the residual has a similar affine dependence structure

$$\mathcal{R}(u, y) = \mathcal{R}_0(u) + \sum_{j=1}^{d_y} y_j \mathcal{R}_j u, \quad \mathcal{R}_0(u) := f - \mathcal{B}_0 u, \quad \mathcal{R}_j = -\mathcal{B}_j. \tag{2.9}$$

For the example (2.6) such a structure is encountered for *affine* parametric representations of the diffusion coefficients

$$a(x, y) = a_0(x) + \sum_{j=1}^{d_y} y_j \theta_j(x), \quad (x, y) \in \Omega \times \mathcal{Y}, \tag{2.10}$$

i.e., the field a is expanded in terms of some given spatial basis functions θ_j. As indicated earlier, the pressure equation in Darcy's law for porous media flow is an example for (2.6) where the diffusion coefficient $a(y)$ of the form (2.10) may arise from a stochastic model for permeability via a Karhunen-Loève expansion. In this case (upon proper normalization) $y \in [-1, 1]^{\mathbb{N}}$ has, in principle, *infinitely* many entries, that is $d_y = \infty$. However, due to (2.7), the θ_j should then have some decay as $j \to \infty$ which means that the parameters become less and less important when j increases. Another example is electron impedance tomography involving the same type of elliptic operator where parametric expansions represent possible variations of conductivity often modeled as piecewise constants, i.e., the θ_j could be characteristic functions subordinate to a partition of Ω. In this case data are acquired through sensors that act through trace functionals greatly adding to ill-posedness.

A central role in the subsequent discussion is played by the solution manifold

$$\mathcal{M} = u(\mathcal{Y}) := \{u(y) \ : \ y \in \mathcal{Y}\} \tag{2.11}$$

which is then the range of the *parameter-to-solution map* $u : y \mapsto u(y)$ comprised of all states that can be attained when y traverses $\mathcal{Y}$. Without further mention, $\mathcal{M}$ will be assumed to be compact which actually follows under standard assumptions met in all above mentioned examples.

Estimating states in $\mathcal{M}$ or corresponding parameters from measurements requires the efficient approximation of elements in $\mathcal{M}$. A common challenge encountered in all such models lies in the inherent *high-dimensionality* of the states $u = u(\cdot, y)$ as functions of d_x spatial variables $x \in \Omega$ and $d_y \gg 1$ parametric variables $y \in \mathcal{Y}$. In particular, when $d_y = \infty$ any calculation, of course, has to work with finitely many "activated" parameters whose number, however, has to be coordinated with the spatial resolution of a numerical scheme to retain *model-consistency*. It is especially this issue that hinders standard approaches based on *first discretizing* the parametric model because rigorously balancing spatial and parametric uncertainties becomes then difficult.

What renders such problem scenarios nevertheless numerically tractable is a further property that will be implicitly assumed in what follows, namely that the *Kolmogorov n-widths* of the solution manifold

$$d_n(\mathcal{M})_{\mathbb{U}} := \inf_{\dim \mathbb{U}_n = n} \sup_{u \in \mathcal{M}} \inf_{v \in \mathbb{U}_n} \|u - v\|_{\mathbb{U}} \tag{2.12}$$

exhibits at least some algebraic decay

$$d_n(\mathcal{M})_{\mathbb{U}} \lesssim n^{-s} \tag{2.13}$$

for some $s > 0$, see [13] for a comprehensive account.

For instance, this is known to be the case for elliptic models (2.6) with (2.7), as a consequence of the results of sparse polynomial approximation of the parameter to solution map $y \mapsto u(y)$ established e.g. in [15]. More generally, (2.13) can be established under a general holomorphy property of the parameter to solution map, as a consequence of a similar algebraic decay assumed on the n-widths of the parameter set, see [14]. For a fixed finite number $d_y < \infty$ of parameters, under certain structural assumptions on the parameter representations (e.g. piecewise constants on checkerboard partitions) one can even establish (sub-) exponential decay rates, see [2] for more details. Assuming s in (2.13) to have a "substantial" size for any range of d_y, is therefore justified.

In summary, the results discussed below are valid and practically feasible for well posed linear models (2.4) with affine parameter dependence (2.9) whose solution manifolds have rapidly decaying n-widths (2.13).

2.2 *The Data*

Suppose we are given data $\mathbf{w} = (w_1, \dots, w_m)^\top \in \mathbb{R}^m$ representing observations of an unknown state $u \in \mathbb{U}$ obtained through m linearly independent linear functionals $\ell_i \in \mathbb{U}'$, i.e.,

$$w_i = \ell_i(u), \quad i = 1, \dots, m. \tag{2.14}$$

Since in real applications data acquisition may be costly or harmful we assume that m is *fixed*. The central task to be discussed in what follows is to recover from this information an estimate for the observed unknown state u, based on the prior assumption that u belongs to $\mathcal{M}$ or is close to $\mathcal{M}$. Moreover, to bring out the essence of this estimation task we assume for the moment that the data are noise-free.

Following [4, 20], we first recast the data in a "compliant" metric, by introducing the Riesz representers $\psi_i \in \mathbb{U}$, defined by

$$(\psi_i, v)_{\mathbb{U}} = \ell_i(v), \quad v \in \mathbb{U}, \quad i = 1, \dots, m,$$

The ψ_i now span the m-dimensional subspace $\mathbb{W} \subset \mathbb{U}$ which we refer to as *measurement space*, and the information carried by the $\ell_i(u)$ is equivalent to that of the orthogonal projection $P_{\mathbb{W}}u$ of u to $\mathbb{W}$. The decomposition

$$u = P_{\mathbb{W}}u + P_{\mathbb{W}^\perp}u, \quad u \in \mathbb{U}, \tag{2.15}$$

thus contains a first term that is “seen” by the sensors and a second (infinite-dimensional) term which cannot be detected. The decomposition (2.15) may be seen as a sensor-induced “coordinate system” thereby opening up a *geometric perspective* that will prove very useful in what follows. State estimation can then be viewed as learning from samples $w := P_{\mathbb{W}}u$ the unknown “labels” $P_{\mathbb{W}^\perp}u \in \mathbb{W}^\perp$.

In this article, we are interested in how well we can approximate u from the information that $u \in \mathcal{M}$ and $P_{\mathbb{W}}u = w$ with w given to us. Any such approximation is given by a mapping $A : w \to A(w) \in \mathbb{U}$. The overall performance of recovery on all of $\mathcal{M}$ by the mapping A is typically measured in the worst case setting, that is,

$$E_{\text{wc}}(A, \mathcal{M}, \mathbb{W}) = \sup_{u \in \mathcal{M}} \|u - A(P_{\mathbb{W}}u)\|_{\mathbb{U}}. \tag{2.16}$$

The optimal recovery error on $\mathcal{M}$ is then defined as

$$E_{\text{wc}}(\mathcal{M}, \mathbb{W}) := \inf_A E_{\text{wc}}(A, \mathcal{M}, \mathbb{W}), \tag{2.17}$$

where the infimum is over all possible recovery maps. Let us observe that the construction of recovery maps can be restricted to be of the form

$$A : w \to A(w), \quad A(w) = w + B(w), \quad \text{with } B : \mathbb{W} \to \mathbb{W}^\perp. \tag{2.18}$$

Indeed, given any recovery mapping A, we can write $A(w) = P_{\mathbb{W}}A(w) + P_{\mathbb{W}^\perp}A(w)$ and the performance of the recovery can only be improved if we replace the first term by w. In other words, $A(w)$ should belong to the affine space

$$\mathbb{U}_w := w + \mathbb{W}^\perp, \tag{2.19}$$

that contains u. The mappings B are commonly referred to as liftings into $\mathbb{W}^\perp$.

2.3 Optimality Criteria and Numerical Recovery

Finding a best recovery map A attaining (2.17) is known as *optimal recovery*. The best mapping has a well-known simple theoretical description, see e.g. [21], that we now describe. Note first that a precise recovery of the unknown state u from the given information is generally impossible. Indeed, the best we can say about u is that it lies in the *manifold slice*

$$\mathcal{M}_w := \{u \in \mathcal{M} : P_{\mathbb{W}} u = w\} = \mathcal{M} \cap \mathbb{U}_w, \tag{2.20}$$

which is comprised of all elements in $\mathcal{M}$ sharing the same measurement $w \in \mathbb{W}$. The Chebyshev ball $B(\mathcal{M}_w)$ is the smallest ball in $\mathbb{U}$ that contains $\mathcal{M}_w$. The best recovery algorithm is then given by the mapping

$$A^*(w) := \mathrm{cen}(\mathcal{M}_w), \tag{2.21}$$

that assigns to each $w \in \mathcal{M}$ the center $\mathrm{cen}(\mathcal{M}_w)$ of $B(\mathcal{M}_w)$, called the Chebyshev center of $\mathcal{M}_w$. Then, the radius $\mathrm{rad}(\mathcal{M}_w)$ of $B(\mathcal{M}_w)$ is the best worst case error over the class $\mathcal{M}_w$. The best worst case error over $\mathcal{M}$, which is achieved by A^*, is thus given by

$$E_{\mathrm{wc}}(\mathcal{M}, W) = E_{\mathrm{wc}}(A^*, \mathcal{M}, \mathbb{W}) = \max_{w \in \mathbb{W}} \mathrm{rad}(\mathcal{M}_w). \tag{2.22}$$

While the above mapping A^* gives a nice theoretical description of the optimal recovery algorithm, it is typically not numerically implementable since the Chebyshev center $\mathrm{cen}(\mathcal{M}_w)$ is not easily found. Moreover, such an optimal algorithm is highly nonlinear and possibly discontinuous. The purpose of this section is to formulate a more modest goal for the performance of a recovery algorithm with the hope that this more modest goal can be met with a numerically realizable algorithm. The remaining sections of the paper introduce numerically implementable recovery mappings, analyze their performance, and evaluate the numerical cost in constructing these mappings.

The search for a numerically realizable algorithm must out of necessity lessen the performance criteria. A first possibility is to weaken the performance criteria to *near best* algorithms. This means that we search for an algorithm A such that

$$E_{\mathrm{wc}}(A, \mathcal{M}, \mathbb{W}) \le C_0 E_{\mathrm{wc}}(\mathcal{M}, \mathbb{W}), \tag{2.23}$$

with a reasonable value of $C_0 > 1$. For example, any mapping A which takes w into an element in the Chebyshev ball of $\mathcal{M}_w$ is near best with constant $C_0 = 2$. However, finding near best mappings A also seems to be numerically out of reach.

In order to formulate a more attainable performance criterion, we return to our earlier observations about uncertainty in both the model class $\mathcal{M}$ and in the measurements w. The former is a modeling error while the latter is an inherent measurement error. Both of these uncertainties can be quantified by introducing for each $\varepsilon > 0$, the ε-neighborhood of the manifold

$$\mathcal{M}^\varepsilon := \{v \in \mathbb{U} : \mathrm{dist}\,(v, \mathcal{M})_{\mathbb{U}} \le \varepsilon\}. \tag{2.24}$$

The uncertainty in the model can be thought of as saying the sought after u is in $\mathcal{M}^\varepsilon$ rather than $u \in \mathcal{M}$. Also, we may formulate uncertainty (noise) in the measurements as saying that they are not measurements of a $u \in \mathcal{M}$ but rather some $u \in \mathcal{M}^\varepsilon$. Here the value of ε quantifies these uncertainties.

Our new goal is to numerically construct a recovery map A that is near-optimal on $\mathcal{M}^\varepsilon$, for some given $\varepsilon > 0$. Let us note that $\mathcal{M}^\varepsilon$ is not compact. An algorithm A is worst-case near optimal for $\mathcal{M}^\varepsilon$ if and only if its performance is bounded by a constant multiple of the diameter

$$\delta_\varepsilon(\mathcal{M}, \mathbb{W}) := \max\{\|u - v\|_{\mathbb{U}} : u, v \in \mathcal{M}^\varepsilon,\ P_{\mathbb{W}}(u - v) = 0\}. \tag{2.25}$$

Notice that $\varepsilon = 0$ gives the performance criterion for near optimal recovery over $\mathcal{M}$. One can show that the function $\varepsilon \mapsto \delta_\varepsilon(\mathcal{M}, \mathbb{W})$ is monotone non-decreasing in ε, continuous from the right, and $\lim_{\varepsilon\to 0^+} \delta_\varepsilon(\mathcal{M}, \mathbb{W}) = \delta_0(\mathcal{M}, \mathbb{W})$. The speed at which $\delta_\varepsilon(\mathcal{M}, \mathbb{W})$ approaches $\delta_0(\mathcal{M}, \mathbb{W})$ reflects the "condition" of the estimation problem depending on $\mathcal{M}$ and $\mathbb{W}$. While the practical realization of worst-case near-optimality for $\mathcal{M}^\varepsilon$ is already a challenge, quantifying corresponding computational cost would require assumptions on the condition of the problem.

One central theme, guiding subsequent discussions, is therefore to find recovery maps A_ε that realize an error bound of the form

$$E_{wc}(A_\varepsilon, \mathcal{M}, \mathbb{W}) \le C_0 \delta_\varepsilon(\mathcal{M}, \mathbb{W}). \tag{2.26}$$

Any a priori information on measurement accuracy and model bias might be used to choose a viable tolerance ε.

High parametric dimensionality poses particular challenges to estimation tasks when the targeted error bounds are in the above worst case sense. These challenges can be somewhat mitigated when adopting a Bayesian point of view [24]. The prior information on u is then described by a probability distribution p on $\mathbb{U}$, which is supported on $\mathcal{M}$. Such a measure is typically induced by a probability distribution on $\mathcal{Y}$ that may or may not be known. In the latter case, sampling $\mathcal{M}$, i.e., computing snapshots $u(y^i)$, $i = 1, \dots, N$, for i.i.d. samples $y^i \in \mathcal{Y}$, provides labeled data $(w_i, w_i^\perp) = (P_{\mathbb{W}} u(y^i), P_{\mathbb{W}^\perp} u(y^i))$ according to the sensor-based decomposition (2.15). This puts us into the setting of *regression* in machine learning asking for an estimator that predicts for any new measurement $w \in \mathbb{W}$ its lifting $w^\perp = B(w)$. It is then natural to measure the performance of an algorithm in an averaged sense. The best estimator A that minimizes the mean-square risk

$$E_{\rm ms}(A, p, \mathbb{W}) = \mathbb{E}(\|u - A(P_{\mathbb{W}} u)\|^2) = \int_{\mathbb{U}} \|u - A(P_{\mathbb{W}} u)\|^2 dp(u) \tag{2.27}$$

is given by the conditional expectation

$$A(w) = \mathbb{E}(u | P_{\mathbb{W}} u = w). \tag{2.28}$$

Since always $E_{\rm ms}(A, p, \mathbb{W}) \le E_{\rm wc}(A, \mathcal{M}, \mathbb{W})$, the optimality benchmarks are somewhat weaker. In the rest of this paper, we adhere to the worst case error in the deterministic setting that only assumes membership of u to $\mathcal{M}$ or $\mathcal{M}^\varepsilon$.

The following section is concerned with an important *building block* on a pathway towards achieving (2.26) at quantifiable computational cost. This building block, referred to as *one-space method* is a linear (affine) scheme which is, in principle, simple and easy to numerically implement. It depends on suitably chosen subspaces. We highlight the *regularizing property* of these subspaces as well as ways to *optimize* them. This will reveal certain intrinsic obstructions caused by *parameter dimensionality*. The one-space method by itself will generally not achieve (2.26) but, as indicated earlier, can be used as a building block in a *nonlinear* recovery scheme that may indeed meet the goal (2.26).

3 The One-Space Method

3.1 Subspace Regularization

The one space method can be viewed as a simple regularizer for state estimation. The resulting recovery map is induced by an n-dimensional subspace $\mathbb{U}_n$ of $\mathbb{U}$ for $n \le m$. Assume that, for each $n \ge 0$, we are given a subspace $\mathbb{U}_n \subset \mathbb{U}$ of dimension n whose distance from $\mathcal{M}$ can be assessed

$$\operatorname{dist}(\mathcal{M}, \mathbb{U}_n)_{\mathbb{U}} := \max_{u\in\mathcal{M}} \operatorname{dist}(u, \mathbb{U}_n)_{\mathbb{U}} \le \varepsilon_n. \tag{3.1}$$

Then the cylinder

$$\mathcal{K}(\mathbb{U}_n, \varepsilon_n) := \{u \in \mathbb{U} : \operatorname{dist}(u, \mathbb{U}_n)_{\mathbb{U}} \le \varepsilon_n\} \tag{3.2}$$

contains $\mathcal{M}$ and likewise the cylinder $\mathcal{K}(\mathbb{U}_n, \varepsilon_n + \varepsilon)$ contains $\mathcal{M}^\varepsilon$. Our prior assumption that the observed state belongs to $\mathcal{M}$ or $\mathcal{M}^\varepsilon$ can then be relaxed by assuming membership to these larger but simpler sets.

Remarkably, one can now realize an optimal recovery map quite easily that meets the relaxed benchmark $E_{\rm wc}(\mathcal{K}(\mathbb{U}_n, \varepsilon_n), \mathbb{W})$: in [4] it was shown that the Chebyshev center of the slice

$$\mathcal{K}_w(U_n, \varepsilon_n) := \mathcal{K}(\mathbb{U}_n, \varepsilon_n) \cap \mathbb{U}_w, \tag{3.3}$$

is exactly given by the state in $\mathbb{U}_w$ that is closest to $\mathbb{U}_n$, that is

$$u^* = u^*(w) := \operatorname*{argmin}_{u\in\mathbb{U}_w} \|u - P_{\mathbb{U}_n} u\|_{\mathbb{U}}. \tag{3.4}$$

This minimizer exists and can be shown to be unique as long as $\mathbb{U}_n \cap \mathbb{W}^\perp = \{0\}$. The corresponding optimal recovery map

$$A_{\mathbb{U}_n} : w \mapsto u^*(w) \tag{3.5}$$

was first introduced in [20] as the Parametrized Background Data Weak (PBDW) algorithm, and is referred to as the *one-space* method in [4]. Due to its above minimizing property, it is readily checked that this map is linear and can be determined with the aid of the singular value decomposition of the cross-Gramian between any pair of orthonormal basis for $\mathbb{U}_n$ and $\mathbb{W}$.

The worst case error $E_{\mathrm{wc}}(\mathcal{K}(\mathbb{U}_n, \varepsilon_n), \mathbb{W})$ can be described more precisely by introducing

$$\mu(\mathbb{U}_n, \mathbb{W}) := \sup_{v \in \mathbb{U}_n} \frac{\|v\|_{\mathbb{U}}}{\|P_{\mathbb{W}} v\|_{\mathbb{U}}} \tag{3.6}$$

which is finite if and only if $\mathbb{U}_n \cap \mathbb{W}^\perp = \{0\}$. This quantity, also introduced in a related but slightly different context in [1], is therefore related to the angle between the spaces $\mathbb{U}_n$ and $\mathbb{W}$. It becomes large when $\mathbb{U}_n$ contains elements that are nearly perpendicular to $\mathbb{W}$. It is actually computable: one has $\mu(\mathbb{U}_n, \mathbb{W}) = \beta(\mathbb{U}_n, \mathbb{W})^{-1}$ where

$$\beta(\mathbb{U}_n, \mathbb{W}) := \inf_{v \in \mathbb{U}_n} \sup_{w \in \mathbb{W}} \frac{\langle v, w\rangle_{\mathbb{U}}}{\|v\|_{\mathbb{U}} \|w\|_{\mathbb{U}}}, \tag{3.7}$$

and $\beta(\mathbb{U}_n, \mathbb{W})$ is the smallest singular value of the cross-Gramian between any pair of orthonormal bases for $\mathbb{W}$ and $\mathbb{U}_n$. It has been shown in [4, 20] that the worst case error bound over $\mathcal{K}(\mathbb{U}_n, \varepsilon_n)$ is given by

$$E_{\mathrm{wc}}(A_{\mathbb{U}_n}, \mathcal{K}(\mathbb{U}_n, \varepsilon_n), \mathbb{W}) = E_{\mathrm{wc}}(\mathcal{K}(\mathbb{U}_n, \varepsilon_n), \mathbb{W}) = \mu(\mathbb{U}_n, \mathbb{W})\varepsilon_n. \tag{3.8}$$

The quantity $\mu(\mathbb{U}_n, \mathbb{W})$ also coincides with the norm of the linear recovery map $A_{\mathbb{U}_n}$. Relaxing the prior $u \in \mathcal{M}$ by exploiting information on $\mathcal{M}$ solely through approximability of $\mathcal{M}$ by $\mathbb{U}_n$, thus implicitly *regularizes* the estimation task: whenever $\mu(\mathbb{U}_n, \mathbb{W})$ is finite, the optimal recovery map $A_{\mathbb{U}_n}$ is bounded and hence Lipschitz.

One important observation is that the map $A_{\mathbb{U}_n}$ is actually independent of ε_n. In particular, it achieves optimality for the smallest possible containment cylinder

$$\mathcal{K}(\mathbb{U}_n) := \mathcal{K}(\mathbb{U}_n, \mathrm{dist}(\mathcal{M}, \mathbb{U}_n)_{\mathbb{U}}), \tag{3.9}$$

and therefore, since $E_{\mathrm{wc}}(A_{\mathbb{U}_n}, \mathcal{M}, \mathbb{W}) \le E_{\mathrm{wc}}(A_{\mathbb{U}_n}, \mathcal{K}(\mathbb{U}_n), \mathbb{W}) = E_{\mathrm{wc}}(\mathcal{K}(\mathbb{U}_n), \mathbb{W})$,

$$E_{\mathrm{wc}}(A_{\mathbb{U}_n}, \mathcal{M}, \mathbb{W}) \le \mu(\mathbb{U}_n, \mathbb{W})\mathrm{dist}\,(\mathcal{M}, \mathbb{U}_n)_{\mathbb{U}}. \tag{3.10}$$

Likewise, the containment $\mathcal{M}^\varepsilon \subset \mathcal{K}(\mathbb{U}_n, \mathrm{dist}\,(\mathcal{M}, \mathbb{U}_n)_{\mathbb{U}} + \varepsilon)$ implies that

$$E_{\mathrm{wc}}(A_{\mathbb{U}_n}, \mathcal{M}^\varepsilon, \mathbb{W}) \le \mu(\mathbb{U}_n, \mathbb{W})(\mathrm{dist}\,(\mathcal{M}, \mathbb{U}_n)_{\mathbb{U}} + \varepsilon). \tag{3.11}$$

On the other hand, the recovery map $A_{\mathbb{U}_n}$ may be far from optimal over the sets $\mathcal{M}$ or $\mathcal{M}^\varepsilon$. This is due to the fact that the cylinders $\mathcal{K}(\mathbb{U}_n, \varepsilon_n)$ and $\mathcal{K}(\mathbb{U}_n, \varepsilon_n + \varepsilon)$ may be much larger than $\mathcal{M}$ or $\mathcal{M}^\varepsilon$. In particular, it is quite possible that for a particular observation w, one has $\mathrm{rad}(\mathcal{M}_w) \ll \mathrm{rad}(\mathcal{K}_w(\mathbb{U}_n, \varepsilon_n))$. Therefore, we cannot generally expect that the one space method achieves our goal (2.26). In particular, the condition $n \leq m$, which is necessary to avoid that $\mu(\mathbb{U}_n, \mathbb{W}) = \infty$, limits the dimension of an approximating subspace $\mathbb{U}_n$ and therefore ε_n itself is inherently bounded from below. The "dimension budget" has therefore to be used wisely in order to obtain good performance bounds. This typically rules out "generic approximation spaces" such as finite element spaces, and raises the question which subspace $\mathbb{U}_n$ yields the best estimator when applying the above method.

3.2 *Optimal Affine Recovery*

The results of the previous section bring forward the question as to what is the best choice of the space $\mathbb{U}_n$ for the given $\mathcal{M}$. On the one hand, proximity to $\mathcal{M}$ is desirable since $\mathrm{dist}\,(\mathcal{M}, \mathbb{U}_n)_{\mathbb{U}}$ enters the error bound. However, favoring proximity, may increase $\mu(\mathbb{U}_n, \mathbb{W})$. Before addressing this question systematically, it is important to note that the above results carry over verbatim when $\mathbb{U}_n$ is replaced by an *affine space* $\mathbb{U}_n = \bar{u} + \widetilde{\mathbb{U}}_n$ where $\widetilde{\mathbb{U}}_n \subset \mathbb{U}$ is a linear space. This means the reduced model $\mathcal{K}(\mathbb{U}_n, \varepsilon_n)$ is of the form

$$\mathcal{K}(\mathbb{U}_n, \varepsilon_n) := \bar{u} + \mathcal{K}(\widetilde{\mathbb{U}}_n, \varepsilon_n).$$

The best worst-case recovery bound is now given by

$$E_{\mathrm{wc}}(\mathcal{K}(\mathbb{U}_n, \varepsilon_n), \mathbb{W}) = \mu(\widetilde{\mathbb{U}}_n, \mathbb{W})\varepsilon_n. \tag{3.12}$$

Intuitively, this may help to better control the angle between $\mathbb{W}$ and $\mathbb{U}_n$ by anchoring the affine space at a suitable location (typically near or on $\mathcal{M}$). More importantly, it helps in *localizing* models via parameter domain decompositions that will be discussed later.

The one-space algorithm discussed in the previous section confines the "dimensionality" budget of the approximation spaces $\mathbb{U}_n$ to $n \leq m$. In view of (3.10), to obtain an overall good estimation accuracy, this space can clearly not be chosen arbitrarily but should be well adapted both to the solution manifold $\mathcal{M}$ and to measurement space W, that is, to the given observation functionals giving rise to the data.

A simple way of *adapting* a recovery space to $\mathbb{W}$ is as follows: suppose for a moment that we were able to construct for $n = 1, \dots, m$, a hierarchy of spaces $\mathbb{U}_1^{\mathrm{nb}} \subset \mathbb{U}_2^{\mathrm{nb}} \subset \cdots \subset \mathbb{U}_m^{\mathrm{nb}}$, that approximate $\mathcal{M}$ in a *near-best* way, namely

$$\mathrm{dist}\,(\mathcal{M}, \mathbb{U}_n^{\mathrm{nb}})_{\mathbb{U}} \leq C d_n(\mathcal{M})_{\mathbb{U}}. \tag{3.13}$$

We may compute along the way the quantities $\mu(\mathbb{U}_j^{\rm nb}, \mathbb{W})$, then choose

$$n^* = \operatorname*{argmin}_{n \le m} \mu(\mathbb{U}_n^{\rm nb}, \mathbb{W}) \operatorname{dist}(\mathcal{M}, \mathbb{U}_n^{\rm nb})_{\mathbb{U}}, \tag{3.14}$$

and take the map $A_{\mathbb{U}_{n^*}^{\rm nb}}$. We sometimes refer to this choice as *"poor man's algorithm"*. It is not clear though whether $\mathbb{U}_{n^*}^{\rm nb}$ is indeed a near-best choice for state recovery by the one-space method. In other words, one may question whether

$$E_{\rm wc}(A_{\mathbb{U}_{n^*}^{\rm nb}}, \mathcal{M}, \mathbb{W}) \le C \inf_{\dim \widetilde{\mathbb{U}} \le m} E_{\rm wc}(A_{\widetilde{\mathbb{U}}}, \mathcal{M}, \mathbb{W}), \tag{3.15}$$

holds with a uniform constant $C < \infty$. In fact, numerical tests strongly suggest otherwise, which motivated in [12] the following alternative to the poor man's algorithm.

Recall that a given linear space $\mathbb{U}_n$ determines the linear recovery map $A_{\mathbb{U}_n}$. Likewise a given affine space $\mathbb{U}_n$ determines an affine recovery map $A_{\mathbb{U}_n}$. Conversely, it can be checked that an affine recovery map A determines an affine space $\mathbb{U}_n$ that allows one to interpret the recovery scheme as a one-space method in the sense that $A = A_{\mathbb{U}_n}$. Denoting by $\mathcal{A}$ the class of all affine mappings of the form

$$A(w) = w + z + Bw, \tag{3.16}$$

where $z \in \mathbb{W}^\perp$ and $B \in \mathcal{L}(\mathbb{W}, \mathbb{W}^\perp)$ is linear, we might thus as well directly look for a mapping that minimizes

$$E_{\rm wc}(A, \mathcal{M}, \mathbb{W}) := \sup_{u \in \mathcal{M}} \|u - A(P_{\mathbb{W}} u)\|_{\mathbb{U}} = \sup_{u \in \mathcal{M}} \|P_{\mathbb{W}^\perp} u - z - BP_{\mathbb{W}} u\|_{\mathbb{U}} =: \mathcal{E}(z, B) \tag{3.17}$$

over $\mathcal{A}$, i.e., over all $(z, B) \in \mathbb{W}^\perp \times \mathcal{L}(\mathbb{W}, \mathbb{W}^\perp)$. It can be shown that indeed a minimizing pair (z^*, B^*) exists, i.e.,

$$\mathcal{E}(z^*, B^*) = \min_{A \in \mathcal{A}} E_{\rm wc}(A, \mathcal{M}, \mathbb{W}) =: E_{{\rm wc}, \mathcal{A}}(\mathcal{M}, \mathbb{W}),$$

see [12]. However, the minimization of $E_{\rm wc}(A, \mathcal{M}, \mathbb{W})$ over $(z, B) \in \mathbb{W}^\perp \times \mathcal{L}(\mathbb{W}, \mathbb{W}^\perp)$ is far from practically feasible. In fact, each evaluation of $E_{\rm wc}(A, \mathcal{M}, \mathbb{W})$ requires exploring $\mathcal{M}$ and B can have a range in the infinite dimensional space $\mathbb{W}^\perp$. In order to arrive at a computationally tractable problem, one needs to

(i) Replace $\mathcal{M}$ by a finite set $\widetilde{\mathcal{M}} \subset \mathcal{M}$, that should be sufficiently dense. Denseness can be quantified by requiring that $\widetilde{\mathcal{M}} = \widetilde{\mathcal{M}}^\delta$ is a δ-net for $\mathcal{M}$ for some $\delta > 0$, i.e., for any $u \in \mathcal{M}$, there exists $\tilde{u} \in \widetilde{\mathcal{M}}^\delta$ such that $\|u - \tilde{u}\|_{\mathbb{U}} \le \delta$.

(ii) Choose a finite dimensional space $\mathbb{U}_L \subset \mathbb{U}$ that approximates $\mathcal{M}$ to a desired precision $\operatorname{dist}(\mathcal{M}, \mathbb{U}_L)_{\mathbb{U}} \le \eta$, and replace $\mathbb{W}^\perp$ by the finite dimensional complement

$$\widetilde{\mathbb{W}}^{\perp} := \mathbb{U}_L \ominus \mathbb{W} \tag{3.18}$$

of $\mathbb{W}$ in $\mathbb{U}_L$.

The resulting optimization problem

$$(\tilde{z}, \widetilde{B}) = \operatorname*{argmin}_{(z,B)\in\widetilde{\mathbb{W}}^{\perp}\times\mathcal{L}(\mathbb{W},\widetilde{\mathbb{W}}^{\perp})} \sup_{u\in\widetilde{\mathcal{M}}^{\delta}} \|P_{\mathbb{W}^{\perp}}u - z - BP_{\mathbb{W}}u\|_{\mathbb{U}}. \tag{3.19}$$

can be solved by primal-dual splitting methods providing a $O(1/k)$ convergence rate, [12].

Due to the perturbations (i) and (ii) of the ideal minimization problem, the resulting $(\tilde{z}, \widetilde{B})$ is no longer optimal. However, one can show that

$$E_{\rm wc}(\widetilde{A}, \mathcal{M}, \mathbb{W}) \le E_{{\rm wc},\mathcal{A}}(\mathcal{M}, \mathbb{W}) + \eta + C\delta, \tag{3.20}$$

where the constant C is the operator norm of B minimizing (3.17). On the other hand, since the range of any affine mapping A is an affine space of dimension at most m, therefore contained in a linear space of dimension at most $m+1$, one always has $E_{{\rm wc},\mathcal{A}}(\mathcal{M}, \mathbb{W}) \ge d_{m+1}(\mathcal{M})_{\mathbb{U}}$. Therefore $(\tilde{z}, \widetilde{B})$ satisfies a near-optimal bound

$$E_{\rm wc}(\widetilde{A}, \mathcal{M}, \mathbb{W}) \lesssim E_{{\rm wc},\mathcal{A}}(\mathcal{M}, \mathbb{W}), \tag{3.21}$$

whenever η and δ are picked such that

$$\eta \lesssim d_{m+1}(\mathcal{M})_{\mathbb{U}}, \quad \text{and} \quad \delta \lesssim d_{m+1}(\mathcal{M})_{\mathbb{U}}. \tag{3.22}$$

The numerical tests in [12] for a model problem of the type (2.6) with piecewise constant checkerboard diffusion coefficients and d_y up to $d_y = 64$ show that this recovery map exhibits significantly better accuracy than the method based on (3.14). It even yields smaller error bounds than the affine mean square estimator (2.27). The following section discusses the numerical cost entailed by conditions like (3.22).

3.3 *Rate-Optimal Reduced Bases*

To keep the dimension L of the space $\mathbb{U}_L$ in (3.18) small, a near-best subspace $\mathbb{U}_L^{\rm nb}$ in the sense of (3.13) would be highly desirable. Likewise the poor man's scheme (3.14) would benefit from such subspaces. Unfortunately, such near-best subspaces are not practically accessible. The *reduced basis method* aims to construct subspaces which come close to near-optimality in a sense that we further explain next. The main idea is to generate theses subspaces by a sequence of elements picked from the manifold $\mathcal{M}$ itself, by means of a *weak-greedy algorithm* introduced and studied in [8]. In an idealized form, this algorithm proceeds as follows: given a current space $\mathbb{U}_n^{\rm wg} = \mathrm{span}\{u_1, \dots, u_n\}$, one takes $u_{n+1} = u(y_{n+1})$ such that, for some fixed $\gamma \in]0, 1]$, $\|u_{n+1} - P_{\mathbb{U}_n} u_{n+1}\|_{\mathbb{U}} \geq \gamma \max_{u \in \mathcal{M}} \|u - P_{\mathbb{U}_n} u\|_{\mathbb{U}}$, or equivalently

$$\|u(y_{n+1}) - P_{\mathbb{U}_n} u(y_{n+1})\|_{\mathbb{U}} \geq \gamma \max_{y \in \mathcal{Y}} \|u(y) - P_{\mathbb{U}_n} u(y)\|_{\mathbb{U}}. \tag{3.23}$$

Then, one defines $\mathbb{U}_{n+1}^{\rm wg} = \mathrm{span}\{u_1, \dots, u_{n+1}\}$. While unfortunately, the weak greedy algorithm does in general not produce spaces satisfying (3.13), it does come close. Namely, it has been shown in [3, 19] that the spaces $\mathbb{U}_n^{\rm wg}$ are *rate-optimal* in the following sense:

(i) For any $s > 0$ one has

$$d_n(\mathcal{M})_{\mathbb{U}} \leq C(n+1)^{-s}, \ n \geq 0 \implies \mathrm{dist}\,(\mathcal{M}, \mathbb{U}_n^{\rm wg})_{\mathbb{U}} \leq \widetilde{C}(n+1)^{-s}, \ n \geq 0, \tag{3.24}$$

where $\widetilde{C}$ depends on C, s, γ.

(ii) For any $\beta > 0$, one has

$$d_n(\mathcal{M})_{\mathbb{U}} \leq Ce^{-cn^\beta}, \ n \geq 0 \implies \mathrm{dist}\,(\mathcal{M}, \mathbb{U}_n^{\rm wg})_{\mathbb{U}} \leq \widetilde{C}e^{-\tilde{c}n^\beta}, \ n \geq 0, \tag{3.25}$$

where the constants $\tilde{c}, \widetilde{C}$ depend on c, C, β, γ.

In the form described above, the weak-greedy concept seems infeasible since it would, in principle, require computing the solution $u(y)$ for all values of $y \in \mathcal{Y}$ exactly, exploring the whole exact solution manifold. However, its practical applicability is facilitated when there exists a *tight* surrogate $R(y, \mathbb{U}_n)$, satisfying

$$c_R R(y, \mathbb{U}_n) \leq \|u(y) - P_{\mathbb{U}_n} u(y)\|_{\mathbb{U}} = \mathrm{dist}\,(u(y), \mathbb{U}_n) \leq C_R R(y, \mathbb{U}_n), \quad y \in \mathcal{Y}, \tag{3.26}$$

for uniform constants $0 < c_R \leq C_R < \infty$, which can be evaluated at affordable cost. Then, maximization of $R(y, \mathbb{U}_n)$ over $\mathcal{Y}$ amounts to the weak-greedy step (3.23) with $\gamma := \frac{c_R}{C_R}$. According to [18], the validity of the following two conditions indeed allows one to derive computable surrogates that satisfy (3.26):

(i) The underlying parametric family of PDEs (2.1) permits a uniformly *stable variational formulation* (2.4), and one has *affine parameter dependence* (2.9);
(ii) The discrete projection $\Pi_{\mathbb{U}_n}$ (of Galerkin or Petrov-Galerkin type) has the *best approximation property*, i.e., resulting errors are uniformly comparable to the best approximation error.

Conditions (i) and (ii) ensure, in view of (2.5), that $\|u(y) - P_{\mathbb{U}_n} u(y)\|_{\mathbb{U}} \sim \|\mathcal{R}(y, \Pi_{\mathbb{U}_n} u(y))\|_{\mathbb{V}'}$ holds uniformly in $y \in \mathcal{Y}$. Thus,

$$R(y, \mathbb{U}_n) := \|\mathcal{R}(y, \Pi_{\mathbb{U}_n} u(y))\|_{\mathbb{V}'} = \sup_{v \in \mathbb{V}} \frac{\mathcal{R}(y, \Pi_{\mathbb{U}_n} u(y))(v)}{\|v\|_{\mathbb{V}}} \tag{3.27}$$

satisfies (3.26) and is therefore a tight surrogate for $\mathrm{dist}\,(\mathcal{M}, \mathbb{U}_n)_{\mathbb{U}}$. In the elliptic case (2.6) under assumption (2.7), (i) and (ii) hold and the above comments reflect standard practice. For the wider scope of stable but *unsymmetric* variational formulations [6, 16, 23] the inf-sup conditions (2.4) imply (i), but the Galerkin projection in (ii) needs to be replaced by a stable *Petrov-Galerkin* projection with respect to suitable test spaces $\mathbb{V}_n$ accompanying the reduced trial spaces $\mathbb{U}_n$. It has been shown in [18] how to generate such test spaces with the aid of a *double-greedy* strategy, see also [16].

The main pay-off of using the surrogate $R(y, \mathbb{U}_n)$ is that one no longer needs to compute $u(y)$ but only the low-dimensional projection $\Pi_{\mathbb{U}_n} u(y)$ by solving for each y an $n \times n$ system, which itself can be rapidly assembled thanks to the affine parameter dependence [22]. However, one still faces the problem of its exact maximization over $y \in \mathcal{Y}$. A standard approach is to maximize instead over a *discrete* training set $\widetilde{\mathcal{Y}}_n \subset Y$, which in turn induces a discretization of the solution manifold

$$\widetilde{\mathcal{M}}_n = \{u(y) \;:\; y \in \widetilde{\mathcal{Y}}_n\}. \tag{3.28}$$

The resulting weak-greedy algorithm can be shown to remain rate optimal in the sense of (3.24) and (3.25) if the discretization is fine enough so that $\widetilde{\mathcal{M}}_n$ constitutes an ε_n-approximation net of $\mathcal{M}$ where ε_n does not exceed $c\,\mathrm{dist}\,(\mathcal{M}, \mathbb{U}_n^{\mathrm{wg}})_{\mathbb{U}}$ for a suitable constant $0 < c < 1$. In the current regime of *large or even infinite parameter dimensionality*, this becomes prohibitive because $\#\widetilde{\mathcal{Y}}_n$ would then typically scale like $O\big(\varepsilon_n^{-cd_y}\big)$, [10].

As a remedy it has been proposed in [10] to use training sets $\widetilde{\mathcal{Y}}_n$ that are generated by *randomly sampling* $\mathcal{Y}$, and ask that the objective of rate optimality is met with high probability. This turns out to be achievable with training sets of much less prohibitive size. In an informal and simplified manner the main result can be stated as follows.

Theorem 1 *Given any target accuracy $\varepsilon > 0$ and some $0 < \eta < 1$, then the weak greedy reduced basis algorithm based on choosing at each step $N = N(\varepsilon, \eta) \sim |\ln \eta| + |\ln \varepsilon|$ randomly chosen training points in $\mathcal{Y}$ has the following properties with probability at least $1 - \eta$: it terminates with dist $(\mathcal{M}, \mathbb{U}_{n(\varepsilon)})_{\mathbb{U}} \le \varepsilon$ as soon as the maximum of the surrogate over the current training set falls below $c\varepsilon^{1+a}$ for some $c, a > 0$. Moreover, if $d_n(\mathcal{M})_{\mathbb{U}} \le Cn^{-s}$, then $n(\varepsilon) \lesssim \varepsilon^{-\frac{1}{s}-b}$. The constants c, a, b depend on the constants in (3.26), as well as on the rate r of polynomial approximability of the parameter to solution map $y \mapsto u(y)$. The larger s and r, the smaller a and b, and the closer the performance becomes to the ideal one.*

4 Nonlinear Models

4.1 Piecewise Affine Reduced Models

As already noted, schemes based on linear or affine reduced models of the form $\mathcal{K}(\mathbb{U}_n, \varepsilon)$ can, in general, not be expected to realize the benchmark (2.26), discussed earlier in Sect. 2. The convexity of the containment set $\mathcal{K}(\mathbb{U}_n, \varepsilon)$ may cause the reconstruction error to be significantly larger than $\delta_\varepsilon(\mathcal{M}, \mathbb{W})$. Another way of understanding this limitation is that in order to make ε small, one is enforced to raise the dimension n of $\mathbb{U}_n$, making the quantity $\mu(\mathbb{U}_n, \mathbb{W})$ larger and eventually infinite if $n > m$.

To overcome this principal limitation one needs to resort to *nonlinear* models that better capture the non-convex geometry of $\mathcal{M}$. One natural approach consists in replacing the single space $\mathbb{U}_n$ by a family $(\mathbb{U}^k)_{k=1,\dots,K}$ of affine spaces

$$\mathbb{U}^k = \bar{u}_k + \widetilde{\mathbb{U}}^k, \quad \dim(\widetilde{\mathbb{U}}^k) = n_k \le m, \tag{4.1}$$

each of which aims to approximate a *portion* $\mathcal{M}_k$ of $\mathcal{M}$ to a prescribed target accuracy simultaneously controlling $\mu(\mathbb{U}^k, \mathbb{W})$: fixing $\varepsilon > 0$, we assume that we have at hand a partition of $\mathcal{M}$ into portions

$$\mathcal{M} = \bigcup_{k=1}^{K} \mathcal{M}_k \tag{4.2}$$

such that

$$\operatorname{dist}(\mathcal{M}_k, \mathbb{U}^k)_{\mathbb{U}} \le \varepsilon_k, \quad \text{and} \quad \mu(\widetilde{\mathbb{U}}^k, \mathbb{W})\varepsilon_k \le \varepsilon, \quad k = 1, \dots, K. \tag{4.3}$$

One way of obtaining such a partition is through a greedy splitting procedure of the domain $\mathcal{Y} = [-1, 1]^{d_y}$ which is detailed in [9]. The procedure terminates when for each cell $\mathcal{Y}_k$ the corresponding portion of the manifold $\mathcal{M}_k$ can be associated to an affine $\mathbb{U}_k$ satisfying these properties. We are ensured that this eventually occurs since for a sufficiently fine cell $\mathcal{Y}_k$ one has $\operatorname{rad}(\mathcal{M}_k) \le \varepsilon$ which means that we could then use a zero dimensional affine space $\mathbb{U}_k = \{\bar{u}_k\}$ for which we know that $\mu(\widetilde{\mathbb{U}}^k, \mathbb{W}) = 1$. In this piecewise affine model, the containment property is now

$$\mathcal{M} \subset \bigcup_{k=1}^{K} \mathcal{K}(\mathbb{U}_k, \varepsilon_k). \tag{4.4}$$

and the cardinality K of the partition depends on the prescribed ε.

For a given measurement $w \in \mathbb{W}$, we may now compute the state estimates

$$u_k^*(w) = A_{\mathbb{U}^k}(w), \quad k = 1, \dots, K, \tag{4.5}$$

by the affine variant of the one-space method from (3.4). Since $u \in \mathcal{M}_{k_0}$ for some value k_0, we are ensured that

$$\|u - u_{k_0}^*(w)\|_{\mathbb{U}} \le \varepsilon, \tag{4.6}$$

for this particular choice. However k_0 is unknown to us and one has to rely on the data w in order to decide which one among the affine models is most appropriate for the recovery. One natural *model selection* criterion can be derived if for any $\overline{u} \in \mathbb{U}$ we have at our disposal a computable surrogate $S(\overline{u})$ that is equivalent to the distance from $\overline{u}$ to $\mathcal{M}$, that is

$$cS(\bar{u}) \le \operatorname{dist}(\bar{u}, \mathcal{M})_{\mathbb{U}} \le CS(\bar{u}), \quad \operatorname{dist}(\bar{u}, \mathcal{M})_{\mathbb{U}} = \min_{y \in \mathcal{Y}} \|\bar{u} - u(y)\|_{\mathbb{U}}, \tag{4.7}$$

for some fixed $0 < c \le C$. We give an instance of such a computable surrogate in Sect. 4.2 below. The selection criterion then consists in picking k^* minimizing this surrogate between the different available state estimates, that is,

$$u^*(w) := u^*_{k^*}(w) = \operatorname{argmin}\{S(u^*_k(w)) \,:\, k = 1, \dots, K\}. \tag{4.8}$$

The following result, established in [9], shows that this estimator now realizes the benchmark (2.26) up to a multiplication of ε by $\kappa := C/c$, where c, C are the constants from (4.7).

Theorem 2 *Assume that* (4.2) *and* (4.3) *hold. For any $u \in \mathcal{M}$, if $w = P_{\mathbb{W}}u$, one has*

$$\|u - u^*(w)\| \le \delta_{\kappa\varepsilon}(\mathcal{M}, \mathbb{W}), \tag{4.9}$$

where $\delta_\varepsilon(\mathcal{M}, \mathbb{W})$ is given by (2.25).

4.2 Approximate Metric Projection and Parameter Estimation

A practically affordable realization of the surrogate $S(\overline{u})$, providing a *near-metric projection distance* to $\mathcal{M}$, is a key ingredient of the above nonlinear recovery scheme. Since it has further useful implications we add a few comments on that matter.

As already observed in (2.5), whenever (2.1) admits a stable variational formulation with respect to a suitable pair $(\mathbb{U}, \mathbb{V})$ of trial and test spaces, the distance of any $\overline{u} \in \mathbb{U}$ to any $u(y) \in \mathcal{M}$ is uniformly equivalent to the residual of the PDE in $\mathbb{V}'$

$$c\|\mathcal{R}(\bar{u}, y)\|_{\mathbb{V}'} \le \|u(y) - \bar{u}\|_{\mathbb{U}} \le C\|\mathcal{R}(\bar{u}, y)\|_{\mathbb{V}'}, \tag{4.10}$$

with $c = C_b^{-1}$, $C = c_b^{-1}$ from (2.5). Assume in addition that $\mathcal{R}(u, y)$ depends *affinely* on $y \in \mathcal{Y}$, according to (2.9). Then, minimizing $\|\mathcal{R}(\bar{u}, y)\|_{\mathbb{V}'}$ over y is equivalent to solving a *constrained least squares* problem

$$\bar{y} = \operatorname*{argmin}_{y \in \mathcal{Y}} \|\mathbf{g} - \mathbf{M}y\|_2, \tag{4.11}$$

where $\mathbf{M}$ is a matrix of size $d_y \times d_y$ resulting from Riesz-lifts of the functionals $\mathcal{R}_j(\bar{u})$.

The solution to this problem therefore satisfies

$$\|\bar{u} - u(\bar{y})\|_{\mathbb{U}} \le \kappa \inf_{y \in \mathcal{Y}} \|\bar{u} - u(y)\|_{\mathbb{U}} = \kappa \operatorname{dist}(\bar{u}, \mathcal{M})_{\mathbb{U}}. \tag{4.12}$$

where $\kappa = C/c = C_b/c_b$ is the quotient between the equivalence constants in (4.10). The surrogate

$$S(\bar{u}) := \|\mathcal{R}(\bar{u}, y)\|_{\mathbb{V}'} \tag{4.13}$$

for the metric projection distance of $\overline{u}$ onto $\mathcal{M}$ obviously satisfies (4.7). It is indeed computable at affordable cost using (an approximation to) its Riesz-lifted version $\|e(\bar{u}, y)\|_{\mathbb{V}} = \|\mathcal{R}(\bar{u}, y)\|_{\mathbb{V}'}$ (in $\mathbb{V}_h \subset \mathbb{V}$) assembled from the Riesz-lifts of the components $\mathcal{R}_j(\bar{u})$, see [9] for details in the affine expansion (2.9).

Since solving the above problem provides an admissible parameter value $\bar{y} \in \mathcal{Y}$, this also has some immediate bearing on *parameter estimation*. Suppose we wish to estimate from $w = P_{\mathbb{W}} u(y^*)$ the unknown parameter $y^* \in \mathcal{Y}$. Assume further that A is any given linear or nonlinear recovery map. Computing along the above lines

$$\bar{y}_w = \operatorname*{argmin}_{y \in \mathcal{Y}} \|\mathcal{R}(A(w), y)\|_{\mathbb{V}'}$$

we have

$$\begin{aligned} \|u(y^*) - u(\bar{y}_w)\|_{\mathbb{U}} &\le \|u(y^*) - A(w)\|_{\mathbb{U}} + \|A(w) - u(\bar{y}_w)\|_{\mathbb{U}} \\ &\le E_{\mathrm{wc}}(A, \mathcal{M}, \mathbb{W}) + \kappa \operatorname{dist}(A(w), \mathcal{M})_{\mathbb{U}} \le (1+\kappa) E_{\mathrm{wc}}(A, \mathcal{M}, \mathbb{W}). \end{aligned} \tag{4.14}$$

We consider now the specific elliptic model (2.6) with affine diffusion coefficients $a(y)$ given by (2.10). For this model, it was established in [5] that for strictly positive f and certain regularity assumptions on $a(y)$ as functions of $x \in \Omega$, parameters may be estimated by states. Specifically, when $a(y) \in H^1(\Omega)$ uniformly in $y \in \mathcal{Y}$, one has an inverse stability estimate of the form

$$\|a(y) - a(\tilde{y})\|_{L_2(\Omega)} \le C \|u(y) - u(\tilde{y})\|_{\mathbb{U}}^{1/6}. \tag{4.15}$$

Thus, whenever the recovery map A satisfies (4.9) for some prescribed $\varepsilon > 0$, we obtain a parameter estimation bound of the form

$$\|a(y^*) - a(\bar{y}_w)\|_{L_2(\Omega)} \le C \delta_{\kappa\varepsilon}(\mathcal{M}, \mathbb{W})^{1/6},$$

Note that when the basis functions θ_j are L_2-orthogonal, $\|a(y^*) - a(\bar{y}_w)\|_{L_2(\Omega)}$ is equivalent to a (weighted) ℓ_2 norm of $y^* - \bar{y}_w$.

4.3 Concluding Remarks

The affine or piecewise affine recovery scheme hinges on the ability to approximate a solution manifold effectively by linear or affine spaces, globally or locally. As explained earlier this is true for problems of elliptic or parabolic type that may include convective terms as long as they are dominated by diffusion. This may however no longer be the case when dealing with pure transport equations or models involving strongly dominating convection.

An interesting alternative would then be to adopt a stochastic model according to (2.27) and (2.28) that allows one to view the construction of the recovery map as a regression problem. In particular, when dealing with transport models, a natural candidate for parametrizing a reduced model are *deep neural networks*. However, properly adapting the architecture, regularization and training principles pose wide open questions addressed in current work in progress.

Acknowledgements A. C. was supported by ERC Adv Grant BREAD. W. D. was supported in part by the NSF-grants DMS 17-20297, DMS-2012469, by the Smart State Program of the State of South Carolina, and the Williams-Hedberg Foundation. R. D. was supported by the NSF grant DMS 18-17603. A portion of this work was completed when the authors were supported as visitors to the Isaac Newton Institute of Cambridge University.

References

1. Adcock, B., Hansen, A.C., Poon, C.: Beyond consistent reconstructions: optimality and sharp bounds for generalized sampling, and application to the uniform resampling problem. SIAM J. Math. Anal. **45**, 3132–3167 (2013)
2. Bachmayr, M., Cohen, A.: Kolmogorov widths and low-rank approximations of parametric elliptic PDEs. Math. Comp. **86**, 701–724 (2017)
3. Binev, P., Cohen, A., Dahmen, W., DeVore, R., Petrova, G., Wojtaszczyk, P.: Convergence rates for greedy algorithms in reduced basis methods. SIAM J. Math. Anal. **43**, 1457–1472 (2011)
4. Binev, P., Cohen, A., Dahmen, W., DeVore, R., Petrova, G., Wojtaszczyk, P.: Data assimilation in reduced modeling. SIAM J. Uncert Quantif **5**(1), 1–29 (2017)
5. Bonito, A., Cohen, A., DeVore, R., Petrova, G., Welper, G.: Diffusion coefficients estimation for elliptic partial differential equations. SIAM J. Math. Anal. **49**(2), 1570–1592 (2017)
6. Broersen, D., Stevenson, R.: A robust Petrov-Galerkin discretisation of convection-diffusions equations. Comput. Math. Appl. **68**(11), 1605–1618 (2014)
7. Broersen, D., Dahmen, W., Stevenson, R.: On the stability of DPG formulations of transport equations. Math. Comp. **87**(311), 1051–1082 (2018)
8. Buffa, A., Maday, Y., Patera, A.T., Prud'homme, C., Turicini, G.: A piori convergence of the greedy algorithm for the parametrized reduced bases. ESAIM Math. Model. Numer. Anal. **46**(03), 595–603 (2012)
9. Cohen, A., Dahmen, W., Mula, O., Nichols, J.: Nonlinear reduced models for state and parameter estimation (to appear in SIAM J. Uncert. Quantif). https://arxiv.org/submit/3356244
10. Cohen, A., Dahmen, W., DeVore, R., Nichols, J.: Reduced basis greedy selection using random training sets. ESAIM: M2AN **54**(5), 1509–1524. https://doi.org/10.1051/m2an/2020004, http://arxiv.org/abs/1810.09344 [math.NA]
11. Cohen, A., Dahmen, W., Welper, G.: Adaptivity and variational stabilization for convection-diffusion equations. ESAIM: Math. Model. Numer. Anal. **46**(5), 1247–1273 (2012)

12. Cohen, A., Dahmen, W., DeVore, R., Fadili, J., Mula, O., Nichols, J.: Optimal reduced model algorithms for data-based state estimation. SIAM J. Numer. Anal. **58**(6), 3355–3381 (2020). http://arxiv.org/abs/1903.07938
13. Cohen, A., DeVore, R.: Approximation of high-dimensional PDEs. Acta Numer. **24**, 1–159 (2015)
14. Cohen, A., DeVore, R.: Kolmogorov widths under holomorphic mappings. IMA J. Numer. Anal. **36**(1), 1–12 (2016)
15. Cohen, A., DeVore, R., Schwab, C.: Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. Found. Comput. Math. **10**(6), 615–646 (2010)
16. Dahmen W.: How to best sample a solution manifold? In: Pfander, G.E. (eds.) Sampling Theory, a Renaissance, Applied and Numerical Harmonic Analysis. Birkhäuser. ISBN 978-3-319-19748-7. https://doi.org/10.1007/978-3-319-19749-4_11, http://arxiv.org/abs/1503.00307 [math.NA]
17. Dahmen, W., Huang, C., Schwab, C., Welper, G.: Adaptive Petrov-Galerkin methods for first order transport equations. SIAM J. Numer. Anal. **50**(5), 2420–2445 (2012)
18. Dahmen, W., Plesken, C., Welper, G.: Double greedy algorithms: reduced basis methods for transport dominated problems. ESAIM: Math. Model. Numer. Anal. **48**(3), 623–663 (2014)
19. DeVore R., Petrova, G., Wojtaszczyk, P.: Greedy algorithms for reduced bases in Banach spaces. Constr. Approx. **37**(3), 455–466 (2013)
20. Maday, Y., Patera, A.T., Penn, J.D., Yano, M.: A parametrized-background data-weak approach to variational data assimilation: formulation, analysis, and application to acoustics. Int. J. Numer. Methods Eng. Spec. Issue Model Reduct. **102**(5), 931–1292 (2015)
21. Micchelli, C.A., Rivlin, T.J.: A survey of optimal recovery. In: Micchelli, C.A., Rivlin, T.J. (eds.) Optimal Estimation in Approximation Theory, pp. 1–54. Plenum, NY (1977)
22. Rozza, G., Huynh, D.B.P., Patera, A.T.: Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations: application to transport and continuum mechanics. Arch. Comput. Methods Eng. **15**(3), 229–275 (2008). https://doi.org/10.1007/s11831-008-9019-9
23. Stevenson, R., Westerdiep, J.: Stability of Galerkin discretizations of a mixed space-time variational formulation of parabolic evolution equations. IMA J. Numer. Anal. **41**(1) (2021). https://doi.org/10.1093/imanum/drz069
24. Stuart, A.M.: Inverse problems: a Bayesian perspective. Acta Numer. **19**, 451–559 (2010). https://doi.org/10.1017/S0962492910000061

Pattern Formation Inside Living Cells

Leah Edelstein-Keshet

Abstract While most of our tissues appear static, in fact, cell motion comprises an important facet of all life forms, whether in single or multicellular organisms. Amoeboid cells navigate their environment seeking nutrients, whereas collectively, streams of cells move past and through evolving tissue in the development of complex organisms. Cell motion is powered by dynamic changes in the structural proteins (actin) that make up the cytoskeleton, and regulated by a circuit of signaling proteins (GTPases) that control the cytoskeleton growth, disassembly, and active contraction. Interesting mathematical questions we have explored include (1) How do GTPases spontaneously redistribute inside a cell? How does this determine the emergent polarization and directed motion of a cell? (2) How does feedback between actin and these regulatory proteins create dynamic spatial patterns (such as waves) in the cell? (3) How do properties of single cells scale up to cell populations and multicellular tissues given interactions (adhesive, mechanical) between cells? Here I survey mathematical models studied in my group to address such questions. We use reaction-diffusion systems to model GTPase spatiotemporal phenomena in both detailed and toy models (for analytic clarity). We simulate single and multiple cells to visualize model predictions and study emergent patterns of behavior. Finally, we work with experimental biologists to address data-driven questions about specific cell types and conditions.

1 Introduction: Motile Cells and Their Inner Workings

Many types of cells are endowed with the ability to move purposefully. As an example, neutrophils, shown in Fig. 1a, are white blood cells that make up part of our immune system, in charge of patrolling tissues for pathogens or sites of injury. The motion of unicellular organisms such as bacteria, while interesting in its own right, is governed by distinct mechanisms that will not be discussed here.

L. Edelstein-Keshet (✉)
University of British Columbia, Vancouver, BC, Canada
e-mail: keshet@math.ubc.ca

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_5

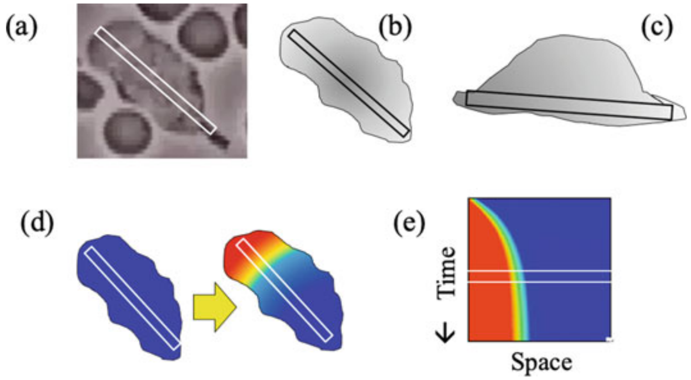


Fig. 1 Cell motility and cell polarization: from biology to mathematical model: **a** A white blood cell (neutrophil) moving between red blood cells (disk-shaped objects) from a 1950s movie clip by David Rogers. The 1D band represents a transect of the cell from front to back. We are concerned with how the cell breaks symmetry and polarizes to define such a front-back axis. **b**, **c** Sketch of a cell in top-down **b** and side **c** views, indicating the same 1D axis. **d** In our mathematical model, we aim to explain how regulatory proteins in the cell (called GTPases) spontaneously polarize and form hot spots of activity that define the front and back of the cell. **e** In our abstract "wave-pinning" model, this same process is depicted as a 1D pattern-formation event, with a wave that stalls to produced a polarized distribution

In a movie dating to the 1950s' David Rogers (then at Vanderbilt University) captured the amoeboid movements of a neutrophil as it navigates between red blood cells (disk shaped objects in Fig. 1a). In this movie, which can be seen on a popular YouTube site, we see a crawling cell, with dynamic shape—a broad front that pushes outwards, and a thin tail that is pulled along as the cell moves. Figure 1b, c are two projections of cell shape (top down in (b) and side view in (c)) that we later utilize in modeling cell polarization.

It is worth pointing out the sizes and timescales that concern us here. In contrast to some papers (e.g. Prof. Marsha Berger's whose work describes geological size scales and timescales of hours and days [1]), here we deal with the micro-world of cells, whose diameter is on the order of 10–30 μm. The time-scale of relevance is on the order of seconds. As summarized in Table 1, the process of cell polarization, which defines the front and back of the cell and specifies its direction of motion, take place over seconds across the tiny cell diameter. Also noteworthy is the fact that the production of new copies of proteins (i.e. protein synthesis) does not suffice to explain how protein activity becomes concentrated at some parts of a cell, since synthesis takes hour(s), while the response times of a cell to stimuli that polarize it is known to take only seconds for fast-moving cells like neutrophils.

Here the purpose is to explain an important first step in cell motility: the symmetry breaking that creates a front and a back in the cell (Fig. 1d), namely the polarization of the cell. But before embarking on the mathematics that describes this process, we first discuss the important cellular components that are involved.

Table 1 Typical sizes and speeds of cells, and typical time-scales of protein synthesis and activation

Cell part or process	Typical size
Cell diameter	10–30 μm
Cell thickness	0.1 μm
Cell speed (WBC)	0.1–0.2 μm/s
Response time to stimuli	Few seconds
Protein synthesis time	Hour(s) (!!)
Protein activation time	Few seconds
Diffusion rates (proteins)	0.1–10 μm^2/s

Recall that 1 μm = 10^{-6} m. *WBC* white blood cell (neutrophil)

1.1 Actin Powers Cell Motility

Unlike plants and bacteria, animal cells have no tough outer cell wall. They are enclosed in a lipid membrane that envelopes the interior, which in turn includes the fluid cytosol and many organelles. Most organelles, including the cell's nucleus are not directly involved in powering cellular motion.

Without some structural components, the cell would be essentially a bag of fluids. An internal "skeleton" (called the *cytoskeleton*) is formed by a meshwork of filamentous actin (F-actin), a dynamic biopolymer protein structure that is assembled at what becomes the cell front. The polymerization of actin leads to protrusion of the cell front [23]. Meanwhile, in association with the motor protein myosin, contraction of actomyosin leads to retraction of the rear portion of the cell [33], Fig. 2a.

Due to the abundance of actin monomers at excess concentration in every cell, actin assembly would be an explosive process were it not tightly controlled by many interacting regulatory cellular proteins. Many of those proteins, discovered and characterized experimentally over the last decades [27, 34], interact with actin to make it branch, to cut or cap its growing ends, to sequester or to recycle its monomeric subunits. Other proteins play the role of master-regulators that control the components of the cytoskeleton [30].

1.2 GTPases Are Master Regulators

One important class of proteins that regulate the cytoskeleton is the class of Rho GTPases, among which Rac and Rho are well known [3]. In the schematic Fig. 2, GTPases are shown to promote the assembly of filamentous actin, and the activity of myosin contraction. The GTPase Rac does the former, while the GTPase Rho enables the latter. Hence, if we can explain how Rac and Rho activities concentrate at one or another part of the cell, we can also explain the localizations of a front and rear cellular axis, and hence cell polarization. This then, is the main focus of our approach.

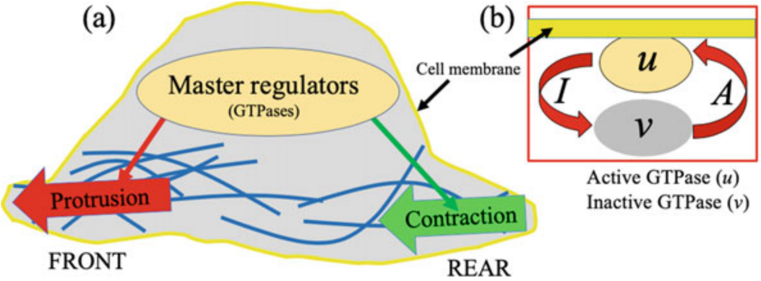


Fig. 2 Schematic diagram of the cell's motility machinery: **a** Actin filaments (F-actin), represented as blue curves, assemble at what becomes the cell front. Actin polymerization leads to protrusion at the front edge of the cell. In the cell rear, myosin motors (not shown) associate with F-actin to contract and pull up the "tail". Proteins in the class known as Rho GTPases are master regulators. These proteins control where and when actin assembly and myosin contraction take place. GTPases play an essential role in cell polarization. **b** Each GTPase has an active and an inactive state, modeled by the variables u, v. Only when bound to the cell membrane (shown in yellow) is the GTPase active. A, I denote rates of activation and inactivation

Interestingly, proteins in the family of Rho GTPases have a curious life-cycle. They occur in active and inactive forms, with only the active forms exerting the effects mentioned above [8]. Moreover, the active forms are always bound to the fatty membrane that forms the outer cell envelop (shown in yellow in Fig. 2). Hence, the small GTPases spend their cellular lives shuttling between the cell membrane (where part of their structure gets embedded when active) and the cell interior (where they are entirely inactive). This basic idea is illustrated in Fig. 2b. The GTPases act as cellular switches that are "ON" when active and "OFF" otherwise.

A natural question one could ask, is what is the functional purpose of the GTPase cycling between the cell's membrane and the cell's interior? As we shall see, mathematics may have something to contribute towards answering such questions. A second question is what property of the cellular machinery account for the spontaneous polarization of the cell? That is, how do GTPases redistribute so that their levels of activity differ between the front and rear of a cell [2].

2 Mathematical Models

In our earliest works on cell polarization, we attempted to account for many known features of the GTPase activity and their crosstalk and interactions [6, 18, 20]. Such models were largely computational, as it was a challenge to analyse them mathematically. It was clear that more basic model variants would be useful for mathematical progress to be feasible.

As described in Mori et al. [24, 25], we simplify a very complicated cellular process to allow for mathematical tractability. We thereby hope to identify key elements

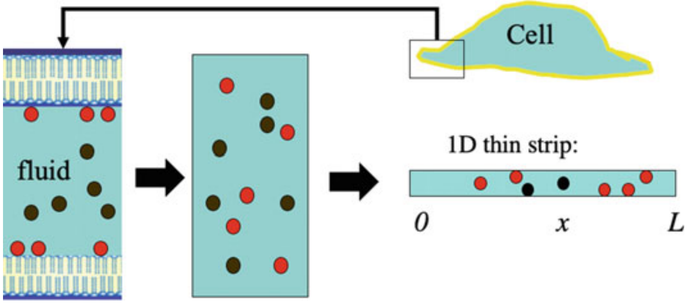


Fig. 3 **Model geometry**: The complicated cell geometry is simplified into a 1D domain (transect along the cell diameter) with active and inactive proteins distributed along that axis, but with distinct rates of diffusion, $D_u \ll D_v$

that allow for spontaneous cell polarization. First, we consider just one GTPase (say Rac), rather than the entire network (Cdc42, Rac and Rho). We ask which biological attributes account for spontaneous symmetry breaking and polar pattern formation. To investigate this, we construct the following mathematical model.

We define $u(t)$, $v(t)$ to be the concentrations of the active and inactive forms of the GTPase. Then, based on the schematic diagram in Fig. 2b, it follows that

$$\frac{du}{dt} = Av - Iu, \quad \frac{dv}{dt} = -Av + Iu.$$

This is not yet enough, since spatial distribution is a vital aspect. Hence, we require a spatial variable, and need to account for the localization of each of u, v. To do so, we also need to define the geometry of interest.

As argued earlier, and noted in Fig. 1, to explain symmetry breaking for polarization, a 1D model along the front-back axis suffices. And while the detailed residence of the proteins on the membrane or cell interior is important, it proves helpful to simplify this too, in the steps shown in Fig. 3. In that figure, we first idealize the cell as a thin sheet of uniform thickness, surrounded top and bottom by a membrane (yellow outline). Zooming in on a small portion of the cell, we might see active (red) and inactive (black) copies of the GTPase associated with the membrane or the fluid cell interior. We homogenize these compartments, treating both u and v as dependent variables on a 1D spatial domain $0 \le x \le L$ where L is the cell diameter. We do however, take into account the very different rates of diffusion of a protein in the membrane ($D_u \approx 0.01–0.1\,\mu\text{m}^2/s$) versus the fluid cell interior ($D_v \approx 10\,\mu\text{m}^2/s$) [28]. As we shall see, this huge disparity in diffusion plays a significant role.

The model becomes

$$\frac{\partial u}{\partial t} = D_u \frac{\partial^2 u}{\partial x^2} + Av - Iu, \tag{1a}$$

$$\frac{\partial v}{\partial t} = D_v \frac{\partial^2 v}{\partial x^2} - Av + Iu. \tag{1b}$$

In principle, the rates of activation and inactivation A, I, are not merely constant. If they were, then Eq. (1) would be linear in u, v, and would have fairly uninteresting steady state solutions. Some nonlinearity is essential, and this also requires feedback—something that can only depend on levels of active proteins. (Recall that the inactive GTPases do not participate in any interactions.) We have considered models where many other proteins influence each of the state transitions [14, 18, 21], and in that case, the model would expand in complexity,

$$\frac{\partial u_1}{\partial t} = D_u \frac{\partial^2 u_1}{\partial x^2} + A(u_1, u_2, \dots)v_1 - I(u_1, u_2, \dots)u_1, \tag{2a}$$

$$\frac{\partial v_1}{\partial t} = D_v \frac{\partial^2 v_1}{\partial x^2} - A(u_1, u_2, \dots)v_1 + I(u_1, u_2, \dots)u_1, \tag{2b}$$

$$\frac{\partial u_2}{\partial t} = \dots \tag{2c}$$

Such examples, considered in the context of biological experiments, are briefly discussed further on, but mathematically, they are harder to analyze.

Our ultimate purpose, mathematically, is to strip away such complexity and focus on the most elementary example, where a single GTPase polarizes on its own. To do so, we considered the version

$$\frac{\partial u}{\partial t} = D_u \frac{\partial^2 u}{\partial x^2} + A(u)v - Iu, \tag{3a}$$

$$\frac{\partial v}{\partial t} = D_v \frac{\partial^2 v}{\partial x^2} - A(u)v + Iu, \tag{3b}$$

with feedback exclusively in the activation rate $A(u)$ and a constant rate of inactivation I. This specific choice is somewhat arbitrary, as shown in [18], since it is possible to obtain essentially the same behaviour with nonlinearity introduced by assuming that $I = I(u)$ with A constant, or by other variants where both A and I depend on u. The biological interpretation is somewhat different, since distinct proteins in cells play the role of activating (GEFs) and inactivating (GAPS) the GTPases. In the case of constant I, we can rescale time, so that $I = 1$. Altogether, then, the single-GTPase system consists of the pair of PDEs

$$\frac{\partial u}{\partial t} = D_u \frac{\partial^2 u}{\partial x^2} + f(u, v), \tag{4a}$$

$$\frac{\partial v}{\partial t} = D_v \frac{\partial^2 v}{\partial x^2} - f(u, v), \tag{4b}$$

with

$$f(u, v) = \left(b + \gamma \frac{u^n}{1 + u^n}\right) v - u, \tag{4c}$$

where b is the basal rate of activation and γ is an additional rate of activation depicting positive feedback from u to its own activation. The constant $n \geq 2$ is the so-called "Hill coefficient". Larger values of n result in sharper switching between states.

We also assume Neumann boundary conditions, namely,

$$u_x(0, t) = 0, \quad u_x(L, t) = 0, \quad v_x(0, t) = 0, \quad v_x(L, t) = 0. \tag{4d}$$

This signifies that no material leaks out of the ends of the 1D domain, i.e. that the cell ends are sealed.

Notably, on the timescale of interest (a few seconds), no protein is made or lost, it is merely exchanged between the active and inactive states (see Table 1). This is captured by the model, since it is easy to see that the total amount of protein in the domain is conserved, that is,

$$\text{Mean total concentration} = \frac{1}{L} \int_0^L (u(x, t) + v(x, t))dx = \text{constant} \tag{5}$$

As shown in [24, 25], the following properties are necessary and sufficient to ensure that a unimodal pattern (depicting a polarized distribution) will exist as a nonuniform steady state of the model:

1. There is some range of values $v_1 \leq v \leq v_2$ for which the function $f(u, v)$ has three roots, $u_a < u_m < u_b$. (We refer to this range of v as the bistable regime.)
2. Of these three roots, the outer two (u_a, u_b) are stable fixed points of the spatially homogeneous variant of (4).
3. For some value, v^* in $v_1 \leq v \leq v_2$, there is a change in the sign of the integral

$$\int_{u_a}^{u_b} f(u, v)du.$$

4. The rates of diffusion of u and v are sufficiently different: $D_u \ll D_v$.

It is interesting to contrast the system (4) with a related one consisting of (4a), (4c) and (4d) but with $v \equiv$ constant, that is, with a single bistable reaction-diffusion equation in one variable, u. The latter is known to sustain traveling wave solutions, as

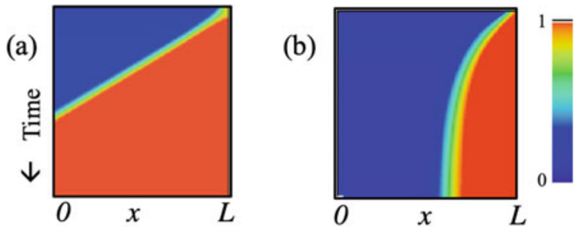


Fig. 4 **Travelling waves versus wave-pinning**: **a** A single reaction-diffusion equation (4a) (for constant v) with kinetics of type (4c) is known to sustain traveling wave solutions for $u(x, t)$. **b** In contrast, the system of Eqs. (4a)–(4d) with conservation and distinct rates of diffusion ($D_u \ll D_v$) results in waves that stop inside the domain, a phenomenon we termed "wave-pinning"

shown in Fig. 4a. In contrast, the two-variable system (4a)–(4d) leads to waves that decelerate and stop inside the domain (once the sign condition above is satisfied) as demonstrated in Fig. 4b. We refer to this behaviour as "wave-pinning". We see that Fig. 4a fails to explain polarization, because the cell diameter is eventually uniformly active. Figure 4b is consistent with polarization, since the two ends of the domain develop distinct levels of activity as time goes by. In this sense, wave-pinning is a simple caricature of cell polarization.

2.1 How Wave-Pinning Works

Full details of the analysis of such dynamics are described in [25]. Here it suffices to briefly mention the key asymptotic analysis ideas used in establishing the result.

The system (4) is rescaled to exploit the existence of a small parameter

$$\epsilon^2 = \frac{D_u}{rL^2},$$

where r is a typical kinetics rate constant with units of 1/time (e.g., $r = \gamma$). We then examine the short and intermediate time-scales of the rescaled system.

On a short time-scale ($t_s = t/\epsilon$), it can be shown that to leading order, at various sites in the domain, u approaches its steady state values u_a, u_b. This means that the domain is "carved up" into plateaus of high and of low activity levels u separated by transition layers between them.

To make progress, we consider the case of a single interface separating a low and a high plateau. Let the position of the interface be $\phi(t)$. We go on to seek the intermediate time scale behaviour. We construct an inner and an outer solution next to the transition layer and show that, to leading order, the variable v is roughly spatially constant on the two sides of the interface $v \approx V_0(t)$, while it is depleted in time as u evolves.

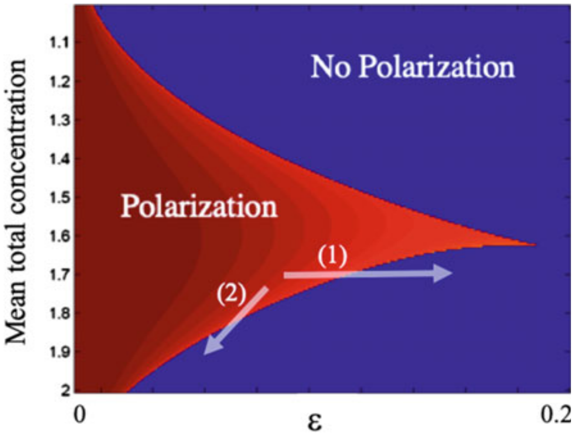


Fig. 5 **Regimes of wave-pinning**: Wave-pinning, which represents cell polarization, depends on a balance between the total amount of GTPase (5) and the size of the small parameter $\epsilon = D_u/(rL^2)$. If the total amount is too small, the wave of activity collapses, whereas if it is too large, the wave sweeps across the entire domain, and a net homogenous state results. Polarization can also be lost in several ways (1) If the cell size decreases too much, and hence ϵ increases, the system leaves the polarization regime. (2) If cell size increases so that the mean total GTPase becomes too "diluted", polarization can also be lost. Image credit: Alexandra JIlkine

Using well-known analysis for wave-speed, we construct the speed of the wave, finding it to be described by a ratio of two integrals

$$\text{speed} = \frac{\int_{u_a}^{u_b} f(u, v)du}{I_2}.$$

Here u_a, u_b depend on $V_0(t)$, and I_2 is a strictly positive integral. We argue that the wave stops when the numerator vanishes, which is guaranteed to happen at some point by Condition 3, a Maxwell condition. Indeed, once v is depleted sufficiently, to the level v^*, the integral in the numerator vanishes. Details and discussion of the steps appear in [25]. Regimes of polarization are shown in (Fig. 5).

Intuitively, the result can be explained as follows: at the transition zone, the high u plateau activates an adjoining site by virtue of local diffusion and positive feedback. The spread of u, however, is at the expense of the inactive form v, which gets depleted as the wave of activity spreads. Once v is sufficiently depleted, the spread of the activity wave can no longer be sustained. At that point, the wave freezes.

It is also interesting to note that the fast diffusion of v means that it acts as a "global messenger" in the sense that it rapidly stores domain-wide information about the level of activity in the cell. Hence, local activation (of u by itself) and global depletion (of v) synergize to produce the polarization of activity in the domain.

3 Recent Work: Analysis, Simulation, and Contact with Experiments

The wave-pinning equations are merely a prototype of the dynamics of a protein in the small GTPase family. Related systems with greater levels of biological detail have also been explored [12, 14, 21]. Indeed insights by AFM Marée in [20] contributed to the understanding that led to the mathematical treatment of wave-pinning in [24, 25].

3.1 Analysis of Slow-Fast Reaction Diffusion Systems: LPA

While studying systems of reaction-diffusion equations (RDEs) for cell polarization, we have benefitted from a number of recent methods that result in shortcuts for quick diagnosis of pattern-formation regimes. Among these, the "Local Perturbation Analysis" (LPA) is a method to track local and global variables in RDEs using ODEs that approximate the fate of a small peak of activity (u_L). This method was invented by AFM Mareé and V Grieneisen [9, 36], and popularized in several papers [11, 12, 15]. It has helped us to identify approximate regimes where a nonuniform pattern could form by a finite perturbation of a spatially uniform state in a fast-slow reaction diffusion system.

Figure 6 illustrates a typical LPA bifurcation result, and its interpretation. The method identifies the existence of a spatially uniform global branch (in black), and parameter regimes where this branch is stable (solid) or unstable (dot-dashed curve). Even when the global homogeneous steady state is stable, a polarized pattern can be established with large enough stimulus. The local variable u_L represents a thin local peak of active u. That peak could grow (and lead to a polar pattern) in the regime where the solid red curve is present. The LPA diagram demonstrates that a sufficiently large stimulus peak is needed, that its size has to exceed a threshold (dashed red curve), and that some parameter regimes allow for patterning in response to arbitrarily small stimuli (dot-dashed black curve). The latter regimes can be identified with Turing instabilities. The former regimes are not discoverable by the usual linear stability analysis (LSA) for Turing pattern formation, and are a helpful aspect of LPA that goes beyond LSA.

In our experience, solving the full PDEs with insights gained from LPA diagrams makes it easier to identify the interesting parameter regimes. Details of the method and its uses has been extensively described in [15]. Other useful shortcuts have included "sharp-switch" approximations (Hill functions replaced by piecewise constant functions), as in [12], and analysis of plateaus described in [36]. None of these replace the need for simulating the PDEs, but all of them help to gain familiarity with possible expected behaviours of the reaction-diffusion systems we have investigated. Most recently, Andreas Buttenschön has created full numerical bifurcation software for PDEs that permits much greater accuracy in tracking solution branches

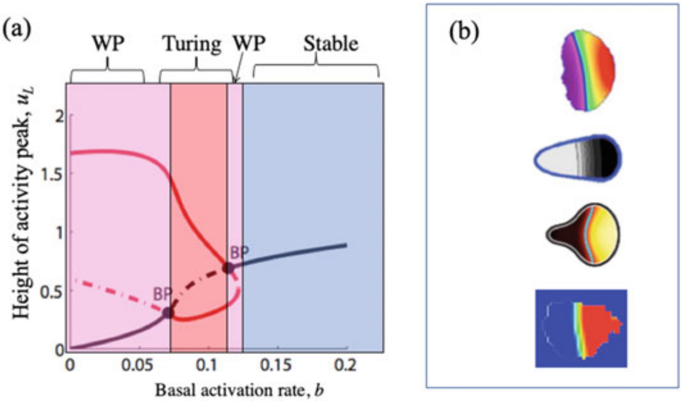


Fig. 6 **Methods of analysis and simulations**: **a** Local perturbation analysis (LPA), a shortcut bifurcation method has helped to detect regimes of patterning in slow-fast reaction-diffusion systems. Here we show an example of how the basal activation rate b influences potential regimes of wave-pinning and of Turing-type instability. See text and references [11, 12, 15] for details. **b** A number of methods have been used to simulate polarization in 2D deforming domains representing the "top-down" view of a cell (as in Fig. 1b). From top to bottom: A cellular-Potts model simulation by A. F. M. Mareé of a 2D deforming cell with an internal reaction-diffusion signaling circuit (and an implicit reaction-diffusion solver) that includes GTPases, interacting lipids, actin, and other components [21], the wave-pinning system (4) solved in an immersed-boundary method simulation by Ben Vanderlei [35], by the level set and moving boundary node method by Zajac [7], and using CompuCell3D by undergraduate summer research student Zachary Pellegrin

[4]. The software builds on state of the art well-conditioned collocation techniques to discretize functions and their operators. Solution branches are continued using a matrix-free Newton-Gauss method, for which rigorous convergence estimates are available.

3.2 *Simulating the PDEs in Dynamic Cell-Shaped Domains*

So far, analytic results were described in 1D domains that represent a cell transect. It is instructive to ask how the same systems behave in domains whose shape more closely relates to that of cells, and in particular, where the internal chemistry affects (and is affected by) the deforming cell. Based on the fact that cell fragments (radius $\approx$ 5–10 μm) without a nucleus, and with overall uniform thickness ($\approx$0.2 µm) are capable of motility, we take the liberty of reducing cell shape to its two-dimensional "top-down" projection shown in Fig. 1b, d. We solve the governing equations (4) or more detailed versions, in the 2D domain, and assume that the boundary of the domain is influenced by the local chemical activity level. For example, if u represents the level of activity of the GTPase Rac, it causes the boundary to be pushed outwards (via F-actin assembly), whereas Rho has the opposite effect (activating contraction via myosin).

A number of results obtained over the years by group members are illustrated in Fig. 6b. In general, we found that the simplest system to understand analytically (4), is not as robust computationally as other variants. Cross-talk between GTPases results in larger parameter regimes for polarization. As an example, models consisting of four PDEs that describe the mutual antagonism between Rac and Rho [12] lead to greater robustness in 2D computations. An even more detailed variant, that includes several GTPases (Rac, Rho, Cdc42), as well as their effects on actin assembly and myosin contraction was capable of realistic behaviour such as directed motility (chemotaxis) [20]. The addition of a layer of signaling lipids (phosphoinositides) also permitted a simulated cell to rapidly select one front despite conflicting or competing stimuli [21].

Simulating the reaction-diffusion systems for GTPase signaling in deforming domains also reveals that evolving domain shape and level curves of the chemical system influence one another: the zero-flux boundary conditions impose constraints on the level curves that also accelerate the dynamics of the chemical redistribution when the domain deforms. Such findings were discussed in detail in [21].

For practical reasons, it is harder to simulate the same systems in 3D. However, recent work by the group of Anotida Madzvamuse [5] has extended these results to a coupled bulk-surface wave-pinning computation in a 3D cell-shaped static domain.

3.3 Contact with Biological Experiments

While details are beyond the scope of this summary, it is worth noting several directions in which the mathematical modeling has contributed to understanding of experimental cell biology.

Willian Bement (U Wisconsin) studies the patterns of GTPases (Rho and Cdc42) that form spontaneously around sites of laser-inflicted wounds in frog eggs (Xenopus oocytes). The connectivity of these GTPases, and their crosstalk with proteins that activate or inactivate them (e.g. Abr) has been modeled by group members, including Cory Simon, Laura Liao, and William R Holmes. Combining models with experiments has helped to build an understanding of the biology [12, 13, 32].

The polarization of HeLa cells exposed to gradients that stimulate a graded response by the GTPase Rac were studied experimentally by Benjamin Lin, in the Lab of Andre Levchenko [19]. A model for Cdc42, Rac, and Rho, interacting with one another and with the phosphoinositides PIP, PIP$_2$ and PIP$_3$ explained the timing and strength of the response, and predicted results of experimental manipulations that affect parts of the crosstalk [14, 19].

Experiments have been carried out on melanoma cells grown on microfabricated surfaces that mimic the natural environment of cells (“extracellular matrix”). JinSeok Park, of the Levchenko Lab at Yale University found three typical motility phenotypes, including persistently polarized, random, and oscillatory front-back cycling,

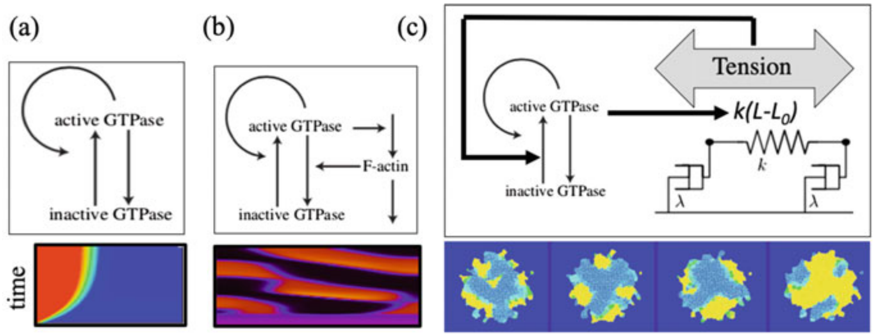


Fig. 7 Extensions of the minimal model: **a** The simplest basic wave-pinning model of Eq. (4) can produce a polarized pattern. **b** When the GTPase promotes assembly of F-actin, which then promotes GTPase inactivation, waves and other exotic dynamics can be observed, provided the negative feedback is on a slow time-scale [10, 22]. In **a**, **b** time increases along the vertical axis and space is on the horizontal axis. **c** Some GTPases cause the cell to spread (Rac) or to shrink (Rho), affecting cell tension. If the tension also affects GTPase activity, interesting dynamics are observed. Shown is a time sequence (left to right) of a "tissue" composed of ≈370 cells, colour coded by their internal GTPase activity. The cell size is correlated to that activity, as described in [37]

depending on levels of adhesion to the substrate, and manipulations that affect activities of the GTPases or their downstream targets. We were able to account for the observed phenotypes by a model for Rac-Rho mutual antagonism, weighted by signals from the extracellular matrix substrate [16, 26, 29].

4 Extending the Minimal Model

The wave-pinning model has been used as a nucleus from which we have expanded to larger circuits, and greater levels of biological detail. We showed that some properties of the system (4) is shared by a circuit of the mutually antagonistic GTPases Rac-Rho [12]. A notable common feature is the existence of parameter regimes in which several states coexist. These include states of uniformly low activity, uniformly high activity, or polarized levels of activity. Which of these develops then depends on initial conditions. A recent contribution [38] extends these findings to more general model variants.

A hallmark of the kinetics we described above is the presence of bistability in some parameter regimes, i.e. the existence of two stable steady states separated by an unstable one. Such systems also display hysteresis, or a kind of history-dependence: slowly increasing a parameter results in a sudden appearance of a new steady state at some transition point, but to reverse the process, the same parameter has to be decreased much beyond the transition point. The addition of feedback from a third dynamic variable in such cases, is known to produce the possibility of oscillations.

We examined several cases of this type, motivated by biological observations. In one case, we studied feedback from F-actin to the inactivation of a GTPase, as observed, for example, in [31]. Assuming slow negative feedback from F-actin (to the inactivation of the GTPase), as shown in Fig. 7b leads to interesting dynamics of traveling waves and pulses in the domain [10, 22]. Feedback between the Rac-Rho circuit and the extracellular matrix also results in oscillations, as previously described [16]. More recently, we also modeled the interplay between mechanical tension in the cell and the activity of GTPases, as observed experimentally by [17]. Here we assumed that GTPase such as Rho and Rac can affect cell spreading, which changes the tension on the cell and feeds back to the activation of the GTPase. A typical circuit of this type is shown in Fig. 7c. As expected, such negative feedback is also consistent with regimes of oscillatory dynamics in individual cells, as demonstrated in [37]. Moreover, when cells with such behaviour are coupled to one another in 1D or in 2D (simulations in Fig. 7c), one observes waves of chemical activity coupled to cell-size changes as the "model tissue" undergoes the spatio-temporal dynamics so created.

5 Discussion

Cell biology presents an unlimited source of inspiring problems. The links between mathematics and cell biology are relatively recent, and not yet fully recognized. But the need for quantitative methods, computational platforms, and mathematical analysis of cellular phenomena promises to grow with time, presenting many opportunities for young applied mathematicians looking for problems to study.

Here I have mainly described a toy model that we constructed to help us understand cell polarization. The simplicity of the model made it mathematically tractable. Its analysis reveals several insights that were not a priori evident. First, with the right kind of positive feedback, we showed that a single GTPase could, on its own, lead to spontaneous polarization that explains cell directionality. In other words, it is not essential to have networks of such proteins to achieve this cellular process. Second, there is a functional purpose for the curious biology of GTPases: their cycling between membrane and cytosol is not a mere evolutionary artifact. We argue that this transition sets up the differences in diffusion between active and inactive GTPases—a difference that is crucial for polarization to be possible, according to our mathematical model.

The motivation of cell polarity led us to mathematics with a surprising twist, uncovering the phenomenon of decelerating waves and wave-pinning that were not widely recognized before in the literature on reaction-diffusion systems. From this standpoint, we could argue that biology inspires new mathematics. The efforts to understand models that were so developed also resulted in a variety of methods that ease the analysis, among them LPA. Extensions of the basic wave-pinning model led to variants with more exotic patterns and waves. These were investigated in various geometries, in single cells, and finally, in interacting groups of cells to identify causes for cell size fluctuations in a tissue and for a variety of emergent phenomena in single and collective cell motility. Finally, developing simple theoretical models and in parallel considering biologically-inspired detailed models are not mutually exclusive. Our experience in the former helps us with the later, and vice versa.

Many still-unanswered questions can be posed. Among these are some of the following: How does the internal GTPase state of a cell affect the outcome of interactions between cells, and how does contact between cells change their GTPase state? What are reasonable ways to model such cell-cell interactions leading to cell adhesion or cell separation? How is cell state coordinated in a multicellular tissue? What aspects of cell adhesion, mechanics, deformation, chemical secretion, and environmental topography (to name a few) affect and are affected by GTPase activities, and how should these be modelled? What methods of analysis can we develop to help with larger, more realistic models that have many interacting components? What aspects of 3D cell shape, and of cell motion in a 3D matrix lead to new phenomena, and what numerical methods should be developed to address such behaviours? Is there a compromise between large-scale computations and mathematical analysis in these more challenging scenarios? In conclusion, the motility and interactions of cells is a rich scientific area calling for investigation by applied mathematicians. Pattern formation inside living cells is merely one facet, while many other fundamental challenges are at hand.

Acknowledgements LEK gratefully acknowledges the contributions of many group members to this research over the years. Among these, special thanks go to A. F. M. Mareé, Y. Mori, W. R. Holmes, A. Jilkine, A. T. Dawes, C. Zmurchok, A. Buttenschön and E. G. Rens. LEK is supported by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Berger, M., Goodman, J.: Airburst-generated tsunamis. Pure Appl. Geophys. **175**(4), 1525–1543 (2018)
2. Burridge, K., Doughman, R.: Front and back by Rho and Rac. Nat. Cell Biol. **8**(8), 781 (2006)
3. Burridge, K., Wennerberg, K.: Rho and Rac take center stage. Cell **116**(2), 167–179 (2004)
4. Buttenschön, A.: Reaction-diffusion bifurcation methods (in preparation) (2021)
5. Cusseddu, D., Edelstein-Keshet, L., Mackenzie, J.A., Portet, S., Madzvamuse, A.: A coupled bulk-surface model for cell polarisation. J. Theoret. Biol. **481**, 119–135 (2019)
6. Dawes, A.T., Edelstein-Keshet, L.: Phosphoinositides and Rho proteins spatially regulate actin polymerization to initiate and maintain directed movement in a one-dimensional model of a motile cell. Biophys. J. **92**(3), 744–768 (2007)
7. Edelstein-Keshet, L., Holmes, W.R., Zajac, M., Dutot, M.: From simple to detailed models for cell polarization. Philos. Trans. R. Soc. B Biol. Sci. **368**(1629), 20130003 (2013)
8. Etienne-Manneville, S., Hall, A.: Rho GTPases in cell biology. Nature **420**(6916), 629 (2002)
9. Grieneisen, V.: Dynamics of auxin patterning in plant morphogenesis. Ph.D. thesis. University of Utrecht (2009)
10. Holmes, W., Carlsson, A., Edelstein-Keshet, L.: Regimes of wave type patterning driven by refractory actin feedback: transition from static polarization to dynamic wave behaviour. Phys. Biol. **9**(4), 046005 (2012)

11. Holmes, W.R.: An efficient, nonlinear stability analysis for detecting pattern formation in reaction diffusion systems. Bull. Math. Biol. **76**(1), 157–183 (2014)
12. Holmes, W.R., Edelstein-Keshet, L.: Analysis of a minimal Rho-GTPase circuit regulating cell shape. Phys. Biol. **13**, 046001 (2016)
13. Holmes, W.R., Liao, L., Bement, W., Edelstein-Keshet, L.: Modeling the roles of protein kinase Cβ and η in single-cell wound repair. Mol. Biol. Cell **26**(22), 4100–4108 (2015)
14. Holmes, W.R., Lin, B., Levchenko, A., Edelstein-Keshet, L.: Modelling cell polarization driven by synthetic spatially graded Rac activation. PLoS Comput. Biol. **8**(6), e1002366 (2012). https://doi.org/10.1371/journal.pcbi.1002366
15. Holmes, W.R., Mata, M.A., Edelstein-Keshet, L.: Local perturbation analysis: a computational tool for biophysical reaction-diffusion models. Biophys. J. **108**(2), 230–236 (2015)
16. Holmes, W.R., Park, J., Levchenko, A., Edelstein-Keshet, L.: A mathematical model coupling polarity signaling to cell adhesion explains diverse cell migration patterns. PLoS Comput. Biol. **13**(5), e1005524 (2017)
17. Houk, A.R., Jilkine, A., Mejean, C.O., Boltyanskiy, R., Dufresne, E.R., Angenent, S.B., Altschuler, S.J., Wu, L.F., Weiner, O.D.: Membrane tension maintains cell polarity by confining signals to the leading edge during neutrophil migration. Cell **148**(1–2), 175–188 (2012)
18. Jilkine, A., Marée, A.F., Edelstein-Keshet, L.: Mathematical model for spatial segregation of the Rho-Family GTPases based on inhibitory crosstalk. Bull. Math. Biol. **69**, 1943–1978 (2007)
19. Lin, B., Holmes, W.R., Wang, J., Ueno, T., Harwell, A., Edelstein-Keshet, L., Takanari Inoue, A.L.: Synthetic spatially graded Rac activation drives directed cell polarization and locomotion. PNAS **109**(52), E3668–E3677 (2012)
20. Marée, A.F., Jilkine, A., Dawes, A., Grieneisen, V.A., Edelstein-Keshet, L.: Polarization and movement of keratocytes: a multiscale modelling approach. Bull. Math. Biol. **68**, 1169–1211 (2006)
21. Marée, A.F.M., Grieneisen, V.A., Edelstein-Keshet, L.: How cells integrate complex stimuli: the effect of feedback from phosphoinositides and cell shape on cell polarization and motility. PLoS Comput. Biol. **8**, e1002402 (2012)
22. Mata, M.A., Dutot, M., Edelstein-Keshet, L., Holmes, W.R.: A model for intracellular actin waves explored by nonlinear local perturbation analysis. J. Theoret. Biol. **334**, 149–161 (2013)
23. Mogilner, A., Oster, G.: Cell motility driven by actin polymerization. Biophys. J. **71**(6), 3030–3045 (1996)
24. Mori, Y., Jilkine, A., Edelstein-Keshet, L.: Wave-pinning and cell polarity from a bistable reaction-diffusion system. Biophys. J. **94**(9), 3684–3697 (2008)
25. Mori, Y., Jilkine, A., Edelstein-Keshet, L.: Asymptotic and bifurcation analysis of wave-pinning in a reaction-diffusion model for cell polarization. SIAM J. Appl. Math. **71**, 1401–1427 (2011)
26. Park, J., Holmes, W.R., Lee, S.H., Kim, H.N., Kim, D.H., Kwak, M.K., Wang, C.J., Edelstein-Keshet, L., Levchko, A.: A mechano-chemical feedback underlies co-existence of qualitatively distinct cell polarity patterns within diverse cell populations. PNAS **114**(28), E5750–59 (2017)
27. Pollard, T.D., Blanchoin, L., Mullins, R.D.: Actin dynamics. J. Cell Sci. **114**(1), 3 (2001)
28. Postma, M., Bosgraaf, L., Loovers, H.M., Van Haastert, P.J.: Chemotaxis: signalling modules join hands at front and tail. EMBO Rep. **5**(1), 35–40 (2004)
29. Rens, E.G., Edelstein-Keshet, L.: Cellular tango: how extracellular matrix adhesion choreographs Rac-Rho signaling and cell movement. Phys. Biol. **18**, 066005 (2021)
30. Ridley, A.J., Schwartz, M.A., Burridge, K., Firtel, R.A., Ginsberg, M.H., Borisy, G., Parsons, J.T., Horwitz, A.R.: Cell migration: integrating signals from front to back. Science **302**(5651), 1704–1709 (2003)
31. Robin, F.B., Michaux, J.B., McFadden, W.M., Munro, E.M.: Excitable RhoA dynamics drive pulsed contractions in the early *C. elegans* embryo. BioRxiv, p. 076356 (2016)
32. Simon, C.M., Vaughan, E.M., Bement, W.M., Edelstein-Keshet, L.: Pattern formation of Rho GTPases in single cell wound healing. Mol. Biol. Cell **24**(3), 421–432 (2013)
33. Small, J.V., Resch, G.P.: The comings and goings of actin: coupling protrusion and retraction in cell motility. Curr. Opin. Cell Biol. **17**(5), 517–523 (2005)

34. Svitkina, T.M., Borisy, G.G.: Arp2/3 complex and actin depolymerizing factor/cofilin in dendritic organization and treadmilling of actin filament array in lamellipodia. J. Cell Biol. **145**(5), 1009–1026 (1999)
35. Vanderlei, B., Feng, J.J., Edelstein-Keshet, L.: A computational model of cell polarization and motility coupling mechanics and biochemistry. Multiscale Model. Simul. **9**(4), 1420–1443 (2011)
36. Walther, G.R., Marée, A.F., Edelstein-Keshet, L., Grieneisen, V.A.: Deterministic versus stochastic cell polarisation through wave-pinning. Bull. Math. Biol. **74**(11), 2570–2599 (2012)
37. Zmurchok, C., Bhaskar, D., Edelstein-Keshet, L.: Coupling mechanical tension and GTPase signaling to generate cell and tissue dynamics. Phys. Biol. **15**(4), 046004 (2018)
38. Zmurchok, C., Holmes, W.R.: Modeling cell shape diversity arising from complex Rho GTPase dynamics. bioRxiv, p. 561373 (2019)

Private AI: Machine Learning on Encrypted Data

Kristin Lauter

Abstract This paper gives an overview of my Invited Plenary Lecture at the International Congress of Industrial and Applied Mathematics (ICIAM) in Valencia in July 2019.

1 Motivation: Privacy in Artificial Intelligence

These days more and more people are taking advantage of cloud-based artificial intelligence (AI) services on their smart phones to get useful predictions such as weather, directions, or nearby restaurant recommendations based on their location and other personal information and preferences. The AI revolution that we are experiencing in the high tech industry is based on the following value proposition: you input your private data and agree to share it with the cloud service in exchange for some useful prediction or recommendation. In some cases the data may contain extremely personal information, such as your sequenced genome, your health record, or your minute-to-minute location.

This quid pro quo may lead to the unwanted disclosure of sensitive information or an invasion of privacy. Examples during the year of ICIAM 2019 include the case of the Strava fitness app which revealed the location of U.S. army bases world-wide, or the case of the city of Los Angeles suing IBM's weather company over deceptive use of location data. It is hard to quantify the potential harm from loss of privacy, but employment discrimination or loss of employment due to a confidential health or genomic condition are potential undesirable outcomes. Corporations also have a need to protect their confidential customer and operations data while storing, using, and analyzing it.

To protect privacy, one option is to lock down personal information by encrypting it before uploading it to the cloud. However, traditional encryption schemes do not allow for any computation to be done on encrypted data. In order to make useful

K. Lauter (✉)
Cryptography and Privacy Research, Microsoft Research, Redmond, USA
e-mail: klauter@microsoft.com

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_6

predictions, we need a new kind of encryption which maintains the structure of the data when encrypting it so that meaningful computation is possible. Homomorphic encryption allows us to switch the order of encryption and computation: we get the same result if we first encrypt and then compute, as if we first compute and then encrypt.

The first solution for a homomorphic encryption scheme which can process any circuit was proposed in 2009 by Gentry [21]. Since then, many researchers in cryptography have worked hard to find schemes which are both practical and also based on well-known hard math problems. In 2011, my team at Microsoft Research collaborated on the homomorphic encryption schemes [8, 9] and many practical applications and improvements [30] which are now widely used in applications of Homomorphic Encryption. Then in 2016, we had a surprise breakthrough at Microsoft Research with the now widely cited CryptoNets paper [22], which demonstrated for the first time that evaluation of neural network predictions was possible on encrypted data.

Thus began our Private AI project, the topic of my Invited Plenary Lecture at the International Congress of Industrial and Applied Mathematics in Valencia in July 2019. Private AI refers to our Homomorphic Encryption-based tools for protecting the privacy of enterprise, customer, or patient data, while doing Machine Learning (ML)-based AI, both learning classification models and making valuable predictions based on such models.

You may ask, “What is Privacy?” Preserving “Privacy” can mean different things to different people or parties. Researchers in many fields including social science and computer science have formulated and discussed definitions of privacy. My favorite definition of privacy is: a person or party should be able to control how and when their data is used or disclosed. This is exactly what Homomorphic Encryption enables.

1.1 Real-World Applications

In 2019, the British Royal Society released a report on Protecting privacy in practice: Privacy Enhancing Technologies in data analysis. The report covers Homomorphic Encryption (HE) and Secure Multi-Party Computation (MPC), but also technologies not built with cryptography, including Differential Privacy (DP) and secure hardware hybrid solutions. Our homomorphic encryption project was featured as a way to protect “Privacy as a human right” at the Microsoft Build world-wide developers conference in 2018 [39]. Private AI forms one of the pillars of Responsible ML in our collection of Responsible AI research and Private Prediction notebooks were released in Azure ML at Build 2020.

Over the last 8 years, my team has created demos of Private AI in action, running private analytics services in the Azure cloud. I showed a few of these demos in my talk at ICIAM in Valencia. Our applications include an encrypted fitness app, which is a cloud service which processes all your workout and fitness data and locations in the cloud in encrypted form, and displays your summary statistics to you on your phone after decrypting the results of the analysis locally. Another application shows an encrypted weather prediction app, which takes your encrypted zip-code and returns encrypted versions of the weather at your location to be decrypted and displayed to you on your phone. The cloud service never learns your location or what weather data was returned to you. Finally, I showed a private medical diagnosis application, which uploads an encrypted version of your Chest X-Ray image, and the medical condition is diagnosed by running image recognition algorithms on the encrypted image in the cloud, and returned in encrypted form to the doctor.

Over the years, my team[1] has developed other Private AI applications, enabling private predictions such as sentiment analysis in text, cat/dog image classification, heart attack risk based on personal health data, neural net image recognition of hand-written digits, flowering time based on the genome of a flower, and pneumonia mortality risk using intelligible models. All of these operate on encrypted data in the cloud to make predictions, and return encrypted results in a matter of fractions of a second.

Many of these demos and applications have been inspired by collaborations with researchers in Medicine, Genomics, Bioinformatics, and Machine Learning. We have worked together with finance experts and pharmaceutical companies to demonstrate a range of ML algorithms operating on encrypted data. The UK Financial Conduct Authority (FCA) ran an international Hackathon in August 2019 to combat money-laundering with encryption technologies by allowing banks to share confidential information with each other. Since 2015, the annual iDASH competition has attracted teams from around the world to submit solutions to the Secure Genome Analysis Competition. Participants include researchers at companies such as Microsoft and IBM, start-up companies, and academics from the U.S., Korea, Japan, Switzerland, Germany, France, etc. The results provide benchmarks for the medical research community of the performance of encryption tools for preserving privacy of health and genomic data.

2 What Is Homomorphic Encryption?

I could say, "Homomorphic Encryption is encryption which is homomorphic." But that is not very helpful without further explanation. Encryption is one of the building blocks of cryptography: encryption protects the confidentiality of information. In mathematical language, encryption is just a map which transforms plaintexts (unencrypted data) into ciphertexts (encrypted data), according to some recipe. Examples of encryption include blockciphers, which take sequences of bits and process them in blocks, passing them through an S-box which scrambles them, and iterating that process many times. A more mathematical example is RSA encryption, which raises

[1] My collaborators on the SEAL team include: Kim Laine, Hao Chen, Radames Cruz, Wei Dai, Ran Gilad-Bachrach, Yongsoo Song, Shabnam Erfani, Sreekanth Kannepalli, Jeremy Tieman, Tarun Singh, Hamed Khanpour, Steven Chith, James French, with substantial contributions from interns Gizem Cetin, Kyoohyung Han, Zhicong Huang, Amir Jalali, Rachel Player, Peter Rindal, Yuhou Xia as well.

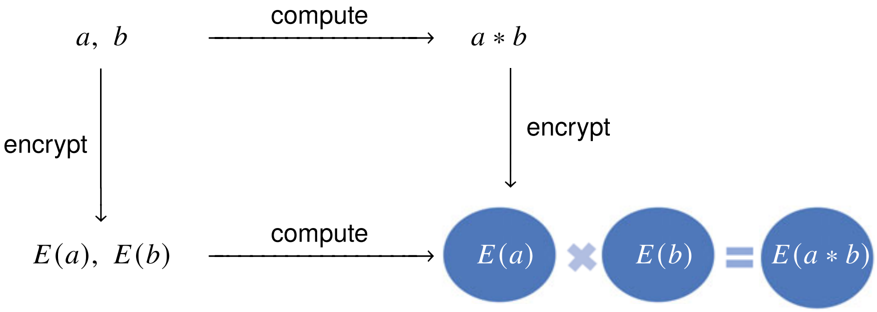


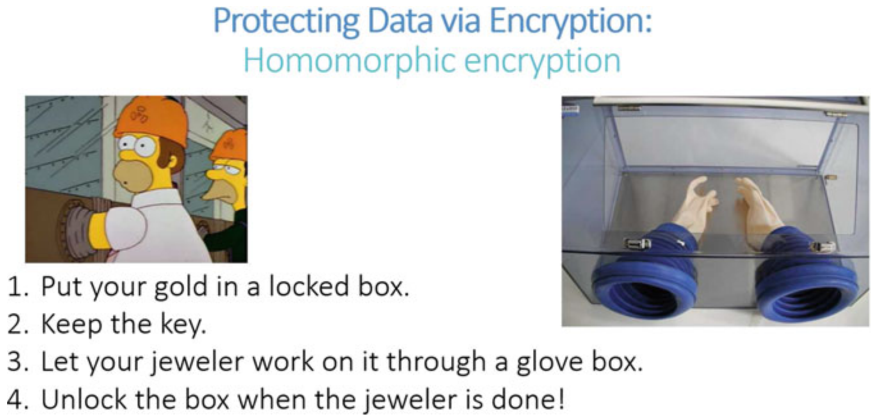
Fig. 1 Homomorphic encryption

a message to a certain power modulo a large integer N, whose prime factorization is secret, $N = p \cdot q$, where p and q are large primes of equal size with certain properties.

A map which is *homomorphic* preserves the structure, in the sense that an operation on plaintexts should correspond to an operation on ciphertexts. In practice that means that switching the order of operations preserves the outcome after decryption: i.e. *encrypt-then-compute* and *compute-then-encrypt* give the same answer. This property is described by the following diagram:

Starting with two pieces of data, a and b, the functional outcome should be the same when following the arrows in either direction, across and then down (*compute-then-encrypt*), or down and then across (*encrypt-then-compute*): $E(a + b)\ E(a) + E(b)$. If this diagram holds for two operations, addition and multiplication, then any circuit of AND and OR gates encrypted under map the encryption map E. It is important to note that homomorphic encryption solutions provide for randomized encryption, which is an important property to protect against so-called dictionary attacks. This means that new randomness is used each time a value is encrypted, and it should not be computationally feasible to detect whether two ciphertexts are the encryption of the same plaintext or not. Thus the ciphertexts in the bottom right corner of the diagram need to be decrypted in order to detect whether they are equal.

The above description gives a mathematical explanation of homomorphic encryption by defining its properties. To return to the motivation of Private AI, another way to describe homomorphic encryption is to explain the functionality that it enables. Figure 2 shows Homer-morphic encryption, where Homer Simpson is a jeweler tasked with making jewelry given some valuable gold. Here the gold represents some private data, and making jewelry is analogous to analyzing the data by applying some AI model. Instead of accessing the gold directly, the gold remains in a locked box, and the owner keeps the key to unlock the box. Homer can only handle the gold through gloves inserted in the box (analogous to handling only encrypted data). When Homer completes his work, the locked box is returned to the owner who unlocks the box to retrieve the jewelry.



Protecting Data via Encryption:
Homomorphic encryption

1. Put your gold in a locked box.
2. Keep the key.
3. Let your jeweler work on it through a glove box.
4. Unlock the box when the jeweler is done!

Fig. 2 Homer-morphic encryption

To connect to Fig. 1 above, outsourcing sensitive work to an untrusted jeweler (cloud) is like following the arrows down, across, and then up. First the data owner encrypts the data and uploads it to the cloud, then the cloud operates on the encrypted data, then the cloud returns the output to the data owner to decrypt.

2.1 History

Almost 5 decades ago, we already had an example of encryption which is homomorphic for one operation: the RSA encryption scheme [36]. A message m is encrypted by raising it to the power e modulo N for fixed integers e and N. Thus the product of the encryption of two messages m_1 and m_2 is $m_1^e m_2^e = (m_1 m_2)^e$. It was an open problem for more than thirty years to find an encryption scheme which was homomorphic with respect to two (ring) operations, allowing for the evaluation of any circuit. Boneh-Goh-Nissim [3] proposed a scheme allowing for unlimited additions and one multiplication, using the group of points on an elliptic curve over a finite field, along with the Weil pairing map to the multiplicative group of a finite field.

In 2009, Gentry proposed the first homomorphic encryption scheme, allowing in theory for evaluation of arbitrary circuits on encrypted data. However it took several years before researchers found schemes which were implementable, relatively practical, and based on known hard mathematical problems. Today all the major homomorphic encryption libraries world-wide implement schemes based on the hardness of lattice problems. A lattice can be thought of as a discrete linear subspace of Euclidean space, with the operations of vector addition, scalar multiplication, and inner product, and its dimension, n, is the number of basis vectors.

2.2 Lattice-Based Solutions

The high-level idea behind current solutions for homomorphic encryption is as follows. Building on an old and fundamental method of encryption, each message is *blinded*, by adding a random inner product to it: the inner product of a secret vector with a randomly generated vector. Historically, blinding a message with fresh randomness was the idea behind encryption via *one-time pads*, but those did not satisfy the homomorphic property. Taking inner products of vectors is a linear operation, but if homomorphic encryption involved only addition of the inner product, it would be easy to break using linear algebra. Instead, the encryption must also add some freshly generated noise to each blinded message, making it difficult to separate the noise from the secret inner product. The noise, or *error*, is selected from a fairly narrow Gaussian distribution. Thus the hard problem to solve becomes a noisy decoding problem in a linear space, essentially Bounded Distance Decoding (BDD) or a Closest Vector Problem (CVP) in a lattice. Decryption is possible with the secret key, because the decryptor can subtract the secret inner product and then the noise is small and is easy to cancel.

Although the above high-level description was formulated in terms of lattices, in fact the structure that we use in practice is a polynomial ring. A vector in a lattice of n dimensions can be thought of as a monic polynomial of degree n, where the coordinates of the vector are the coefficients of the polynomial. Any number ring is given as a quotient of $\mathbb{Z}[x]$, the polynomial ring with integer coefficients, by a monic irreducible polynomial $f(x)$. The ring can be thought of as a lattice in $\mathbb{R}^n$ when embedded into Euclidean space via the canonical embedding. To make all objects finite, we consider these polynomial rings modulo a large prime q, which is often called the ciphertext modulus.

2.3 Encoding Data

When thinking about practical applications, it becomes clear that real data first has to be embedded into the mathematical structure that the encryption map is applied to, the *plaintext space*, before it is encrypted. This encoding procedure must also be homomorphic in order to achieve the desired functionality. The encryption will be applied to the polynomial ring with integer coefficients modulo q, so real data must be embedded into this polynomial ring.

In a now widely cited 2011 paper, “Can Homomorphic Encryption be Practical?” ([30, Sect. 4.1]), we introduced a new way of encoding real data in the polynomial space which allowed for efficient arithmetic operations on real data, opening up a new direction of research focusing on practical applications and computations. The encoding technique was simple: embed an integer m as a polynomial whose ith coefficient is the ith bit of the binary expansion of m (using the ordering of bits so that the least significant bit is encoded as the constant term in the polynomial).

This allows for direct multiplication of real integers, represented as polynomials, instead of encoding and encrypting data bit-by-bit, which requires a deep circuit just to evaluate simple integer multiplication. When using this approach, it is important to keep track of the growth of the size of the output to the computation. In order to assure correct decryption, we limit the total size of the polynomial coefficients to t. Note that each coefficient was a single bit to start with, and a sum of k of them grows to at most k. We obtain the correct decryption and decoding as long as $q > t > k$, so that the result does not wrap around modulo t.

This encoding of integers as polynomials has two important implications, for performance and for storage overhead. In addition to enabling multiplication of floating point numbers via direct multiplication of ciphertexts (rather than requiring deep circuits to multiply data encoded bit wise), this technique also saves space by packing a large floating point number into a single ciphertext, reducing the storage overhead. These encoding techniques help to squash the circuits to be evaluated, and make the size expansion reasonable. However, they limit the possible computations in interesting ways, and so all computations need to be expressed as polynomials. The key factor in determining the efficiency is the degree of the polynomial to be evaluated.

2.4 *Brakerski/Fan-Vercauteren Scheme (BFV)*

For completeness, I will describe one of the most widely used homomorphic encryption schemes, the Brakerski/Fan-Vercauteren Scheme (BFV) [7, 20], using the language of polynomial rings.

2.4.1 Parameters and Notation

Let $q \gg t$ be positive integers and n a power of 2. Denote $\Delta = \lfloor q/t \rfloor$. Define

$$R = \mathbb{Z}[x]/(x^n + 1),$$

$$R_q = R/qR = (\mathbb{Z}/q\mathbb{Z})[x]/(x^n + 1),$$

and $R_t = \mathbb{Z}/t\mathbb{Z}[x]/(x^n + 1)$, where $\mathbb{Z}[x]$ is the set of polynomials with integer coefficients and $(\mathbb{Z}/q\mathbb{Z})[x]$ is the set of polynomials with integer coefficients in the range $[0, q - 1)$.

In the BFV scheme, plaintexts are elements of R_t, and ciphertexts are elements of $R_q \times R_q$. Let χ denote a narrow (centered) discrete Gaussian error distribution. In practice, most implementations of homomorphic encryption use a Gaussian distribution with standard deviation $\sigma[\chi] \approx 3.2$. Finally, let U_k denote the uniform distribution on $\mathbb{Z} \cap [-k/2, k/2)$.

2.4.2 Key Generation

To generate a public key, pk, and a corresponding secret key, sk, sample $s \leftarrow U_3^n$, $a \leftarrow U_q^n$, and $e \leftarrow \chi^n$. Each of s, a, and e is treated as an element of R_q, where the n coefficients are sampled independently from the given distributions. To form the public key–secret key pair, let

$$\texttt{pk} = ([-(as+e)]_q, a) \in R_q^2, \ \texttt{sk} = s$$

where $[\cdot]_q$ denotes the (coefficient-wise) reduction modulo q.

2.4.3 Encryption

Let $m \in R_t$ be a plaintext message. To encrypt m with the public key $\texttt{pk} = (p_0, p_1) \in R_q^2$, sample $u \leftarrow U_3^n$ and $e_1, e_2 \leftarrow \chi^n$. Consider u and e_i as elements of R_q as in key generation, and create the ciphertext

$$\texttt{ct} = ([\Delta m + p_0 u + e_1]_q, [p_1 u + e_2]_q) \in R_q^2.$$

2.4.4 Decryption

To decrypt a ciphertext $\texttt{ct} = (c_0, c_1)$ given a secret key $\texttt{sk} = s$, write

$$\frac{t}{q}(c_0 + c_1 s) = m + v + bt,$$

where $c_0 + c_1 s$ is computed as an integer coefficient polynomial, and scaled by the rational number t/q. The polynomial b has integer coefficients, m is the underlying message, and v satisfies $\|v\|_\infty \ll 1/2$. Thus decryption is performed by evaluating

$$m = \left\lfloor \frac{t}{q}(c_0 + c_1 s) \right\rceil_t,$$

where $\lfloor \cdot \rceil$ denotes rounding to the nearest integer.

2.4.5 Homomorphic Computation

Next we see how to enable addition and multiplication of ciphertexts. Addition is easy: we define an operation $\oplus$ between two ciphertexts $\texttt{ct}_1 = (c_0, c_1)$ and $\texttt{ct}_2 = (d_0, d_1)$ as follows:

$$\texttt{ct}_1 \oplus \texttt{ct}_2 = ([c_0 + d_0]_q, [c_1 + d_1]_q) \in R_q^2.$$

Denote this homomorphic sum by $\mathtt{ct}_{\text{sum}} = (c_0^{\text{sum}}, c_1^{\text{sum}})$, and note that if

$$\frac{t}{q}(c_0 + c_1 s) = m_1 + v_1 + b_1 t, \quad \frac{t}{q}(d_0 + d_1 s) = m_2 + v_2 + b_2 t,$$

then

$$\frac{t}{q}(c_0^{\text{sum}} + c_1^{\text{sum}} s) = [m_1 + m_2]_t + v_1 + v_2 + b_{\text{sum}} t,$$

As long as $\|v_1 + v_2\|_\infty < 1/2$, the ciphertext $\mathtt{ct}_{\text{sum}}$ is a correct encryption of $[m_1 + m_2]_t$.

Similarly, there is an operation $\otimes$ between two ciphertexts that results in a ciphertext decrypting to $[m_1 m_2]_t$, as long as $\|v_1\|_\infty$ and $\|v_2\|_\infty$ are small enough. Since $\otimes$ is more difficult to describe than $\oplus$, we refer the reader to [20] for details.

2.4.6 Noise

In the decryption formula presented above the polynomial v with rational coefficients is assumed to have infinity-norm less than $1/2$. Otherwise, the plaintext output by decryption will be incorrect. Given a ciphertext $\mathtt{ct} = (c_0, c_1)$ which is an encryption of a plaintext m, let $v \in \mathbb{Q}[x]/(x^n + 1)$ be such that

$$\frac{t}{q}(c_0 + c_1 s) = m + v + bt.$$

The infinity norm of the polynomial v called the noise, and the ciphertext decrypts correctly as long as the noise is less than $1/2$.

When operations such as addition and multiplication are applied to encrypted data, the noise in the result may be larger than the noise in the inputs. This noise growth is very small in homomorphic additions, but substantially larger in homomorphic multiplications. Thus, given a specific set of encryption parameters (n, q, t, χ), one can only evaluate computations of a bounded size (or bounded multiplicative depth).

A precise estimate of the noise growth for the YASHE scheme was given in [4] and these estimates were used in [5] to give an algorithm for selecting secure parameters for performing any given computation. Although the specific noise growth estimates needed for this algorithm do depend on which homomorphic encryption scheme is used, the general idea applies to any scheme.

2.5 Other Homomorphic Encryption Schemes

In 2011, researchers at Microsoft Research and Weizmann Institute published the (BV/BGV [8, 9]) homomorphic encryption scheme which is used by teams around the world today. In 2013, IBM released HELib, a homomorphic encryption library for research purposes, which implemented the BGV scheme. HELib is written in C++ and uses the NTL mathematical library. The Brakerski/Fan-Vercauteren (BFV) scheme described above was proposed in 2012. Alternative schemes with different security and error-growth properties were proposed in 2012 by Lopez-Alt, Tromer, and Vaikuntanathan (LTV [33]), and in 2013 by Bos, Lauter, Loftus, and Naehrig (YASHE [4]). The Cheon-Kim-Kim-Song (CKKS [14]) scheme was introduced in 2016, enabling approximate computation on ciphertexts.

Other schemes [16, 19] for general computation on bits are more efficient for logical tasks such as comparison, which operate bit-by-bit. Current research attempts to make it practical to switch between such schemes to enable both arithmetic and logical operations efficiently ([6]).

2.6 Microsoft SEAL

Early research prototype libraries were developed by the Microsoft Research (MSR) Cryptography group to demonstrate the performance numbers for initial applications such as those developed in [4, 5, 23, 29]. But due to requests from the biomedical research community, it became clear that it would be very valuable to develop a well-engineered library which would be widely usable by developers to enable privacy solutions. The Simple Encrypted Arithmetic Library (SEAL) [37] was developed in 2015 by the MSR Cryptography group with this goal in mind, and is written in C++. Microsoft SEAL was publicly released in November 2015, and was released open source in November 2018 for commercial use. It has been widely adopted by teams worldwide and is freely available online (http://sealcrypto.org).

Microsoft SEAL aims to be easy to use for non-experts, and at the same time powerful and flexible for expert use. SEAL maintains a delicate balance between usability and performance, but is extremely fast due to high-quality engineering. SEAL is extensively documented, and has no external dependencies. Other publicly available libraries include HELib from IBM, PALISADE by Duality Technologies, and HEAAN from Seoul National University.

2.7 Standardization of Homomorphic Encryption [1]

When new public key cryptographic primitives are introduced, historically there has been roughly a 10-year lag in adoption across the industry. In 2017, Microsoft Research Outreach and the MSR Cryptography group launched a consortium for advancing the standardization of homomorphic encryption technology, together with our academic partners, researchers from government and military agencies, and partners and customers from various industries: Homomorphic Encryption.org. The first workshop was hosted at Microsoft in July 2017, and developers for all the existing implementations around the world were invited to demo their libraries.

At the July 2017 workshop, we worked in groups to draft three white papers on Security, Applications, and APIs. We then worked with all relevant stakeholders of the HE community to revise the Security white paper [11] into the first draft standard for homomorphic encryption [1]. The Homomorphic Encryption Standard (HES) specifies secure parameters for the use of homomorphic encryption. The draft standard was initially approved by the HomomorphicEncryption.org community at the second workshop at MIT in March 2018, and then was finalized and made publicly available at the third workshop in October 2018 at the University of Toronto [1]. A study group was initiated in 2020 at the ISO, the International Standards Organization, to consider next steps for standardization.

3 What Kind of Computation Can We Do?

3.1 Statistical Computations

In early work, we focused on demonstrating the feasibility of statistical computations on health and genomic data, because privacy concerns are obvious in the realm of health and genomic data, and statistical computations are an excellent fit for efficient HE because they have very low depth. We demonstrated HE implementations and performance numbers for statistical computations in genomics such as the chi-square test, Cochran-Armitage Test for Trend, and Haplotype Estimation Maximization [29]. Next, we focused on string matching, using the Smith-Waterman algorithm for edit distance [15], another task which is frequently performed for genome sequencing and the study of genomic disease.

3.2 Heart Attack Risk

To demonstrate operations on health data, in 2013 we developed a live demo predicting the risk of having a heart attack based on six health characteristics [5]. We evaluated predictive models developed over decades in the Framingham Heart study, using the Cox proportional Hazard method. I showed the demo live to news reporters at the 2014 AAAS meeting, and our software processed my risk for a heart attack in the cloud, operating on encrypted data, in a fraction of a second.

In 2016, we started a collaboration with Merck to demonstrate the feasibility of evaluating such models on large patient populations. Inspired by our published work on heart attack risk prediction [5], they used SEAL to demonstrate running the heart attack risk prediction on one million patients from an affiliated hospital. Their implementation returned the results for all patients in about 2 h, compared to 10 min for the same computation on unencrypted patient data.

3.3 Cancer Patient Statistics

In 2017, we began a collaboration with a Crayon, a Norwegian company that develops health record systems. The goal of this collaboration was to demonstrate the value of SEAL in a real world working environment. Crayon reproduced all computations in the 2016 Norwegian Cancer Report using SEAL and operating on encrypted inputs. The report processed the cancer statistics from all cancer patients in Norway collected over the last roughly 5 decades.

3.4 Genomic Privacy

Engaging with a community of researchers in bioinformatics and biostatistics who were concerned with patient privacy issues led to a growing interdisciplinary community interested in the development of a range of cryptographic techniques to apply to privacy problems in the health and biological sciences arenas [18]. One measure of the growth of this community over the last five years has been participation in the iDASH Secure Genome Analysis Competition, a series of annual international competitions funded by the National Institutes of Health (NIH) in the U.S. The iDASH competition has included a track on Homomorphic Encryption for the last five years 2015–2019, and our team from MSR submitted winning solutions for the competition in 2015 ([27]) and 2016 ([10]). The tasks were: chi-square test, modified edit distance, database search, training logistic regression models, genotype imputation. Each year, roughly 5–10 teams from research groups around the world submitted solutions for the task, which were bench-marked by the iDASH team. These results provide the biological data science community and NIH with real and evolving measures of the performance and capability of homomorphic encryption to protect the privacy of genomic data sets while in use. Summaries of the competitions are published in [38, 40].

3.5 Machine Learning: Training and Prediction

The 2013 "ML Confidential" paper [23] was the first to propose *training* ML algorithms on homomorphically encrypted data and to show initial performance numbers for simple models such as linear means classifiers and gradient descent. Training is inherently challenging because of the large and unknown amount of data to be processed.

Prediction tasks on the other hand, process an input and model of known size, so many can be processed efficiently. For example, in 2016 we developed a demo using SEAL to predict the flowering time for a flower. The model processed 200, 000 SNPs from the genome of the flower, and evaluated a Fast Linear Mixed Model (LMM).

Including the round-trip communication time with the cloud running the demo as a service in Azure, the prediction was obtained in under a second.

Another demo developed in 2016 using SEAL predicted the mortality risk for pneumonia patients based on 46 characteristics from the medical record for the patient. The model in this case is an example of an intelligible model and consists of 46° 4 polynomials to be evaluated on the patient's data. Data from $4,096$ patients can be batched together, and the prediction for all $4,096$ patients was returned by the cloud service in a few seconds (in 2016).

These two demos evaluated models which were represented by shallow circuits, linear in the first case and degree 4 in the second case. Other models such as deep neural nets (DNNs) are inherently more challenging because the circuits are so deep. To enable efficient solutions for such tasks requires a blend of cryptography and ML research, aimed at designing and testing ways to process data which allow for efficient operations on encrypted data while maintaining accuracy. An example of that was introduced in CryptoNets [22], showing that the activation function in the layers of the neural nets can be approximated with a low-degree polynomial function (x^2) without significant loss of accuracy.

The CryptoNets paper was the first to show the evaluation of a neural net predictions on encrypted data, and used the techniques introduced there to classify hand-written digits from the MNIST [31] data set. Many teams have since worked on improving the performance of CryptoNets, either with hybrid schemes or other optimizations [17, 25, 35]. In 2018, in collaboration with Median Technologies, we demonstrated deep neural net predictions for a medical image recognition task: classification of liver tumors based on medical images.

Returning to the challenge of training ML algorithms, the 2017 iDASH contest task required the teams to train a logistic regression model on encrypted data. The data set provided for the competition was very simple and did not require many iterations to train an effective model (the winning solution used only 7 iterations [26, 28]). The MSR solution [12] computed over 300 iterations and was fully scalable to any arbitrary number of iterations. We also applied our solution to a simplified version of the MNIST data set to demonstrate the performance numbers.

Performance numbers for all computations described here were published at the time of discovery. They would need to be updated now with the latest version of SEAL, or can be estimated. Hardware acceleration techniques using state-of-the-art FPGAs can be used to improve the performance further ([34]).

4 How Do We Assess Security?

The security of all homomorphic encryption schemes described in this article is based on the mathematics of lattice-based cryptography, and the hardness of well-known lattice problems in high dimensions, problems which have been studied for more than 25 years. Compare this to the age of other public key systems such as RSA (1975) or Elliptic Curve Cryptography ECC (1985). Cryptographic applications of Lattice-based Cryptography were first proposed by Hoffstein, Pipher, and Silverman [24] in 1996 and led them to launch the company NTRU. New hard problems such as LWE were proposed in the period of 2004–2010, but were reduced to older problems which had been studied already for several decades: the Approximate Shortest Vector Problem (SVP) and Bounded Distance Decoding.

The best known algorithms for attacking the Shortest Vector Problem or the Closest Vector Problem are called lattice basis reduction algorithms, and they have a more than 30-year history, including the LLL algorithm [32]. LLL runs in polynomial time, but only finds an exponentially bad approximation to the shortest vector. More recent improvements, such as BKZ 2.0 [13], involve exponential algorithms such as sieving and enumeration. Hard Lattice Challenges were created by TU Darmstadt and are publicly available online for anyone to try to attack and solve hard lattice problems of larger and larger size for the record.

Homomorphic Encryption scheme parameters are set such that the best known attacks take exponential time (exponential in the dimension of the lattice, n, meaning roughly 2^n time). These schemes have the advantage that there are no known polynomial time quantum attacks, which means they are good candidates for Post-Quantum Cryptography (PQC) in the ongoing 5-year NIST PQC competition.

Lattice-based cryptography is currently under consideration for standardization in the ongoing NIST PQC Post-Quantum Cryptography competition. Most Homomorphic Encryption deployments use small secrets as an optimization, so it is important to understand the concrete security when sampling the secret from a non-uniform, small distribution. There are numerous heuristics used to estimate the running time and quality of lattice reduction algorithms such as BKZ2.0. The Homomorphic Encryption Standard recommends parameters based on the heuristic running time of the best known attacks, as estimated in the online LWE Estimator [2].

5 Conclusion

Homomorphic Encryption is a technology which allows meaningful computation on encrypted data, and provides a tool to protect privacy of data in use. A primary application of Homomorphic Encryption is secure and confidential outsourced storage and computation in the cloud (i.e. a data center). A client encrypts their data locally, and stores their encryption key(s) locally, then uploads it to the cloud for long-term storage and analysis. The cloud processes the encrypted data without decrypting it, and returns encrypted answers to the client for decryption. The cloud learns nothing about the data other than the size of the encrypted data and the size of the computation. The cloud can process Machine Learning or Artificial Intelligence (ML or AI) computations, either to make predictions based on known models or to train new models, while preserving the client's privacy.

Current solutions for HE are implemented in 5–6 major open source libraries world-wide. The Homomorphic Encryption Standard [1] for using HE securely was approved in 2018 by HomomorphicEncryption.org, an international consortium of researchers in industry, government, and academia.

Today, applied Homomorphic Encryption remains an exciting direction in cryptography research. Several big and small companies, government contractors, and academic research groups are enthusiastic about the possibilities of this technology. With new algorithmic improvements, new schemes, an improved understanding of concrete use-cases, and an active standardization effort, wide-scale deployment of homomorphic encryption seems possible within the next 2–5 years. Small-scale deployment is already happening.

Computational performance, memory overhead, and the limited set of operations available in most libraries remain the main challenges. Most homomorphic encryption schemes are inherently parallelizable, which is important to take advantage of to achieve good performance. Thus, easily parallelizable arithmetic computations seem to be the most amenable to homomorphic encryption at this time and it seems plausible that initial wide-scale deployment may be in applications of Machine Learning to enable Private AI.

Acknowledgements I would like to gratefully acknowledge the contributions of many people in the achievements, software, demos, standards, assets and impact described in this article. First and foremost, none of this software or applications would exist without my collaborators on the SEAL team, including Kim Laine, Hao Chen, Radames Cruz, Wei Dai, Ran Gilad-Bachrach, Yongsoo Song, John Wernsing, with substantial contributions from interns Gizem Cetin, Kyoohyung Han, Zhicong Huang, Amir Jalali, Rachel Player, Peter Rindal, Yuhou Xia as well. The demos described here were developed largely by our partner engineering team in Foundry 99: Shabnam Erfani, Sreekanth Kannepalli, Steven Chith, James French, Hamed Khanpour, Tarun Singh, Jeremy Tieman. I launched the Homomorphic Encryption Standardization process in collaboration with Kim Laine from my team, with Roy Zimmermann and the support of Microsoft Outreach, and collaborators Kurt Rohloff, Vinod Vaikuntanathan, Shai Halevi, and Jung Hee Cheon, and collectively we now form the Steering Committee of HomomorphicEncryption.org. Finally I would like to thank the organizers of ICIAM 2019 for the invitation to speak and to write this article.

References

1. Albrecht, M., Chase, M., Chen, H., Ding, J., Goldwasser, S., Gorbunov, S., Halevi, S., Hoffstein, J., Laine, K., Lauter, K., Lokam, S., Micciancio, Moody, D., Morrison, T., Sahai, A., Vaikuntanathan, V.: Homomorphic encryption security standard. Technical report, HomomorphicEncryption.org, Toronto, Canada, Nov 2018. https://eprint.iacr.org/2019/939
2. Albrecht, M., Player, R., Scott, S.: On the concrete hardness of learning with errors. J. Math. Cryptol. **9**(3), 169–203 (2015)
3. Boneh, D., Goh, E., Nissim, K.: Evaluating 2-dnf formulas on ciphertexts. In: TCC'05: Proceedings of the Second international conference on Theory of Cryptography, vol. 3378. Lecture Notes in Computer Science, pp. 325–341. Springer, Berlin (2005)
4. Bos, J.W., Lauter, K., Loftus, J., Naehrig, M.: Improved security for a ring-based fully homomorphic encryption scheme. In: Cryptography and Coding, pp. 45–64. Springer, Berlin (2013)
5. Bos, J.W., Lauter, K., Naehrig, M.: Private predictive analysis on encrypted medical data. J. Biomed. Inform. **50**, 234–243 (2014)

6. Boura, C., Gama, N., Georgieva, M., Jetchev, D.: Chimera: combining ring-LWE-based fully homomorphic encryption schemes. Cryptology ePrint Archive. https://eprint.iacr.org/2018/758
7. Brakerski, Z.: Fully homomorphic encryption without modulus switching from classical GapSVP. In: Advances in Cryptology–CRYPTO 2012, pp. 868–886. Springer, Berlin (2012)
8. Brakerski, Z., Gentry, C., Vaikuntanathan, V.: (Leveled) fully homomorphic encryption without bootstrapping. In: Proceedings of ITCS, pp. 309–325. ACM (2012)
9. Brakerski, Z., Vaikuntanathan, V.: Efficient fully homomorphic encryption from (standard) LWE. In: 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science, pp. 97–106, Oct 2011
10. Cetin, G.S., Chen, H., Laine, K., Lauter, K., Rindal, P., Xia, Y.: Private queries on encrypted genomic data. BMC Med. Genomics **10**(45) (2017)
11. Chase, M., Chen, H., Ding, J., Goldwasser, S., Gorbunov, S., Hoffstein, J., Lauter, K., Lokam, S., Moody, D., Morrison, T., Sahai, A., Vaikuntanathan, V.: Security of homomorphic encryption. HomomorphicEncryption.org, Redmond WA, Technical report (2017)
12. Chen, H., Gilad-Bachrach, R., Han, K., Huang, Z., Jalali, A., Laine, K., Lauter, K.: Logistic regression over encrypted data from fully homomorphic encryption. BMC Med. Genomics **11**(81) (2018)
13. Chen, Y., Nguyen, P.Q.: BKZ 2.0: better lattice security estimates. In: Lee, D.H., Wang, X. (eds.) Advances in Cryptology—ASIACRYPT 2011, pp. 1–20. Springer, Berlin (2011)
14. Cheon, J.H., Kim, A., Kim, M., Song, Y.: Homomorphic encryption for arithmetic of approximate numbers. In: International Conference on the Theory and Application of Cryptology and Information Security, pp. 409–437. Springer, Berlin (2017)
15. Cheon, J.H., Kim, M., Song, Y.: . Homomorphic computation of edit distance. In: International Conference on Financial Cryptography and Data Security, pp. 194–212. Springer, Berlin (2015)
16. Chillotti, I., Gama, N., Georgieva, M., Izabachène, M.: TFHE: fast fully homomorphic encryption over the torus. J. Cryptol. **33**, 34–91 (2020)
17. Dathathri, R., Saarikivi, O., Chen, H., Laine, K., Lauter, K., Maleki, S., Musuvathi, M., Mytkowicz, T.: CHET: an optimizing compiler for fully-homomorphic neural-network inferencing. In: Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 142–156. ACM (2019)
18. Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M., Wernsing, J.: Manual for using homomorphic encryption for bioinformatics. Proc. IEEE **105**(3), 552–567 (2017)
19. Ducas, L.,Micciancio, D.: FHEW: bootstrapping homomorphic encryption in less than a second. In: Annual International Conference on the Theory and Applications of Cryptographic Techniques, pp. 617–640. Springer, Berlin (2015)
20. Fan, J., Vercauteren, F.: Somewhat practical fully homomorphic encryption. In: IACR Cryptology ePrint Archive 144 (2012). https://eprint.iacr.org/2012/144. Accessed on 9 April 2018
21. Gentry, C.: A fully homomorphic encryption scheme. Stanford University (2009)
22. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., Wernsing, J.: Cryptonets: applying neural networks to encrypted data with high throughput and accuracy. In: International Conference on Machine Learning, pp. 201–210 (2016)
23. Graepel, T., Lauter, K., Naehrig, M.: ML confidential: Machine learning on encrypted data. In: International Conference on Information Security and Cryptology, pp. 1–21. Springer, Berlin (2012)
24. Hoffstein, J., Pipher, J., Silverman, J.H.: NTRU: a ring-based public key cryptosystem. In: Algorithmic number theory (Portland, OR, 1998), vo. 1423. Lecture Notes in Computer Science, pp. 267–288. Springer, Berlin (1998)
25. Juvekar, C., Vaikuntanathan, V., Chandrakasan, A.: GAZELLE: a low latency framework for secure neural network inference. In: 27th USENIX Security Symposium (USENIX Security 18), pp. 1651–1669 (2018)
26. Kim, A., Song, Y., Kim, M., Lee, K., Cheon, J.-H.: Logistic regression model training based on the approximate homomorphic encryption. Cryptology ePrint Archive, Report 2018/254 (2018). https://eprint.iacr.org/2018/254

27. Kim, M., Lauter, K.: Private genome analysis through homomorphic encryption. BMC Med. Inform. Decis. Making **15**(Suppl 5), S3 (2015)
28. Kim, M., Song, Y., Wang, S., Xia, Y., Jiang, X.: Secure logistic regression based on homomorphic encryption. Cryptology ePrint Archive, Report 2018/074 (2018). https://eprint.iacr.org/2018/074
29. Lauter, K., López-Alt, A., Naehrig, M.: Private computation on encrypted genomic data. In: International Conference on Cryptology and Information Security in Latin America, pp. 3–27. Springer, Berlin (2014)
30. Lauter, K., Naehrig, M., Vaikuntanathan, V.: Can homomorphic encryption be practical? In: Proceedings of the 3rd ACM Workshop on Cloud Computing Security Workshop (CCSW '11), New York, NY, USA, pp. 113–124. ACM (2011)
31. LeCun, Y., Cortes, C., Burges, C.J.C.: The MNIST database of handwritten digits (1998). http://yann.lecun.com/exdb/mnist/
32. Lenstra, A.K., Lenstra, H.W., Lovász, L.: Factoring polynomials with rational coefficients. Mathematische Annalen **261**(4), 515–534 (1982)
33. Lopez-Alt, A., Tromer, E., Vaikuntanathan, V.: On-the-fly multiparty computation on the cloud via multikey fully homomorphic encryption. In: Proceedings of STOC, pp. 1219–1234. IEEE Computer Society (2012)
34. Sadegh Riazi, M., Laine, K., Pelton, B., Dai, W.: Heax: high-performance architecture for computation on homomorphically encrypted data in the cloud. *arXiv preprint* arXiv:1909.09731 (2019)
35. Sadegh Riazi, M., Samragh, M., Chen, H., Laine, K., Lauter, K., Koushanfar, F.: XONN: Xnor-based oblivious deep neural network inference. In: 28th USENIX Security Symposium (USENIX Security 19), Santa Clara, CA, pp. 1501–1518. USENIX Association, Aug 2019
36. Rivest, R., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM **21**(2), 120–126 (1978)
37. Microsoft SEAL (release 3.2). https://github.com/Microsoft/SEAL. Microsoft Research, Redmond, WA, Nov 2018
38. Tang, H., Jiang, X., Wang, X., Wang, S., Sofia, H., Fox, D., Lauter, K., Malin, B., Telenti, A., Li, Xi., Ohno-Machado, L.: Protecting genomic data analytics in the cloud: state of the art and opportunities. BMC Med. Genomics **9**(63) (2016)
39. Vanian, J.: 4 Big Takeaways from Satya Nadella's talk at Microsoft Build (2018). https://fortune.com/2018/05/07/microsoft-satya-nadella-build/
40. Wang, S., Jiang, X., Tang, H., Wang, X., Bu, D., Carey, K., Dyke, S.O.M., Fox, D., Jiang, C., Lauter, K., Malin, B., Sofia, H., Telenti, A., Wang, L., Wang, W., Ohno-Machado, L.: A community effort to protect genomic data sharing, collaboration and outsourcing. NPJ Genomic Med. **2**(33) (2017)

Mathematical Approaches for Contemporary Materials Science: Addressing Defects in the Microstructure

Claude Le Bris

Abstract We overview a series of mathematical works that introduce new modeling and computational approaches for non-periodic materials and media. The approaches consider various types of defects embedded in a periodic structure, which can be either deterministic or random in nature. A portfolio of possible computational techniques addressing the identification of the homogenized properties of the material or the determination of the actual multi-scale solution is presented.

1 Introduction

1.1 Contemporary Materials Science

The works outlined in the present review have been motivated by the following two-fold observation. In the past couple of decades, what we believe to be the most spectacular changes in materials science are

(i) **the increasing multi-scale nature of the materials considered**: materials used to be mostly considered at *one single* scale, the effect of the finer scales being only *phenomenologically* accounted for in the model at the largest scale; when absolutely necessary, the effect of some micro-scale structure was explicitly considered, but then it was at most for one such scale and almost exclusively *sequentially*: information was passed from the micro-scale to the macro-scale; modern materials science increasingly *explicitly and concurrently* considers models of a given material at *many* different scales.

(ii) **the increasing imperfect character of the materials considered**: more and more often, deterministic or random sources of disorder are considered within an ordered phase: the simplicity of periodic structures is not a valid approximation any longer for the degree of practical relevance and accuracy that modern materials science requires; crystalline materials are actually polycrystalline materials

C. Le Bris (✉)
Ecole des Ponts and Inria, Paris, France
e-mail: claude.le-bris@enpc.fr

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_7

and consist of mono-crystalline grains, each of them possibly of a different crystalline structure, each crystalline structure being itself flawed because sprinkled of defects and dislocations; the imperfections, or violations of periodicity, affect every possible scale, and actually cut through scales.

As a result, the real materials that contemporary materials scientists have to model have a *multi-scale, imperfect, possibly random* nature. Such materials have several characteristic length-scales that possibly differ from one another by orders of magnitude but must be accounted for simultaneously. At possibly each such scale, they have defects. Their qualitative and quantitative response might therefore differ a lot from the idealized scenario long considered.

Our intent here is to present several mathematical and numerical endeavors that aim to better model, understand and simulate non-periodic multi-scale problems.

The specific theoretical context in which we develop our discussion is homogenization of simple, second order elliptic equations in divergence form with highly oscillatory coefficients:

$$-\operatorname{div}\left[A_\varepsilon(x)\nabla u^\varepsilon\right]=f, \tag{1}$$

in a domain $\mathcal{D}\subset\mathbb{R}^d$, with, say, homogeneous Dirichlet boundary conditions $u^\varepsilon=0$ on $\partial\mathcal{D}$. This particular case is to be thought of as a prototypical case. It is intuitively clear that the same approaches carry over to other settings. Current works are indeed directed toward extending many of the considerations here to other types of equations, as will be clear in the exposition below.

We conclude this introductory section with a quick presentation of the classical theory. The reader familiar with this theory may of course skip the presentation and directly proceed to Sect. 2.

1.2 Basics of Homogenization Theory

1.2.1 Periodic Homogenization

To begin with, we recall some well known, basic ingredients of elliptic homogenization theory in the periodic setting, *see* the classical references [8, 29, 42] for more details, or an overview in [1, Chap. 1] . We consider the problem

$$\begin{cases}-\operatorname{div}\left[A_{per}\left(\frac{x}{\varepsilon}\right)\nabla u^\varepsilon\right]=f \quad \text{in} \quad \mathcal{D},\\ u^\varepsilon=0 \quad \text{on} \quad \partial\mathcal{D},\end{cases} \tag{2}$$

where the matrix A_{per} is $\mathbb{Z}^d$-periodic, bounded and bounded away from zero, and (for simplicity) symmetric. The corrector problem associated to Eq. 2 reads, for $\mathbf{p}$ fixed in $\mathbb{R}^d$,

$$\begin{cases}-\operatorname{div}\left(A_{per}(y)\left(\mathbf{p}+\nabla w_{per,\mathbf{p}}\right)\right)=0,\\ w_{per,\mathbf{p}} \text{ is } \mathbb{Z}^d\text{-periodic}.\end{cases} \tag{3}$$

It has a unique solution up to the addition of a constant. This solution is meant to describe prototypical fine oscillations of the exact solution u^ε for ε small. Then, the homogenized coefficients read

$$[A^*_{per}]_{ij} = \int_Q \mathbf{e}_i^T A_{per}(y) \left(\mathbf{e}_j + \nabla w_{per,\mathbf{e}_j}(y)\right) dy, \tag{4}$$

where Q is the unit cube and $\mathbf{e}_i$, $1 \le i \le d$ are the canonical vectors of $\mathbb{R}^d$. The main result of periodic homogenization theory for Eq. 2 is that, as ε vanishes, the solution u^ε to Eq. 2 converges to u^* solution to

$$\begin{cases} -\mathrm{div}\left[A^*_{per} \nabla u^*\right] = f & \text{in} \quad \mathcal{D}, \\ u^* = 0 \quad \text{on} \quad \partial\mathcal{D}. \end{cases} \tag{5}$$

The convergence holds in $L^2(\mathcal{D})$, and weakly in $H^1_0(\mathcal{D})$. The correctors $w_{per,\mathbf{e}_i}$ may then also be used to "correct" u^* in order to show that, in the strong topology $H^1(\mathcal{D})$, $u^\varepsilon - u^{\varepsilon,1}(x)$ converges to zero, for $u^{\varepsilon,1}(x) = u^*(x) + \varepsilon \sum_{i=1}^d \partial_{x_i} u^*(x)\, w_{per,\mathbf{e}_i}(x/\varepsilon)$. The rate of convergence may also be made precise.

The practical conclusion is that, at the price of only computing the d periodic problems of Eq. 3, the solution to Eq. 2 can be efficiently approached for ε small.

1.2.2 Random Homogenization

A first option to outreach the simplistic setting of periodic structures is to consider random structures. Of course, materials are never random in nature, but randomness is a suitable, practical way to encode the ignorance of, or at best the uncertainty on the intimate microscopic structure of the material considered.

For homogenization, the random setting is a highly non trivial extension of the periodic setting. Many questions, in particular for nonlinear equations, still remain open in the random case although they are solved and well documented in the periodic case. Fortunately, in the case of linear diffusion equations such as Eq. 1, the state of affairs is that, loosely speaking, all the results of convergence still essentially hold true but (a) they are more difficult to prove and (b) the convergence rates are even more difficult to establish.

To fix the ideas, we now give some more formal details on one random case. For brevity, we skip all technicalities related to the definition of the probabilistic setting, which we assume discrete stationary and ergodic (we refer e.g. to [2] for all details). We now fix $A(., \omega)$ a square matrix of size d, again bounded and bounded away from zero, symmetric, which is assumed stationary in the sense

$$\forall \mathbf{k} \in \mathbb{Z}^d, \quad A(x + \mathbf{k}, \omega) = A(x, \tau_{\mathbf{k}}\omega) \text{ almost everywhere in } x, \text{ almost surely} \tag{6}$$

(where τ is an ergodic group action). This amounts to assuming that the law of $A(., \omega)$ is $\mathbb{Z}^d$-periodic. Then we consider the boundary value problem

$$\begin{cases} -\operatorname{div}\left(A\left(\frac{x}{\varepsilon}, \omega\right) \nabla u^\varepsilon\right) = f \quad \text{in} \quad \mathcal{D}, \\ u^\varepsilon = 0 \quad \text{on} \quad \partial\mathcal{D}. \end{cases} \tag{7}$$

Standard results of random homogenization [8, 29] apply and allow to find the homogenized problem for Eq. 7. These results generalize the periodic results recalled in Sect. 1.2.1. The solution u^ε to Eq. 7 converges to the solution to Eq. 5 where the homogenized matrix is now defined as:

$$[A^*]_{ij} = \mathbb{E}\left(\int_Q \mathbf{e}_i^T A\,(y, \cdot)\,\left(\mathbf{e}_j + \nabla w_{\mathbf{e}_j}(y, \cdot)\right)\, dy\right), \tag{8}$$

where for any $\mathbf{p} \in \mathbb{R}^d$, $w_\mathbf{p}$ is the solution (unique up to the addition of a random constant) to

$$\begin{cases} -\operatorname{div}\left[A\,(y, \omega)\,\left(\mathbf{p} + \nabla w_\mathbf{p}(y, \omega)\right)\right] = 0, \quad \text{a.s.on} \quad \mathbb{R}^d, \\ \nabla w_\mathbf{p} \quad \text{is stationary in the sense of Eq. 6}, \\ \mathbb{E}\left(\displaystyle\int_Q \nabla w_\mathbf{p}(y, \cdot)\, dy\right) = \mathbf{0}. \end{cases} \tag{9}$$

A striking difference between the random setting and the periodic setting can be observed comparing Eqs. 3 and 9. In the periodic case, the corrector problem is posed on a *bounded* domain, namely the periodic cell Q. In sharp contrast, the corrector problem in Eq. 9 of the random case is posed *on the whole space* $\mathbb{R}^d$, and cannot be reduced, at the theoretical level, to a problem posed on a bounded domain. The fact that the random corrector problem is posed on the entire space has far reaching consequences both for theory and for numerical practice. To some extent, the unboundedness of the domain on which the corrector problem is posed is a common denominator of all the settings that we will address in the present survey. *This unboundedness of the corrector problem is also a fundamental characteristic feature of the practically relevant problems of materials science*. We cannot emphasize enough this fact.

In order to approximate Eq. 9 numerically, truncations of the problem have to be considered, typically on large domains $Q_N = [0, N]^d$ and using periodic boundary conditions. The actual homogenized coefficients are only captured in the asymptotic regime $Q_N \to \mathbb{R}^d$. Overall, it is fair to consider that the approach is very expensive computationally, and often actually prohibitively expensive. Therefore, in many practical situations, the size of the "large" domain Q_N considered is in fact small, and the number of realizations of the random microstructure considered therein to approach the expectation in Eq. 8 is also dramatically limited. Put differently, *there* *is a large gap looming between the actual practice and the regime where the theory provides relevant information.*

Important theoretical questions about the quality and the rate of the convergence in terms of the truncation size arise: see, in particular, the pioneering works by Bourgeat and Piatnitski [17, 18] and, more broadly and recently, a series of works by F. Otto, A. Gloria, S. Armstrong, Ch. Smart, J.-C. Mourrat and their many collaborators, see e.g. [25, 26] for examples of contributions.

2 A Mathematical Toolbox for "Weakly" Random Problems

We begin with this section our study of homogenization of non-periodic problems. We have already mentioned that one possible option is the random setting. And we have mentioned the practical difficulties it raises. In many practical situations, however, the real material under consideration is not far from being a periodic material. At zero-th order of approximation, the material can be considered periodic, and it is only at a higher order that disorder might play a role. We choose, in this section, to encode this disorder using randomness. When the "material" under study is the geological bedrock, there is of course no reason for this assumption to be valid, and the classical random model of Sect. 1.2.2 might be more relevant. In contrast, the assumption makes a lot of sense when considering manufactured materials, where the defect of periodicity typically owes to flaws in the process: the material was *meant* to be periodic, but it is actually not. The practically relevant question is to understand whether or not, despite its smallness, the microscopic amount of randomness might affect the macroscale at order one. Solving this question requires to come up with a modeling strategy for the imperfect material.

Our purpose here is to outline a modeling strategy that accounts for the presence of randomness in a multi-scale computation, but specifically addresses the case when the amount of randomness present in the system is small. In this case, we call the material *weakly random*. The weakly random material is thus considered as a small perturbation of a periodic material. Our purpose is to introduce a toolbox of possible modeling strategies that all keep the computational workload limited (in comparison to a direct attack of the problem as if, like in Sect. 1.2.2, the randomness was *not* small) and that provides an approximation of the response of the material which one may certify by error estimates.

As mentioned above, the simple diffusion equation Eq. 1 is a perfect prototypical testbed for our toolbox. It is ubiquitous in several, if not all engineering sciences and life sciences. Although we have not developed our theory and computations for other, more general equations and settings, we are convinced that the same line of approach (namely small amount of randomness as compared to a reference periodic setting, plus expansion in the randomness amplitude, and simplified computations) can be useful in many contexts.

2.1 Random Deformations of the Periodic Setting

A first random setting, which has been introduced and studied in [11] and is not, mathematically, a particular case of the classical stationary setting recalled in Sect. 1.2.2, consists of *random deformations* of a periodic structure. As said above, it is motivated by the consideration of random geometries that have some specific proximity to the periodic setting. The periodic setting is here taken as a reference configuration, somewhat similarly to the classical mathematical formalization of continuum mechanics where a reference configuration is used to define the state of the material under study. Another related idea, in a completely different context, is the consideration of a reference element for finite element computations. The real situation is then seen via a *mapping* from the reference configuration to the actual configuration. Here, this mapping is a *random* mapping (otherwise, one would know everything on the material up to a change of coordinates and there would be poor practical interest in the approach). Assuming some regularity of this mapping induces constraints on the sets of geometries that the microstructures of the material can take. Put differently, the material structure, even though it is not entirely known, is not arbitrarily disordered.

We fix some $\mathbb{Z}^d$-periodic A_{per}, assumed to satisfy the usual properties of boundedness and coerciveness, and we consider the following specific form of the coefficient A_ε in Eq. 1

$$A_\varepsilon\,(x,\omega) \;=\; A_{per}\left(\Phi^{-1}\left(\frac{x}{\varepsilon},\omega\right)\right), \tag{10}$$

where the function $\Phi(\cdot,\omega)$ is assumed to be, almost surely, a diffeomorphism from $\mathbb{R}^d$ to $\mathbb{R}^d$. The diffeomorphism, called a *random stationary diffeomorphism*, is assumed to additionally satisfy

$$\operatorname{essinf}_{\omega\in\Omega,\,x\in\mathbb{R}^d}\left[\det(\nabla\Phi(x,\omega))\right] = \nu > 0, \tag{11}$$

$$\operatorname{esssup}_{\omega\in\Omega,\,x\in\mathbb{R}^d}\left(|\nabla\Phi(x,\omega)|\right) = M < \infty, \tag{12}$$

$$\nabla\Phi(x,\omega) \quad \text{is stationary in the sense of Eq. 6.} \tag{13}$$

Note that the first two assumptions enforce the "homogeneity" of the diffeomorphism: the deformed periodic structure does not implode nor explode anywhere.

Homogenization holds for the above problem (the details are made precise in [11]). The homogenized problem again reads as in Eq. 5 with the homogenized matrix given by:

$$\begin{aligned}[A^*]_{ij} = \det&\left(\mathbb{E}\left(\int_Q \nabla\Phi(z,\cdot)dz\right)\right)^{-1}\\ &\times\mathbb{E}\left(\int_{\Phi(Q,\cdot)} \mathbf{e}_i^T A_{per}\left(\Phi^{-1}(y,\cdot)\right)\left(\mathbf{e}_j+\nabla w_{\mathbf{e}_j}(y,\cdot)\right)dy\right),\end{aligned} \tag{14}$$

where for any $\mathbf{p} \in \mathbb{R}^d$, $w_{\mathbf{p}}$ is the solution (unique up to the addition of a random constant and belonging to the suitable functional space) to

$$\begin{cases} -\mathrm{div}\left[A_{per}\left(\Phi^{-1}(y,\omega)\right)\left(\mathbf{p}+\nabla w_{\mathbf{p}}\right)\right]=0, \quad \text{a.s.on} \quad \mathbb{R}^d, \\ w_{\mathbf{p}}(y,\omega)=\tilde{w}_{\mathbf{p}}\left(\Phi^{-1}(y,\omega),\omega\right), \quad \nabla\tilde{w}_{\mathbf{p}} \quad \text{is stationary in the sense of Eq. 6}, \\ \mathbb{E}\left(\displaystyle\int_{\Phi(Q,\cdot)} \nabla w_{\mathbf{p}}(y,\cdot)dy\right)=\mathbf{0}. \end{cases} \tag{15}$$

At first sight, there seems to be no simplification whatsoever in considering the above system Eq. 15, which even looks way more complex than the classical random problem Eq. 9. The key point, though, is that the introduction of a new modeling “parameter”, namely the random diffeomorphism Φ, allows to in some sense introduce a distance between the periodic case ($\Phi = Id$) and the random case ($\Phi \neq Id$) considered. Our next step consists in proceeding in this direction.

2.2 Small Random Perturbations of the Periodic Setting

We now superimpose to the setting defined in the previous section the assumption that the material considered is a *small* perturbation of a periodic material. This is formalized upon writing

$$\Phi(x,\omega) = x + \eta\,\Psi(x,\omega) + O(\eta^2), \tag{16}$$

where Ψ is any random field such that Φ is a random stationary diffeomorphism that satisfies Eqs. 11-13 for η sufficiently small.

It has been shown in [11] that, when Φ is such a perturbation of the identity map (see Fig. 1), the solution to the corrector problem of Eq. 15 may be developed in powers of the small parameter η. It reads $\widetilde{w}_{\mathbf{p}}(x,\omega) = w_{per,\mathbf{p}}(x) + \eta w^1_{\mathbf{p}}(x,\omega) + O(\eta^2)$, where $w_{per,\mathbf{p}}$ is the periodic corrector defined in Eq. 3 and where $w^1_{\mathbf{p}}$ solves

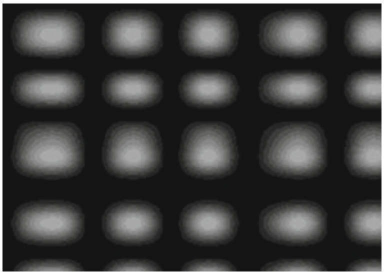


Fig. 1 Small random deformation of a periodic structure. In the unperturbed periodic environment, the inclusions are circular and periodic. The deformation of each inclusion is performed randomly. *Source* [21]

$$\begin{cases} -\mathrm{div}\left[A_{per}\,\nabla w_{\mathbf{p}}^{1}\right] \\ \qquad\qquad = \mathrm{div}\left[-A_{per}\,\nabla\Psi\,\nabla w_{per,\mathbf{p}} - (\nabla\Psi^{T} - (\mathrm{div}\,\Psi)\mathrm{Id})\,A_{per}\,(\mathbf{p}+\nabla w_{per,\mathbf{p}})\right], \\ \nabla w_{\mathbf{p}}^{1} \text{ is stationary and } \mathbb{E}\left(\displaystyle\int_{Q}\nabla w_{\mathbf{p}}^{1}\right) = \mathbf{0}. \end{cases} \tag{17}$$

The problem of Eq. 17 in $w_{\mathbf{p}}^{1}$ is random in nature, but it is in fact easy to see, taking the expectation, that $\overline{w}_{\mathbf{p}}^{1} = \mathbb{E}(w_{\mathbf{p}}^{1})$ is periodic and solves the *deterministic* problem

$$\begin{aligned} &-\mathrm{div}\left[A_{per}\,\nabla\overline{w}_{\mathbf{p}}^{1}\right] \\ &\quad = \mathrm{div}\left[-A_{per}\,\mathbb{E}(\nabla\Psi)\,\nabla w_{per,\mathbf{p}} - (\mathbb{E}(\nabla\Psi^{T}) - \mathbb{E}(\mathrm{div}\,\Psi)\mathrm{Id})\,A_{per}\,(\mathbf{p}+\nabla w_{per,\mathbf{p}})\right]. \end{aligned}$$

This is useful because, on the other hand, the knowledge of $w_{\mathbf{p}}^{0}$ and $\overline{w}_{\mathbf{p}}^{1}$ suffices to obtain a first order expansion (in η) of the homogenized matrix. Indeed, A_{per}^{*} being the periodic homogenized tensor as defined in Eq. 4, and

$$\begin{aligned} A_{ij}^{1} = -\int_{Q}\mathbb{E}(\mathrm{div}\,\Psi)\,[A_{per}^{*}]_{ij} + \int_{Q}(\mathbf{e}_i + \nabla w_{per,\mathbf{e}_i}^{0})^{T} A_{per}\,\mathbf{e}_j\,\mathbb{E}(\mathrm{div}\,\Psi) \\ + \int_{Q}\left(\nabla\overline{w}_{\mathbf{e}_i}^{1} - \mathbb{E}(\nabla\Psi)\nabla w_{per,\mathbf{e}_i}^{0}\right)^{T} A_{per}\,\mathbf{e}_j, \end{aligned}$$

we then have

$$A^{*} = A_{per}^{*} + \eta A^{1} + O(\eta^{2}). \tag{18}$$

For η sufficiently small in function of the accuracy expected, the approach therefore provides a computational strategy to *approximately* compute the homogenized tensor that bypasses the classical random problem and only considers (a sequence of) *deterministic*, periodic problems.

2.3 Rare but Possibly Large Random Perturbations

The previous section has shown that a perturbative approach can be an interesting modeling and computational strategy for cases when the structure of the material is random but "close" to a periodic structure. We now proceed in a similar direction by presenting an alternative perturbative approach, described in full details in [3, 4]. We consider

$$A_{\eta}(x,\omega) = A_{per}(x) + b_{\eta}(x,\omega)\,C_{per}(x), \tag{19}$$

instead of a coefficient $A_{per}\left(\Phi^{-1}(.,\omega)\right)$ with Φ of the form Eq. 16. In Eq. 19, A_{per} is again a periodic matrix modeling the unperturbed material, C_{per} is a periodic matrix

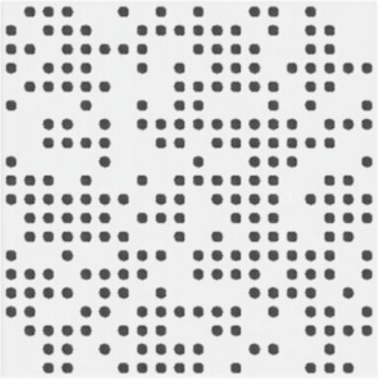


Fig. 2 Defects in a periodic structure. In the unperturbed periodic environment, the inclusions are periodic. The elimination of some of these inclusions are the defects considered. The elimination may be deterministic (as in Sect. 3 below), or random (as in Sect. 1.2.2). One may also consider small probabilities of elimination and construct the corresponding mathematical setting (as in Sect. 2.3). *Source* [3]

modeling the perturbation, and $b_\eta(., \omega)$ is a random field that is, in some sense, small. Consider then the case

$$b_\eta(x, \omega) = \sum_{\mathbf{k}\in\mathbb{Z}^d} \mathbf{1}_{\{Q+\mathbf{k}\}}(x) B_\eta^k(\omega), \tag{20}$$

where the B_η^k are, say, independent identically distributed random variables. One particularly interesting case (see [3, 4] for this case and others) is that when the common law of the B_η^k is a Bernoulli law of parameter η (see Fig. 2).

We now explain *formally* our approach. The mathematical correctness of the approach has been established in the works [23, 40].

To start with, we notice that in the corrector problem

$$-\operatorname{div}\left[A_\eta\,(y, \omega)\left(\mathbf{p} + \nabla w_{\mathbf{p}}(y, \omega)\right)\right] = 0, \tag{21}$$

the only source of randomness comes from the coefficient $A_\eta\,(y, \omega)$. Therefore, in principle, if one knows the law of this coefficient A_η, one knows the law of the corrector function $w_{\mathbf{p}}(y, \omega)$ and therefore may compute the homogenized coefficient A^*, the latter being a function of this law. When the law of A_η is an expansion in terms of a small coefficient, so is the law of $w_{\mathbf{p}}$. Consequently, A_η^* must be attainable using an expansion.

Heuristically, on the cube Q_N and at order 1 in η, the probability to see the perfect periodic material (entirely modeled by the matrix A_{per}) is $(1-\eta)^{N^d} \approx 1 - N^d\eta + O(\eta^2)$, while the probability to see the unperturbed material on all cells except one (where the material has matrix $A_{per} + C_{per}$) is $N^d\,(1-\eta)^{N^d-1}\eta \approx N^d\eta + O(\eta^2)$. All other configurations, with more than two cells perturbed, contribute at orders higher than or equal to η^2. This gives the intuition (indeed confirmed by a mathematical proof) that the first order correction indeed comes from the difference between the material perfectly periodic except on one cell and the perfect material itself: $A_\eta^* = A_{per}^* + \eta A_{1,*} + o(\eta)$ where A_{per}^* is the homogenized matrix for the unperturbed periodic material and

$$A_{1,*}\,\mathbf{e}_i = \lim_{N\to+\infty} \int\limits_{Q_N} \left[(A_{per} + \mathbf{1}_Q C_{per})(\nabla w^N_{\mathbf{e}_i} + \mathbf{e}_i) - A_{per}(\nabla w_{per,\mathbf{e}_i} + \mathbf{e}_i)\right], \tag{22}$$

where $w^N_{\mathbf{e}_i}$ solves

$$-\operatorname{div}\left((A_{per} + \mathbf{1}_Q C_{per})(\mathbf{e}_i + \nabla w^N_{\mathbf{e}_i})\right) = 0 \quad \text{in} \quad Q_N, \quad w^N_{\mathbf{e}_i} \text{ is } Q_N - \text{periodic.} \tag{23}$$

Note that the integral appearing on the right-hand side of Eq. 22 is *not* normalized: it *a priori* scales as the volume N^d of Q_N and has finite limit only because of cancellation effects between the two terms in the integrand.

This perturbative approach has been extensively tested. It has been observed that the large N limit for cubes of size N is already accurately approximated for limited values of N. As in the previous section (Sect. 2.2), the computational efficiency of the approach is clear: solving the two periodic problems with coefficients A_{per} and $A_{per} + \mathbf{1}_Q C_{per}$ for a limited size N is much less expensive than solving the original, random corrector problem for a much larger size N. When the second order term is needed, configurations with two defects have to be computed. They all can be seen as a family of PDEs, parameterized by the geometrical location of the defects (see again Fig. 2). Reduced basis techniques have been shown to allow for a definite speed-up in the computation, *see* [33].

On an abstract level, we note that, in the proposed approach for the "weakly" random regime, the determination of the homogenized tensor for a material containing defects with random locations is reduced to a set of computation of the solutions to correctors problems such as Eq. 23 for materials with defects at *some* particular deterministic locations. This naturally establishes a methodological link with our next section where we indeed consider materials with *deterministic* defects. The link is actually more than methodological: the theoretical results of Sect. 3 establishing that the corrector problems with deterministic defects are uniquely solvable in a suitable class of functions are readily useful in the random setting for the foundation of the approach described here in Sect. 2.

3 Deterministic Defects Within an Otherwise Periodic Structure

We return to the generic multi-scale diffusion equation Eq. 1. Under quite general and mild assumptions on the diffusion (possibly matrix-valued) coefficient A_ε (which needs not be of the form $A_\varepsilon = A_{per}(x/\varepsilon)$ or obey any structural assumption of that type), presumably varying at the tiny scale ε, the equation admits an homogenized limit, which is indeed of the same form as Eq. 1, namely Eq. 5. Celebrated results along these lines are due to S. Spagnolo, E. De Giorgi and L. Tartar and their respec-

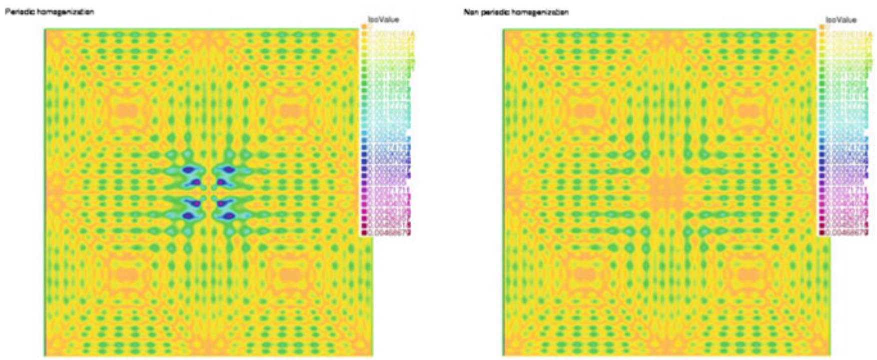


Fig. 3 **Localized defects in a periodic structure**. Some periodic cells in the center of the domain are perturbed. The error $u^{\varepsilon} - u^{\varepsilon,1}$ is displayed when calculating $u^{\varepsilon,1}$ using (left) the periodic corrector $w_{per,\mathbf{p}}$ solution to Eq. 3 and (right) the adjusted corrector $w_{\mathbf{p}}$ solution to Eq. 24. In the former case, the size of the committed error is almost a "defect detector". In the latter case, the error is homogeneous throughout the domain, recovering the quality of the approximation of the unperturbed periodic case. *Source* [12]

tive collaborators, *see* [42]. The strength of such results is their generality. They are obtained by a compactness argument. Schematically the sequence of inverse operators $[-\mathrm{div}(A_{\varepsilon}\nabla.)]^{-1}$ is (weakly) compact in the suitable topology, converges, up to an extraction, and its limit can be proven to be an operator of the same type, namely $[-\mathrm{div}(A^{*}\nabla.)]^{-1}$. On the other hand, and precisely because of the generality, not much is known on the limit A^{*}. This contrasts with periodic homogenization which is both *explicit* (the limit coefficient A^{*} is known by a formula, namely Eq. 4, in function of the, also known, corrector) and *precised* (the rate of convergence of u^{ε} to u^{*} is known for a large variety of norms). Besides their theoretical interest *per se*, the combined two ingredients allow for envisioning, in practice, a numerical approach for the computation of the homogenized limit, certified by a numerical analysis that guarantees a control of the numerical error committed, in function of ε and the discretization parameters.

The question arises to find settings sufficiently general that still allow for the quality of results of the periodic setting. The recent decade has witnessed several mathematical endeavors in this direction. We describe here such an endeavor and give one prototypical example of such a setting, where we illustrate the novelty of the mathematical questions involved (Fig. 3).

Consider Eq. 1 and assume that $A_{\varepsilon} = A(./\varepsilon)$ where the coefficient A models a periodic material perturbed by a localized defect. This setting, mathematically, may be encoded in $A = A_{per} + \tilde{A}$ for $\tilde{A} \in L^{p}(\mathbb{R}^{d})$ for some $p < +\infty$. Clearly, the presence of this defect does not affect the *macroscopic* behavior, that is the homogenized equation for the *same* homogenized coefficient A^{*}, only actually depending on averages of A over large, asymptotically infinite volumes, for which the addition of a function such as $\tilde{A}$ does not matter. On the other hand, when it comes to making this limit more precise, one intuitively realizes, zooming in locally in the material, that the corrector equation that describes the *microscopic* response of the material reads as

$$-\operatorname{div}(A(\mathbf{e}_i + \nabla w_{\mathbf{e}_i})) = 0. \tag{24}$$

This equation is different from Eq. 3, and, in sharp contrast with Eq. 3 (and similarly to what we observed for Eq. 9 in the random setting), *does not reduce* to an equation set on a bounded domain with periodic boundary conditions. Note that, for the particular choice $\tilde{A} = \mathbf{1}_Q C_{per}$, Eq. 23 is a particular instance of Eq. 24 when $N = +\infty$. In essence, Eq. 24 is posed on the entire ambient space $\mathbb{R}^d$, a reflection of the fact that, at the microscopic scale, the defect has broken the periodicity of the environment: the local response is affected by the defect and depends on the state of the *whole* microscopic structure. A considerable mathematical difficulty follows. The classical toolbox for the study of the well-posedness of (here linear) equations on bounded domains: the Lax-Milgram Lemma in the coercive case, the Fredholm Alternative, etc., all techniques that one way or another rely upon the boundedness of the domain or the compactness of the setting, are now ineffective. Should A be random stationary, then Eq. 24 would read as Eq. 9 and admit an equivalent formulation on the abstract probability space. This would make up for compactness, but other significant complications would arise. For Eq. 24, the difficulty must be embraced. A related difficulty is to define the set of admissible functions for solutions, or the variational space in an energetic formulation of the problem. In the specific case $A = A_{per} + \tilde{A}$ with $\tilde{A} \in L^p(\mathbb{R}^d)$, one seeks for the solution to Eq. 24 under the form $w_{\mathbf{e}_i} = w_{per,\mathbf{e}_i} + \tilde{w}_{\mathbf{e}_i}$ that is, *with reference to* the periodic solution $w_{per,\mathbf{e}_i}$, somewhat in echo to what we achieved in Sect. 2.3. Equation 24 then rewrites as

$$-\operatorname{div}\left(A\, \nabla \tilde{w}_{\mathbf{e}_i}\right) = \operatorname{div}\left(\tilde{f}\right),$$

where $\tilde{f} \in L^p(\mathbb{R}^d)$, which, by homogeneity, suggests that the suitable functional space for $\nabla \tilde{w}$ is $L^p(\mathbb{R}^d)$. The question then arises to know whether the operator $[\nabla]\,[\operatorname{div}(A\,\nabla\,.)]^{-1}\,[\operatorname{div}]$ acts continuously in $L^p(\mathbb{R}^d)$. The answer depends on the properties of the coefficient A. In the present setting, it is positive for all $1 < p < +\infty$. The theoretical analysis to reach this conclusion heavily relies upon the celebrated works [5–7] by M. Avellaneda and F. H. Lin for the periodic case (see also [30, 41]).

The consideration of the one-dimensional version of the problem clearly shows (this particular example is worked out in [12]) that when one considers the specific corrector w solution to $-\dfrac{d}{dy}\left((a_{per} + \tilde{a})(y)\left(1 + \dfrac{d}{dy}\,w(y)\right)\right) = 0$, instead of the periodic corrector w_{per} solution to $-\dfrac{d}{dy}\left(a_{per}(y)\left(1 + \dfrac{d}{dy}\,w_{per}(y)\right)\right) = 0$, then the quality of the (two-scale, first order) approximation of the solution u^ε is immediately improved near the defect and at the scale of the defect.

In dimensions higher than or equal to two, the proof is more difficult. Under appropriate conditions, the solution u^ε is well approximated in H^1 norm, both at scale one and at scale ε (thus in particular in L^∞ norm), by the first order expansion

$u^{\varepsilon,1}(x) = u^*(x) + \varepsilon \sum_{i=1}^{d} \partial_{x_i} u^*(x)\, w_{\mathbf{e}_i}(x/\varepsilon)$ constructed using the specific correctors $w_{\mathbf{e}_i}$. The latter approximation property does not in general hold true for the periodic first-order approximation $u^{\varepsilon,1}_{per}(x) = u^*(x) + \varepsilon \sum_{i=1}^{d} \partial_{x_i} u^*(x)\, w_{per,\mathbf{e}_i}(x/\varepsilon)$ constructed using the periodic corrector $w_{per,\mathbf{e}_i}$. One may even make precise the rate of convergence in function of the small parameter ε, and likewise may prove similar convergence for different Sobolev or Hölder norms. The proof of these convergences has first been presented in the case $p = 2$ (and slightly formally) in [12]. All results and extensions are carried out in a series of works [9, 10, 13–15].

The procedure above is not restricted to the linear diffusion problem Eq. 1. One may consider semi-linear equations, quasi-linear equations, systems, etc. And of course it gets all the more delicate as the complexity of the equation increases. One such example, namely an Hamilton-Jacobi equation, is the purpose of the work [19] and also the subject of work in progress by the author and his collaborators, see [16, 20, 28].

Various other cases of defects may be considered for homogenization problems that are otherwise “simple”. They may formally decay at infinity (like the “localized” functions $\tilde{A}$ manipulated above), or not. In the former case, the problem at infinity (that is the problem obtained upon translating the equation far away from the defect) is identical to the underlying periodic problem. In the latter case, the situation may sensitively depend upon what the problem “at infinity” looks like. There may even exist several such problems. Another prototypical example is related to the modeling of *grain boundaries* in materials science: two, different, periodic structures are connected across an interface. The defect is, say, a plane separating the two structures, and at large distances from this interface, different periodic structures are present, depending upon which side of the interface is considered, see [13]. The corresponding mathematical problem is theoretically challenging, and practically relevant. In all cases, the purpose is to identify the homogenized, macroscopic limit, while, in the meantime, retain some of the microscopic features that make the problem relevant.

4 Multi-scale Finite Element Approaches and Nonperiodicity

Multi-scale Finite Element Methods, abbreviated as MsFEM, have proved to be efficient in a number of contexts. In essence, these approaches are based upon choosing, as specific finite dimensional basis to expand the numerical solution upon, a set of functions that themselves are solutions to a highly oscillatory *local* problem, at scale ε, involving the differential operator present in the original equation. This problem-dependent basis set, precomputed (in an offline stage), is likely to better encode the fine-scale oscillations of the solution and therefore allow to capture the solution more accurately. Numerical observation along with mathematical arguments prove that this is indeed generically the case. The versatility of the classical FEM is lost, but with MsFEM, their efficiency is restored for multi-scale problems.

The standard version of the approach has been originally introduced by T. Hou and his collaborators (see the textbook [24] for a general introduction). There exist many variants of such a multi-scale approach, within the formalism of MsFEM or beyond it, and many outstanding numerical analysts and computational analysts have contributed to the field. Classical examples include the Variational multi-scale Method introduced by Hughes et al. the Local Orthogonal Decomposition method by Malqvist and Peterseim, the localization and subspace decomposition method of R. Kornhuber and H. Yserentant, etc. It is not our purpose here to review all these works. We would like to concentrate ourselves here on an issue that is intrinsically related to the context of our discussion, namely breakings of the periodic structure of a material, and its consequence on the accuracy of a dedicated numerical approach.

We recall, on the prototypical multi-scale diffusion problem Eq. 1, that the MsFEM approach, in one of its simplest variant, consists of the following three steps:

1. Introduce a discretization of $\mathcal{D}$ with a coarse mesh; throughout this article, we work with the $\mathbb{P}^1$ Finite Element space

$$V_H = \mathrm{Span}\left\{\phi_i^0,\ 1 \leq i \leq N_{V_H}\right\} \subset H_0^1(\mathcal{D}). \tag{25}$$

2. Solve the local problems (one for each basis function for the coarse mesh)

$$-\operatorname{div}\left(A_\varepsilon \nabla \psi_i^{\varepsilon,\mathbf{K}}\right) = 0 \quad \text{in } \mathbf{K}, \qquad \psi_i^{\varepsilon,\mathbf{K}} = \phi_i^0 \quad \text{on } \partial\mathbf{K}, \tag{26}$$

   on each element $\mathbf{K}$ of the coarse mesh $\mathcal{T}_H$, in order to build the multi-scale basis functions. This is typically performed off-line, using a fine mesh $\mathcal{T}_h$, with $h \ll H$.
3. Apply a standard Galerkin approximation of Eq. 1 on the space

$$\mathrm{Span}\left\{\psi_i^\varepsilon,\ 1 \leq i \leq N_{V_H}\right\} \subset H_0^1(\mathcal{D}), \tag{27}$$

   where ψ_i^ε is such that $\psi_i^\varepsilon\big|_{\mathbf{K}} = \psi_i^{\varepsilon,\mathbf{K}}$ for all $\mathbf{K} \in \mathcal{T}_H$.

The error analysis of this MsFEM method has been performed for $A_\varepsilon = A_{\mathrm{per}}(\cdot/\varepsilon)$ with A_{per} a fixed periodic matrix. Assuming that the basis functions are perfectly determined (that is, $h = 0$), the main error estimate, under the usual assumption of regularity of the data and the mesh, reads as

$$\|u^\varepsilon - u_H^\varepsilon\|_{H^1(\mathcal{D})} \leq C\left(H + \sqrt{\varepsilon} + \sqrt{\frac{\varepsilon}{H}}\right), \tag{28}$$

where C is a constant independent of H and ε.

When the coarse mesh size H is close to the scale ε, a so-called resonance phenomenon, encoded in the term $\sqrt{\varepsilon/H}$ in Eq. 28, occurs and deteriorates the numerical solution. The oversampling method is a popular technique to reduce this effect. In short, the approach, which is non-conforming, consists in setting each local problem on a domain slightly larger than the actual element $\mathbf{K}$ considered, so as to become less sensitive to the arbitrary choice of boundary conditions on that larger domain, and next truncate on the element the functions obtained. That approach allows to significantly improve the results compared to using linear boundary conditions as in Eq. 26. In the periodic case, the following estimate holds

$$\|u^\varepsilon - u_H^\varepsilon\|_{H^1(\mathcal{T}_H)} \leq C\left(H + \sqrt{\varepsilon} + \frac{\varepsilon}{H}\right),$$

where $\|u^\varepsilon - u_H^\varepsilon\|_{H^1(\mathcal{T}_H)} = \sqrt{\sum_{\mathbf{K}\in\mathcal{T}_H} \|u^\varepsilon - u_H^\varepsilon\|^2_{H^1(\mathbf{K})}}$ is the H^1 broken norm of $u^\varepsilon - u_H^\varepsilon$.

The boundary conditions imposed on $\partial\mathbf{K}$ in Eq. 26 are the so-called linear boundary conditions. Besides the linear boundary conditions, and the oversampling technique we have just mentioned, there are many other possible boundary conditions for the local problems. They may give rise to conforming, or non-conforming approximations. The choice sensitively affects the overall accuracy. In an informal way, the whole history of improvements of the original version of MsFEM can be revisited as the history of improvements of the choice of suitable "boundary conditions" for Eq. 26.

The question of how much the choice of boundary conditions for the local problems Eq. 26 alters the overall accuracy is all the more crucial in the context of non-periodic structures. A prototypical case of the difficulty is that of perforated materials. Consider the Poisson problem set on a domain with perforations of size ε. For a generic mesh, the edges (or, alternately, the facets in a three-dimensional setting) of the mesh may intersect the perforations. It is intuitive that difficulties then arise since the (linear or else than linear) *postulated* behavior of the basis functions along the edges has little chance to accurately capture the actual behavior of the exact solution, given the perforations. Of course, one may use oversampling in order to circumvent this difficulty, but then the approach is non conformal and other difficulties arise, besides the increased computational cost. Also, one may consider meshing the domain in such a way that the edges intersect as few perforations as possible. For a periodic array of perforations, this is a decent solution. But in a non-periodic setting, and this is all the more true in a fully disordered array of perforations, this is impractical. A possible option introduced in [34], and extended in [35, 38, 39] and other subsequent works by different authors, is to resort to "*weak*" boundary conditions, in the form of Crouzeix-Raviart boundary conditions. The Dirichlet boundary conditions on $\partial\mathbf{K}$ in Eq. 26 are then replaced by conditions of the type

$$\int_{edge} \psi_i^{\varepsilon,\mathbf{K}} = 0 \quad \text{or} \quad 1,$$

$$n_{edge} \cdot A_\varepsilon \nabla \psi_i^{\varepsilon,\mathbf{K}} = \text{Constant}\ ,$$

on all edges, where the local function $\psi_i^{\varepsilon,\mathbf{K}}$ is now associated to an edge i. For this approach, under technical assumptions, the error estimate is identical to that for linear boundary conditions, namely Eq. 28.

More importantly, upon using such "weak" boundary conditions in the context of a perforated computational domain (and adding other, generic ingredients, such as bubble functions), the accuracy, if not improved, is now significantly more robust with respect to the existence of intersection between edges and perforations. A "stress-test" considering two extreme scenarios illustrates this property: see in [35] the detailed comparison of the results obtained with the MsFEM method and different boundary conditions for the local problems for the shifted meshes in Fig. 4.

Let us conclude this section by emphasizing the formal link between the existence results for the non-periodic corrector $w_{\mathbf{p}}$ that have been examined in the previous section and the actual local basis functions $\psi_i^{\varepsilon,\mathbf{K}}$ of the MsFEM approaches discussed here. Up to irrelevant technicalities and details, the corrector and the local functions are, intrinsically, the same mathematical object: they are obtained by zooming in locally and solving the problem at the scale of its heterogeneities.

5 Homogenization Under Partial Information

One way or another, all the approaches described so far, both at the theoretical level and the numerical level, rely on the full knowledge of the coefficient A_ε. It turns out that there are several *practical* contexts where such a knowledge is incomplete, or sometimes merely unavailable. From an engineering perspective (think e.g. of experiments in Mechanics), there are indeed numerous prototypical situations for Eq. 1 where the response u^ε can be measured for some loadings f, but where A_ε is not completely known, let alone the fact that it is periodic or not. In these situations, it is thus not possible to use homogenization theory, nor to proceed with any MsFEM-type approach or with the similar approaches mentioned above. Finding a pathway alternate to standard approaches is thus a practically relevant question. We are interested in approaches valid for the different regimes of ε, which make no use of the knowledge on the coefficient A_ε, but only use some responses of the medium obtained for certain given solicitations. Questions similar in spirit have been addressed two decades ago by Durlofsky. The point is also to define an effective coefficient only using outputs of the system. They are however different in practice (see [36] for a detailed discussion).

For simplicity, we restrict ourselves to cases when Eq. 1 admits (possibly up to some extraction) a homogenized limit Eq. 5 where the homogenized matrix coefficient A^* is deterministic and constant. This restrictive assumption on the class of A^* (and thus on the structure of the coefficient A_ε in Eq. 1) is useful for our theoretical justifications, but not mandatory for the approach to be applicable.

For any constant matrix $\overline{A}$, we consider generically the problem with constant coefficients

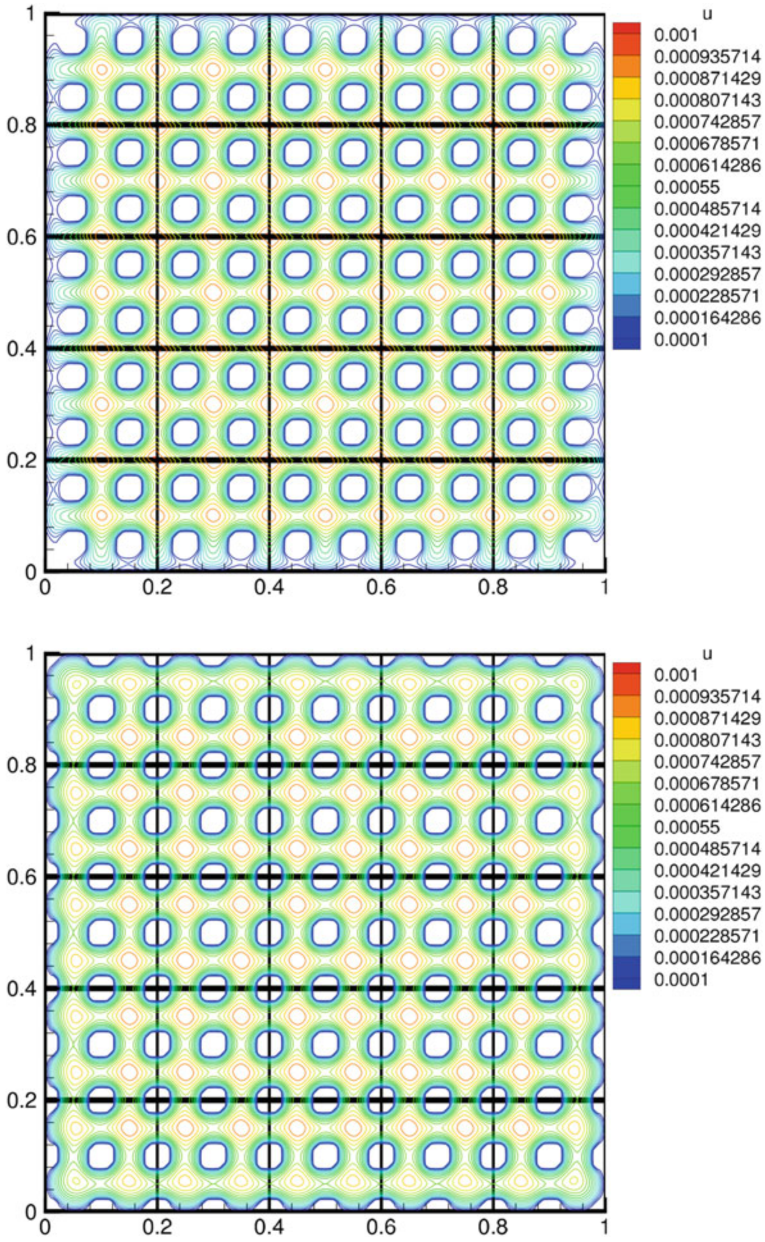


Fig. 4 Two extreme cases of meshes regarding intersections with the perforations: no intersection at all (top), or as many intersections as possible (bottom). The Crouzeix-Raviart version of MsFEM is, roughly, equally accurate in both situations. *Source* [35]

$$-\operatorname{div}\left(\overline{A}\,\nabla\overline{u}\right)=f. \tag{29}$$

We investigate, for any value of the parameter ε, how we may define a constant symmetric matrix such that the solution $u(\overline{A},f)=\overline{u}$ to Eq. 29 with matrix $\overline{A}$ best approximates the solution to Eq. 1. The best constant matrix $\overline{A}$ is (temporarily) defined as a minimizer of

$$I_\varepsilon=\inf_{\text{constant matrix}\overline{A}>0}\quad\sup_{\substack{f\,\in\,L^2(\mathcal{D}),\\ \|f\|_{L^2(\mathcal{D})}=1}}\quad\left\|u^\varepsilon(f)-u(\overline{A},f)\right\|^2_{L^2(\mathcal{D})}, \tag{30}$$

where we have explicitly emphasized the dependency upon the right-hand side f of the solutions to Eq. 1 and Eq. 29. The norm in Eq. 30 is an L^2 norm (and not e.g. an H^1 norm) because, for sufficiently small ε, we wish the best constant matrix $\overline{A}$ to be close to A^*, while u^ε strongly converges to u^* only in the L^2 norm but not in the H^1 norm. The key point is that Eq. 30 is only based on the knowledge of the outputs u^ε (that could be e.g. experimentally measured), and not on that of A_ε itself. The theoretical study of the minimization problem Eq. 30 has been carried out in [36]. In particular it has been proven that, under classical assumptions, the matrices $\overline{A}$ with energy asymptotically close to the infimum I_ε all converge to A^* as ε vanishes. In passing, we note that the approach provides, at least in some settings, a characterization of the homogenized matrix which is an alternative to the standard characterization of homogenization theory. To the best of our knowledge, this characterization, although probably known, has never been made explicit in the literature.

In fact (and this does not alter the above theoretical results), the actual minimization problem we use for the practice reads as

$$I^{\text{pract}}_\varepsilon=\inf_{\text{constant matrix}\overline{A}>0}\quad\sup_{\substack{f\,\in\,L^2(\mathcal{D}),\\ \|f\|_{L^2(\mathcal{D})}=1}}\quad\left\|-\Delta^{-1}\,\left(-\operatorname{div}\overline{A}\,\nabla\,u^\varepsilon(f)-f\right)\right\|^2_{L^2(\mathcal{D})}, \tag{31}$$

where $-\Delta^{-1}$ is the inverse laplacian operator supplied with homogeneous Dirichlet boundary conditions. The function minimized in Eq. 31 is related to the one of Eq. 30 through the application, inside the L^2 norm of the latter, of the zero-order differential operator $\Delta^{-1}\operatorname{div}(\overline{A}\,\nabla\,.\,)$. Note that, in sharp contrast with Eq. 30, the function to minimize in Eq. 31 is now, formally, a second-order polynomial in function of $\overline{A}$. This property significantly speeds up the computations of the infimum. The specific choice Eq. 31 has been suggested to us by Albert Cohen.

Note also that, in practice, we cannot maximize upon all right-hand sides f in $L^2(\mathcal{D})$ (with unit norm) and that we therefore replace the supremum by a maximization upon a finite-dimensional set of thoughtfully selected right-hand sides.

In [36, 37], we have presented a series of numerical experiments using the above approach. Our tests have established that the approach is in particular able to accurately identify the homogenized matrix A^* in the periodic case (with a computational

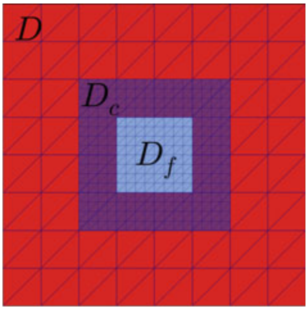


Fig. 5 Homogenization approach within an Arlequin-type coupling: The fine-scale highly oscillatory model and the coarse-grained model (tentatively identical to the homogenized model) co-exist in an overlap region. The three regions described in the body of our text are displayed, along with the fine and coarse meshes. *Source* [27]

time that is much larger than the classical approach, but this is not the point). More importantly, it is also able to complete this task in the random case (where the classical approach can be prohibitively expensive). Finally, and since no particular structure of the coefficient A_ε is used, it may be applied to a large variety of non-periodic structures.

A remark is in order: in both cases of periodic and random homogenization, the classical approach computes the homogenized coefficients by first approximating the corrector function. A fair comparison between the approaches can therefore only be achieved if the above approach also provides some approximation of the corrector function. It is indeed the case: the latter function can also be obtained in our approach, at a reduced additional computational cost, as demonstrated in [36].

A variant of the above approach, originally introduced in [22], is currently under investigation in [27]. The purpose of this variant is also to approximate A^* without explicitly using A_ε, and to achieve this in a robust, engineering-type manner. In a nutshell, the approach consists in considering a domain divided in three regions, see Fig. 5. The inner region and the outer region respectively contain only the oscillatory model of Eq. 1 and the tentative homogenized model of Eq. 29. In between these two regions, an overlap region where both models exist is used for a smooth coupling. Specifically, the coupling is performed using an Arlequin-type approach (see again [22]) but this is not mandatory for the approach to perform. A linear Dirichlet boundary condition, say $u = x_1$, is imposed on the external surface of the domain. It intuitively plays the role of the right-hand side function f in Eq. 31. At ε fixed presumably small, one then solves the minimization problem

$$J_\varepsilon = \inf_{\text{constant matrix}\overline{A}>0} \left\| \nabla(u(\overline{A}) - x_1) \right\|^2_{L^2(\mathcal{D})} . \tag{32}$$

In the limit of ε vanishing, it is established that J_ε also vanishes and the only minimizer is obtained for $\overline{A}\,\mathbf{e}_1 = A^*\,\mathbf{e}_1$, where $\mathbf{e}_1 = \nabla(x_1)$ is the first canonical vector of the ambient space $\mathbb{R}^d$. Repeating this procedure along each dimension of $\mathbb{R}^d$ allows to eventually identify the matrix A^*. Several computational improvements of the original approach are introduced in [27]. A numerical analysis is also presented.

Acknowledgements The author wishes to thank his many collaborators on the topics overviewed in the present article, and in particular: Xavier Blanc, Pierre Cardaliaguet, Olga Gorynina, Rémi Goudey, Ulrich Hetmaniuk, Frédéric Legoll, Pierre-Louis Lions, Alexei Lozinski, Panagiotis Souganidis, Sylvain Wolf. The research of the author is partially supported by ONR under grant N00014-20-1-2691 and by EOARD under grant FA8655-20-1-7043.

References

1. Allaire, G.: Shape optimization by the homogenization method. In: Applied Mathematical Sciences, vol. 146. Springer, New York (2002)
2. Anantharaman, A., Costaouec, R., Le Bris, C. , Legoll, F., Thomines, F.: Introduction to numerical stochastic homogenization and related computational challenges. In: Multi-scale Modeling and Analysis for Materials Simulation. Lecture Notes Series, Institute of Mathematical Science, National University of Singapore, vol. 22, pp. 197–272 (2012)
3. Anantharaman, A., Le Bris, C.: A numerical approach related to defect-type theories for some weakly random problems in homogenization. Multiscale Model. Simul. **9**(2), 513–544 (2011)
4. Anantharaman, A., Le Bris, C.: Elements of mathematical foundations for a numerical approach for weakly random homogenization problems. Commun. Comput. Phys. **11**, 1103–1143 (2011)
5. Avellaneda, M., Lin, F.-H.: Compactness methods in the theory of homogenization. Comm. Pure Appl. Math. **40**(6), 803–847 (1987)
6. Avellaneda, M., Lin, F.-H.: Compactness methods in the theory of homogenization. II: equations in non-divergence form, Commun. Pure Appl. Math. **42**(2), 139–172 (1989)
7. Avellaneda, M., Lin, F.-H.: L^p bounds on singular integrals in homogenization. Commun. Pure Appl. Math. **44**(8–9), 897–910 (1991)
8. Bensoussan A., Lions J.-L., Papanicolaou G.: Asymptotic Analysis for Periodic Structures, 374. American Mathematical Society (2011)
9. Blanc, X., Josien, M., Le Bris, C.: Precised approximations in elliptic homogenization beyond the periodic setting. Asymptot. Anal. **116**(2), 93–137 (2020)
10. Blanc, X., Josien, M, Le Bris, C.: Local precised approximation for multi-scale problems with local defects, C. R. Acad. Sci. Paris Sér. I Math. **357**(2), 167–174 (2019)
11. Blanc, X., Le Bris, C., Lions, P.-L.: Stochastic homogenization and random lattices. J. Math. Pures Appl. **88**, 34–63 (2007)
12. Blanc, X., Le Bris, C., Lions, P-L.: A possible homogenization approach for the numerical simulation of periodic microstructures with defects. Milan J. Math. **80**, 351–367 (2012)
13. Blanc, X., Le Bris, C., Lions, P-L.: Local profiles for eliptic problems at different Scales: defects in, and Interfaces between periodic structures. Comm. Partial Diff. Eq. **40**, 2173–2236 (2015)
14. Blanc, X., Le Bris, C., Lions, P-L.: On correctors for linear elliptic homogenization in the presence of local defects, Comm. Partial Different. Eq. **43**(6), 965–997 (2018)
15. Blanc, X., Le Bris, C., Lions, P-L.: On correctors for linear elliptic homogenization in the presence of local defects: the case of advection-diffusion. J. Maths Pures Appl. **124**, 106–112 (2019)
16. Blanc, X., Wolf, S.: Homogenization of Poisson equation in a non-periodically perforated domain. Asymptot. Anal. **126**(1–2), 129–155 (2021)
17. Bourgeat, A., Piatnitski, A.: Estimates in probability of the residual between the random and the homogenized solutions of one-dimensional second-order operator. Asymptot. Anal. **21**, 303–315 (1999)

18. Bourgeat, A., Piatnitski, A.: Approximation of effective coefficients in stochastic homogenization. Ann I. H. Poincaré - PR **40**(2), 153–165 (2004)
19. Cardaliaguet, P., Le Bris, C., Souganidis, P.: Perturbation problems in homogenization of Hamilton-Jacobi equations. J. Math. Pures Appl. **117**, 221–262 (2018)
20. Cardaliaguet P., Le Bris C., Souganidis P.: work in progress
21. Costaouec, R., Le Bris, C., Legoll, F.: Numerical approximation of a class of problems in stochastic homogenization. C. R. Acad. Sci. Paris Sér. I Math. **348**, 99–103 (2010)
22. Cottereau, R.: Numerical strategy for unbiased homogenization of random materials. Internat. J. Numer. Methods Engrg. **95**(1), 71–90 (2013)
23. Duerinckx, M.: Analyticity of homogenized coefficients under Bernoulli perturbations and the Clausius-Mossotti formulas. Arch. Ration. Mech. Anal. **220**(1), 297–361 (2016)
24. Efendiev, Y., Hou, T.Y.: Multi-scale finite element methods. In: Surveys and Tutorials in the Applied Mathematical Sciences, vol. 4. Springer, New York (2009)
25. Gloria, A., Neukamm, S., Otto, F.: A regularity theory for random elliptic operators. Milan J. Math. **88**(1), 99–170 (2020)
26. Gloria, A., Otto, F.: Quantitative results on the corrector equation in stochastic homogenization. J. Eur. Math. Soc. **19**(11), 3489–3548 (2017)
27. Gorynina, O., Le Bris, C., Legoll, F.: Some remarks on a coupling method for the practical computation of homogenized coefficients. SIAM J. Sci. Comput. **43**(2), A1273–A1304 (2021). See also https://arxiv.org/abs/2106.05202
28. Goudey, R., Le Bris, C.: work in progress
29. Jikov, V., Kozlov, S., Oleinik, O.: Homogenization of Differential Operators and Integral Functionals. Springer Science & Business Media, Berlin (2012)
30. Kenig, C.E., Lin, F.-H., Shen, Z.: Periodic homogenization of Green and Neumann functions. Comm. Pure Appl. Math. **67**(8), 1219–1262 (2014)
31. Le Bris, C., Legoll, F., Thomines, F.: Multi-scale finite element approach for weakly random problems and related issues. ESAIM Math. Model. Numer. Anal. **48**, 815–858 (2014)
32. Le Bris, C., Legoll, F., Thomines, F.: Rate of convergence of a two-scale expansion for some weakly stochastic homogenization problems. Asymptot. Anal. **80**(3–4), 237–267 (2012)
33. Le Bris, C., Thomines, F.: A reduced Basis approach for some weakly stochastic multi-scale problems. Chin. Ann. Math. Ser. B **33**(5), 657–672 (2012)
34. Le Bris, C., Legoll, F., Lozinski, A.: MsFEM à la Crouzeix-Raviart for highly oscillatory elliptic problems. Chin. Ann. Math. Ser. B **34**(1), 113–138 (2013)
35. Le Bris, C., Legoll, F., Lozinski, A.: An MsFEM type approach for perforated domains. Multiscale Model. Simul. **12**(3), 1046–1077 (2014)
36. Le Bris, C., Legoll, F., Lemaire, S.: On the best constant matrix approximating an oscillatory matrix-valued coefficient in divergence-form operators. ESAIM Control Optim. Calc. Var. **24**(4), 1345–1380 (2018)
37. Le Bris, C., Legoll, F., L, K.: Coarse approximation of an elliptic problem with highly oscillatory coefficients. C. R. Acad. Sci. Paris Sér. I Math. **351**, 265–270 (2013)
38. Le Bris, C., Legoll, F., Madiot, F.: A numerical comparison of some MsFEM-type approaches for advection dominated problems in heterogeneous media. ESAIM Math. Model. Numer. Anal. **51**(3), 851–888 (2017)
39. Le Bris, C., Legoll, F., Madiot, F.: Multi-scale finite element methods à la Crouzeix-Raviart for advection-dominated problems in perforated domains. Multiscale Model. Simul. **17**(2), 773–825 (2019)
40. Mourrat, J.C.: First-order expansion of homogenized coefficients under Bernoulli perturbations. J. Math. Pures Appl. **103**(1), 68–101 (2015)
41. Shen, Z.: Periodic homogenization of elliptic systems. In: Operator Theory: Advances and Applications, 269, Advances in Partial Differential Equations. Basel), Birkhäuser. Springer, Cham (2018)
42. Tartar, L.: The general theory of homogenization. A personalized introduction. In: Lecture Notes of the Unione Matematica Italiana, vol. 7. Springer, Berlin (UMI, Bologna) (2009)

Hyperbolic Model Reduction for Kinetic Equations

Zhenning Cai, Yuwei Fan, and Ruo Li

Abstract We make a brief historical review of the moment model reduction for the kinetic equations, particularly Grad's moment method for Boltzmann equation. We focus on the hyperbolicity of the reduced model, which is essential for the existence of its classical solution as a Cauchy problem. The theory of the framework we developed in the past years is then introduced, which preserves the hyperbolic nature of the kinetic equations with high universality. Some lastest progress on the comparison between models with/without hyperbolicity is presented to validate the hyperbolic moment models for rarefied gases.

1 Historical Overview

The moment methods are a general class of modeling methodologies for kinetic equations. We would like to start this paper with a historical review of this topic. However, due to the huge amount of references, a thorough overview would be lengthy and tedious. Therefore, in this section, we only restrict ourselves to the methods related to the hyperbolicity of moment models. Even so, our review in the following paragraphs does not exhaust the contributions in the history.

According to Sir J. H. Jeans [29], the kinetic picture of a gas is "a crowd of molecules, each moving on its own independent path, entirely uncontrolled by forces from the other molecules, although its path may be abruptly altered as regards both

Z. Cai
Department of Mathematics, National University of Singapore, 10 Lower Kent Ridge Road, Singapore 119076, Singapore
e-mail: matcz@nus.edu.sg

Y. Fan
Department of Mathematics, Stanford University, Stanford, CA 94305, USA
e-mail: ywfan@stanford.edu

R. Li (✉)
CAPT, LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China
e-mail: rli@math.pku.edu.cn

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_8

speed and direction, whenever it collides with another molecule or strikes the boundary of the containing vessel." In order to describe the evolution of non-equilibrium gases using the phase-space distribution function, the Boltzmann equation was proposed [1] as a non-linear seven-dimensional partial differential equation. The independent variables of the distribution function include the time, the spatial coordinates, and the velocity.

In most cases, the full Boltzmann equation cannot be solved even numerically. One has to characterize the motion of the gas by resorting to various approximation methods to describe the evolution of macroscopic quantities. One successful way to find approximate solutions is the Chapman-Enskog method [15, 18], which uses a power series expansion around the Maxwellian to describe slightly non-equilibrium gases. The method assumes that the distribution function can be approximated up to any precision only using equilibrium variables and their derivatives. Alternatively, Grad's moment method [24] was developed in the late 1940s. In this method, by taking velocity moments of the Boltzmann equation, transport equations for macroscopic averages are obtained. The difficulty of this method is that the governing equations for the components of the nth velocity moment also depend on components of the $(n + 1)$th moment. Therefore, one has to use a certain closing relation to get a closed system after the truncation.

Among the models given by Grad's method [24], Grad's 13-moment system is the most basic one beyond the Navier-Stokes equations, as any Grad's models with fewer moments do not include either stress tensor or heat transfer. In [23], it was commented that Grad's moment method could be regarded as mathematically equivalent to the Chapman-Enskog method in certain cases. Thus the deduction of Grad's 13-moment system can be regarded as an application of perturbation theory to the Boltzmann equation around the equilibrium. Therefore, it is natural to hope that the 13-moment system will be valid in the vicinity of equilibrium, although it was not expected to be valid far away from the equilibrium distribution [25]. However, due to its complex mathematical expression, it is even not easy to check if the system is hyperbolic, as pointed out in [2]. As late as in 1993, it was eventually verified in [35, 36] that the 1D reduction of Grad's 13-moment equations is hyperbolic around the equilibrium.

In 1958, Grad wrote an article "Principles of the kinetic theory of gases" in Encyclopedia of Physics [26], where he collected his own method in the class of "more practical expansion techniques". However, successful applications of the 13-moment system had been hardly seen within two decades after Grad's classical paper in 1949, as mentioned in the comments by Cercignani [14]. One possible reason was found by Grad himself in [25], where it was pointed out that there may be unphysical sub-shocks in a shock profile for Mach number greater than a critical value. However, the appearance of sub-shocks cannot give any hints on the underlying reason why Grad's moment method does not work for slow flows. Nevertheless, Grad's moment method was still pronounced to "open a new era in gas kinetic theory" [27].

In our paper [5], it was found astonishingly that in the 3D case, the equilibrium is NOT an interior point of the hyperbolicity region of Grad's 13-moment model. Consequently, even if the distribution function is arbitrarily close to the local equilibrium, the local existence of the solution of the 13-moment system cannot be guaranteed as a Cauchy problem of a first-order quasi-linear partial differential system without analytical data. The defects of the 13-moment model due to the lack of hyperbolicity had never been recognized as so severe a problem. The absence of hyperbolicity around local equilibrium is a candidate reason to explain the overall failure of Grad's moment method.

After being discovered, the lack of hyperbolicity is well accepted as a deficiency of Grad's moment method, which makes the application of the moment method severely restricted. "There has been persistent efforts to impose hyperbolicity on Grad's moment closure by various regularizations" [39], and lots of progress has been made in the past decades. For example, Levermore investigated the maximum entropy method and showed in [33] that the moment system obtained with such a method possesses global hyperbolicity. Unfortunately, it is difficult to put it into practice due to the lack of a finite analytical expression, and the equilibrium lies on the boundary of the realizability domain for any moment system containing heat flux [30]. Based on Levermore's 14-moment closure, an affordable 14-moment closure is proposed in [34] as an approximation, which extends the hyperbolicity region to a great extent. Let us mention that actually in [5], we also derived a 13-moment system with hyperbolicity around the equilibrium.

It looks highly non-trivial to gain hyperbolicity even around the equilibrium, while things changed not long ago. Besides the achievement of local hyperbolicity around the equilibrium, the study on the globally hyperbolic moment systems with large numbers of moments was also very successful in the past years. In the 1D case with both spatial and velocity variables being scalar, a globally hyperbolic moment system was derived in [3] by regularization. Motivated by this work, another type of globally hyperbolic moment systems was then derived in [31] using a different strategy. The model in [3] is obtained by modifying only the last equation and the model in [31] revises only the last two equations in Grad's original system. The characteristic fields of these models (genuine nonlinearity, linear degeneracy, and some properties of shocks, contact discontinuities, and rarefaction waves) can be fully clarified, as shows that the wave structures are formally a natural extension of Euler equations.

In [4], the regularization method in [3] is extended to multi-dimensional cases. Here the word "multi-dimension" means that the dimensions of spatial coordinates and velocity are any positive integers and can be different. The complicated multi-dimensional models with global hyperbolicity based on a Hermite expansion of the distribution function up to any degree were systematically proposed in [4]. The wave speeds and the characteristic fields can be clarified, too. Later on, the multi-dimensional model for an anisotropic weight function with global hyperbolicity was derived in [20].

Achieving global hyperbolicity was definitely encouraging, while it sounded like a huge mystery for us how the regularization worked in the aforementioned cases. Particularly, the method cannot be applied to moment systems based on a spherical harmonic expansion of distribution function such as Grad's 13-moment system. As we pointed out, the hyperbolicity is essential for a moment model, while it is hard to obtain by a direct moment expansion of kinetic equations. To overcome such a problem, we in [6] fortunately developed a systematic framework to perform moment model reduction that preserves global hyperbolicity. The framework works not only for the models based on Hermite expansions of the distribution function in the Boltzmann equation, but also works for any ansatz of the distribution function in the Boltzmann equation. Actually, the framework even works for kinetic equations in a fairly general form.

The framework developed in [6] was further presented in the language of projection operators in [19], where the underlying mechanism of how the hyperbolicity is preserved during the model reduction procedure was further clarified. This is the basic idea of our discussion in the next section.

2 Theoretical Framework

In this section, we briefly review the framework in [19] to construct globally hyperbolic moment system from kinetic equations, as well as its variants and some further development. To clarify the statement, we first present the definition of the hyperbolicity as follows:

Definition 1 The first-order system of equations

$$\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{d=1}^{D} \mathbf{A}_d(\boldsymbol{w}) \frac{\partial \boldsymbol{w}}{\partial x_d} = 0, \quad \boldsymbol{w} \in \mathbb{G}$$

is hyperbolic at $\boldsymbol{w}_0$, if for any unit vector $\boldsymbol{n} \in \mathbb{R}^D$, the matrix $\sum_{d=1}^{D} n_d \mathbf{A}_d(\boldsymbol{w}_0)$ is real diagonalizable; the system is called globally hyperbolic if it is hyperbolic for any $\boldsymbol{w} \in \mathbb{G}$.

Based on this definition, the analysis of the hyperbolicity of moment systems reduces to a problem of linear algebra: the analysis of the real diagonalizablity of the coefficient matrices. Without knowing the exact values of the matrix entries, the real diagonalizability of a matrix has to be studied by some sufficient conditions. Some of them are

Condition 1 *All its eigenvalues are real and it has n linearly independent eigenvectors.*

Condition 2 *All the eigenvalues of the matrix are real and distinct.*

Condition 3 *The matrix is symmetric or similar to a symmetric matrix.*

Grad [24] investigated the characteristic structure of the 1D reduction of Grad's 13-moment system, whose hyperbolicity was further studied in [36] based on the Condition 2. Afterwards, this condition is adopted in the proof of the hyperbolicity of the regularized moment system for the 1D case in [3]. It is worth noting that using Condition 2 usually requires us to compute the characteristic polynomial of the coefficient matrix of the moment system, and for large moment systems, this may be complicated or even impractical. Even if the characteristic polynomial is computed, showing that the eigenvalues are real and distinct is still highly nontrivial. This severely restricts the use of this condition in kinetic model reduction.

To study the hyperbolicity in multi-dimensional cases, we have applied Condition 1 in [5] to show that Grad's 13-moment system loses hyperbolicity even in an arbitrarily small neighborhood of the equilibrium, and in [4] to prove the global hyperbolicity of the regularized moment system for the multi-dimensional case. Due to the requirement on the eigenvectors, both proofs based on Condition 1 are complicated and tedious. By contrast, it is much easier to check Condition 3, based on which Levermore provided a concise and clear proof of the hyperbolicity of the maximum entropy moment system in [33]. In [19], we re-studied the hyperbolicity of the regularized moment system in [3, 4] based on the Condition 3 and then generalized it to a framework. Below we will start our discussion from a review of these hyperbolic moment systems.

2.1 Review of Globally Hyperbolic Moment System

Let us consider the Boltzmann equation:

$$\frac{\partial f}{\partial t} + \sum_{d=1}^{D} v_d \frac{\partial f}{\partial x_d} = Q(f), \tag{1}$$

and denote the *local equilibrium* by f_{eq}, which satisfies $Q(f_{eq}) = 0$ and $f_{eq} > 0$. The key idea of Grad's moment method is to expand the distribution as

$$f(t, \boldsymbol{x}, \boldsymbol{v}) = \sum_{|\alpha| \le M} f_{eq}(t, \boldsymbol{x}, \boldsymbol{v}) f_\alpha(t, \boldsymbol{x}) He_\alpha(t, \boldsymbol{x}, \boldsymbol{v}) = \sum_{|\alpha| \le M} f_\alpha(t, \boldsymbol{x}) \mathcal{H}_\alpha(t, \boldsymbol{x}, \boldsymbol{v}) \tag{2}$$

for a given integer $M \ge 2$, where for the multi-dimensional index $\alpha \in \mathbb{N}^D$, $|\alpha| = \sum_{d=1}^{D} \alpha_d$, and the basis function $\mathcal{H}_\alpha$ is defined by $\mathcal{H}_\alpha = f_{eq} He_\alpha$, with He_α being the orthonormal polynomials of $\boldsymbol{v}$ with weight function f_{eq}. When f_{eq} is the local Maxwellian, He_α can be obtained by translation and scaling of Hermite polynomials. Grad's moment system can then be obtained by substituting the expansion into the Boltzmann equation and matching the coefficients of $\mathcal{H}_\alpha$ with $|\alpha| \le M$. To clearly describe this procedure, we assume that the distribution function f is defined on a space $\mathbb{H}$ spanned by the basis functions $\mathcal{H}_\alpha$ for all $\alpha \in \mathbb{N}^D$, and we let $\mathbb{H}_M := \operatorname{span}\{\mathcal{H}_\alpha : |\alpha| \le M\}$ be the subspace for our model reduction. Then one can introduce the projection from $\mathbb{H}$ to $\mathbb{H}_M$ as

$$\mathcal{P}f = \sum_{|\alpha| \leq M} f_\alpha \mathcal{H}_\alpha \text{ with } f_\alpha = \langle f, \mathcal{H}_\alpha \rangle, \tag{3}$$

where the inner product is defined as $\langle f, g \rangle = \int_{\mathbb{R}^D} fg/f_{eq} \, \mathrm{d}\boldsymbol{v}$. The projection accurately describes Grad's expansion (2) and provides a tool to study the operators in the space $\mathbb{H}_M$. For example, matching the coefficients of the basis $\mathcal{H}_\alpha$ with $|\alpha| \leq M$ can be understood as projecting the system into the space $\mathbb{H}_M$. Hence, Grad's moment system is written as

$$\mathcal{P}\frac{\partial \mathcal{P}f}{\partial t} + \sum_{d=1}^{D} \mathcal{P}v_d \frac{\partial \mathcal{P}f}{\partial x_d} = \mathcal{P}Q(\mathcal{P}f). \tag{4}$$

Let $\boldsymbol{\mathcal{H}}$ be the vector whose components are all the basis functions $\mathcal{H}_\alpha$ with $|\alpha| \leq M$ listed in a given order. Since $\mathcal{P}f$ is a function in $\mathbb{H}_M$, one can collect all the independent variables in $\mathcal{P}f$ and denote it by $\boldsymbol{w}$ with its length equal to the dimension of $\mathbb{H}_M$. Thanks to the definition of the projection operator $\mathcal{P}$, there exist the square matrices $\mathbf{D}$ and $\mathbf{B}_d$, $d = 1, \dots, D$ such that

$$\mathcal{P}\frac{\partial \mathcal{P}f}{\partial t} = \boldsymbol{\mathcal{H}}^T \mathbf{D} \frac{\partial \boldsymbol{w}}{\partial t}, \quad \mathcal{P}v_d \frac{\partial \mathcal{P}f}{\partial x_d} = \boldsymbol{\mathcal{H}}^T \mathbf{B}_d \frac{\partial \boldsymbol{w}}{\partial x_d}. \tag{5}$$

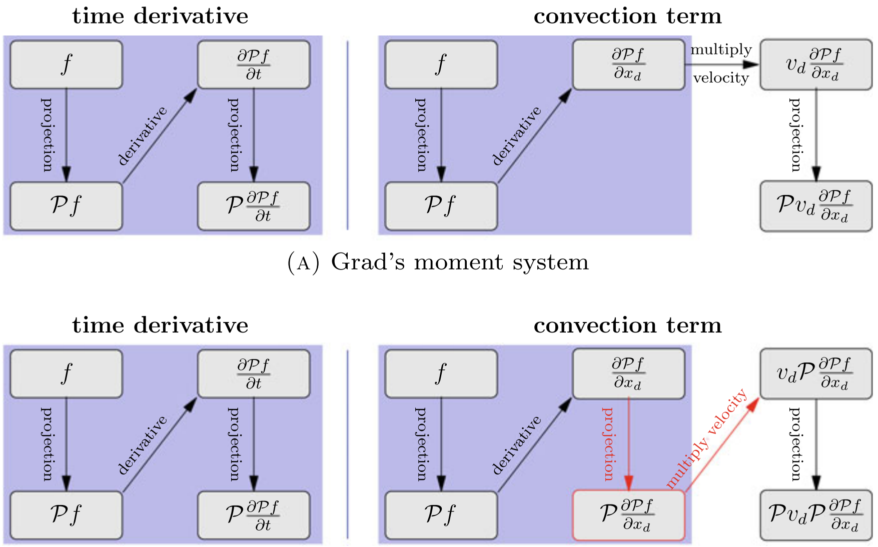
Accordingly, letting $\boldsymbol{Q}$ be the vector such that $\mathcal{P}Q(\mathcal{P}f) = \boldsymbol{\mathcal{H}}^T \boldsymbol{Q}$, one can rewrite Grad's moment system as

$$\mathbf{D}\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{d=1}^{D} \mathbf{B}_d \frac{\partial \boldsymbol{w}}{\partial x_d} = \boldsymbol{Q}. \tag{6}$$

Actually, the system (6) is the vector form of (4) in $\mathbb{H}_M$ with the basis $\mathcal{H}_\alpha$. By comparing these equations, we have the following correspondences

$$\boldsymbol{w} \leftrightarrow \mathcal{P}f, \quad \mathbf{D}\frac{\partial}{\partial t} \leftrightarrow \mathcal{P}\frac{\partial}{\partial t}, \quad \mathbf{B}_d \frac{\partial}{\partial x_d} \leftrightarrow \mathcal{P}v_d \frac{\partial}{\partial x_d}, \quad \boldsymbol{Q} \leftrightarrow \mathcal{P}Q(\mathcal{P}f). \tag{7}$$

Furthermore, we can diagram the procedure to derive Grad's moment system in Fig. 1a. It is noticed in [19] that the time derivative and the spatial derivative are treated differently in such a process, as a projection operator is applied directly to the time derivative, while for the spatial derivative, this projection operator appears only after the velocity v is multiplied. This difference causes the loss of hyperbolicity. By such observation, we have drawn a key conclusion in [19] that one should add a projection operator right in front of the spatial derivative to regain hyperbolicity, as is illustrated in Fig. 1b. The corresponding moment system is



(A) Grad's moment system

(B) Hyperbolic regularized moment system

Fig. 1 Diagram of the procedure of Grad's and regularized moment system

$$\mathcal{P}\frac{\partial \mathcal{P}f}{\partial t} + \sum_{d=1}^{D} \mathcal{P}v_d\mathcal{P}\frac{\partial \mathcal{P}f}{\partial x_d} = \mathcal{P}Q(\mathcal{P}f), \tag{8}$$

where the additional projection operator is labeled in red. Using (5), one can claim that there exist the square matrices $\mathbf{M}_d$, $d = 1, \ldots, D$ such that

$$\mathcal{P}v_d\mathcal{P}\frac{\partial \mathcal{P}f}{\partial x_d} = \boldsymbol{\mathcal{H}}^T\mathbf{M}_d\mathbf{D}\frac{\partial \boldsymbol{w}}{\partial x_d}, \tag{9}$$

and obtain the vector form of the regularized moment system as

$$\mathbf{D}\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{d=1}^{D}\mathbf{M}_d\mathbf{D}\frac{\partial \boldsymbol{w}}{\partial x_d} = \boldsymbol{Q}. \tag{10}$$

Similar to (7), we have one more correspondence:

$$\mathbf{M}_d \leftrightarrow \mathcal{P}v_d, \tag{11}$$

that is to say, the matrices $\mathbf{M}_d$ are the representation of the operators $\mathcal{P}v_d$ on $\mathbb{H}_M$. It is not difficult to check that the matrices $\mathbf{M}_d$ are symmetric due to the orthonormality of the basis $\mathcal{H}_\alpha$, so that any linear combination of the matrices $\mathbf{M}_d$ is real diagonalizable. One can also check the matrix $\mathbf{D}$ is invertible. Hence $\mathbf{D}^{-1}\mathbf{M}_d\mathbf{D}$ is similar to $\mathbf{M}_d$ so that the system (10) is globally hyperbolic. Moreover, if one multiplies $\mathbf{D}^T$ on both sides of (10), the resulting system

$$\mathbf{D}^T\mathbf{D}\frac{\partial \boldsymbol{w}}{\partial t} + \sum_{d=1}^{D} \mathbf{D}^T\mathbf{M}_d\mathbf{D}\frac{\partial \boldsymbol{w}}{\partial x_d} = \mathbf{D}^T\boldsymbol{Q} \tag{12}$$

turns out to be a symmetric hyperbolic system of balance laws.

2.2 Hyperbolic Regularization Framework

Till now, the hyperbolicity of (10) has been proved using the Condition 3. Looking back on the whole procedure, one can find that the key point of the hyperbolic regularization is the extra projection operator in front of the spatial differentiation operator in (8). Meanwhile, the underlying mechanism to obtain hyperbolicity can be extended to much more general cases. For example, the radiative transfer equation has the form

$$\begin{aligned}\frac{\partial f(t,\boldsymbol{x},\theta,\varphi)}{\partial t} + \boldsymbol{\xi}(\theta,\varphi)\cdot\nabla_{\boldsymbol{x}} f(t,\boldsymbol{x},\theta,\varphi) =& Q(f)(t,\boldsymbol{x},\theta,\varphi),\\ & \boldsymbol{x}\in\mathbb{R}^3,\quad \theta\in[0,\pi),\quad \varphi\in[0,2\pi),\end{aligned}$$

where the velocity is given by $\boldsymbol{\xi}(\theta,\varphi) = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)^T$. To derive reduced models, one can replace the local equilibrium f_{eq} in (2) by a nonnegative weight function ω, and correspondingly, the orthogonal polynomials He_α should be replaced by the orthogonal basis functions ϕ_α for the L^2 space weighted by ω, so that the basis functions $\mathcal{H}_\alpha$ become $\Phi_\alpha := \omega\phi_\alpha$. By letting $\mathbb{H}_M := \operatorname{span}\{\Phi_\alpha : |\alpha| \le M\}$, one can similarly define the projection operator $\mathcal{P}$ as in (3). As an extension of the globally hyperbolic moment system, we obtain

$$\mathcal{P}\frac{\partial \mathcal{P}f}{\partial t} + \sum_{d=1}^{D}\mathcal{P}\xi_d(\theta,\varphi)\mathcal{P}\frac{\partial \mathcal{P}f}{\partial x_d} = \mathcal{P}Q(\mathcal{P}f). \tag{13}$$

Again, if the corresponding matrix $\mathbf{D}$ as in (6) is invertible, the resulting moment system is globally hyperbolic. We refer the readers to [6, 19, 21] for more details of such applications in radiative transfer equations.

This framework provides a concise and clear procedure to derive the hyperbolic moment system from a broad range of kinetic equations. It has been applied to many fields, including anisotropic hyperbolic moment system for Boltzmann equation [20], semiconductor device simulation [7], plasma simulation [11], density functional theory [8], quantum gas kinetic theory [16], and rarefied relativistic Boltzmann equation [32].

2.3 Further Progress

The above framework provides an approach to handling the hyperbolicity of the moment system. However, the hyperbolicity is not the only concerned property. Preserving the hyperbolicity and other properties at the same time is often required in model reduction. Below we will list some recent attempts in this direction.

One of the interesting properties is to recover the asymptotic limits of the kinetic equations. For example, the first-order asymptotic hydrodynamic limit of the Boltzmann equation is the Navier-Stokes equations, and therefore it is desirable that the moment equations can preserve such a limit. For the classical Boltzmann equation, most moment systems can automatically preserve the Navier-Stokes limit if the stress tensor and heat flux are included. However, for the quantum Boltzmann equation, the equilibrium has a very special form, so that the moment system directly derived from the framework by taking the equilibrium as the weight function disobeys the Navier-Stokes limit [16]. In this case, the authors of [16] proposed a method called *local linearization* to regularize the moment system. Specifically, we assume the Grad-type system has the form as (6) and define $\hat{\mathbf{M}}_d(\boldsymbol{w}) = \mathbf{B}_d(\boldsymbol{w})\mathbf{D}(\boldsymbol{w})^{-1}$. In the regularization, the matrix $\hat{\mathbf{M}}_d(\boldsymbol{w})$ is replaced by $\mathbf{M}_d := \hat{\mathbf{M}}_d(\boldsymbol{w}_{eq})$ with $\boldsymbol{w}_{eq}$ being the local equilibrium of the state $\boldsymbol{w}$. Such a method allows us to acquire both the hyperbolicity and Navier-Stokes limit simultaneously. The symmetry of $\mathbf{M}$ is thereby lost so that one has to use Condition 1 to prove the hyperbolicity.

Another relevant work is the nonlinear moment system for radiative transfer equation in [21, 22]. In order to retain the diffusion limit (similar to the Navier-Stokes limit for the Boltzmann equation), the authors pointed out that the projection operators in (13) at different places do not have to be same and revised (13) to be

$$\tilde{\mathcal{P}}\frac{\partial \mathcal{P} f}{\partial t} + \sum_{d=1}^{D} \tilde{\mathcal{P}}\xi_d(\theta,\varphi)\tilde{\mathcal{P}}\frac{\partial \mathcal{P} f}{\partial x_d} = \tilde{\mathcal{P}}Q(\mathcal{P} f). \tag{14}$$

The operators $\mathcal{P}$ and $\tilde{\mathcal{P}}$ are orthogonal projections onto different subspaces of $\mathbb{H}$. By a careful choice of the subspace for the operator $\tilde{\mathcal{P}}$, the diffusion limit can be achieved, and meanwhile, the symmetry of $\mathbf{M}$ corresponding to that in (10) is preserved, leading again to global hyperbolicity. This generalization has broadened the application the hyperbolic regularization framework and also permits us to take more properties of the kinetic equations into account.

Besides the hyperbolicity for the convection term, one may also be interested in the wellposedness of the complete moment system including the collision term. One related property is Yong's first stability condition [38], which includes the constraints on the convection term, collision term, and the coupling of both. This stability condition is shown to be critical for the existence of the solutions in [37]. In [17], the authors have studied multiple Grad-type moment systems and confirmed that all of these systems satisfy Yong's first stability condition.

Under this concise and flexible framework, one may wonder what is sacrificed for the hyperbolicity. By writing out the equations, one can immediately observe that the form of balance law is ruined by the hyperbolic regularization. A natural question is: how to define the discontinuity in the solution? More generally, one may ask: what is the effect of such a regularization on the accuracy of the model? In the following section, we will provide some clues using numerical experiments.

3 Numerical Validation

The application of the framework in the gas kinetic theory has been investigated in a number of works [3, 9, 10, 12], where many one- and two-dimensional examples have been numerically studied to show the validity of hyperbolic moment equations. However, these globally hyperbolic models, as an improvement of Grad's original models, have never been compared with Grad's models in terms of the modeling accuracy. The only direct comparison seen in the literature is in [10], wherein for a shock tube problem with a density ratio of 7.0, the simulation of Grad's moment equations breaks down and the corresponding hyperbolic moment equations appear to be stable. Without running numerical tests for the same problem for which both models work and comparing the results, it could be questioned whether we lose accuracy when fixing the hyperbolicity. Such doubt may arise since the globally hyperbolic models can be considered as a partial linearization of Grad's models about the local Maxwellians.

In this section, we will make such straightforward comparison using the same numerical examples for both methods. For simplicity, we only consider the one-dimensional physics, for which both x and v are scalars. In this case, the characteristic polynomial for the Jacobian of the flux function has an explicit formula [3], so that the hyperbolicity of Grad's equation can be easily checked. The underlying kinetic equation used in our test is the Boltzmann-BGK equation with a constant relaxation time

$$\frac{\partial f}{\partial t} + v\frac{\partial f}{\partial x} = \frac{1}{Kn}(f_{eq} - f). \tag{15}$$

The ansatz of the distribution function is given by (3), so that (4) stands for Grad's moment system, and (8) stands for the hyperbolic moment system. Below we are going to use two benchmark tests to show the performance of both types of models. In general, both Grad's moment equations and the hyperbolic moment equations are solved by the first-order finite volume method with local Lax-Friedrichs numerical flux. Time splitting is applied to solve the advection part and the collision part separately, and for each part, the forward Euler method is applied. The CFL condition is utilized to determine the time step, and the Courant number is chosen as 0.9. For Grad's moment method, the maximum characteristic speed is obtained by solving the roots of the characteristic polynomial of the Jacobian, and the explicit expression of the charateristic polynomial has been given in [3]. For the hyperbolic moment method, the maximum characteristic speeds have been computed in [3]. The explicit form of the hyperbolic moment system (given in [3]) shows that its last equation contains a non-conservative product, which is discretized by central difference. In all the numerical examples, the number of grid cells is 1000 if not otherwise specified. We have done the convergence test showing that for smooth solutions, such a resolution can provide solutions sufficiently close to the solutions on a much finer grid, so that their difference is invisible to the naked eye. When exhibiting the numerical results, we will mainly focus on the equilibrium variables including density ρ, velocity u, and temperature θ, which are defined by

$$\begin{aligned}
\rho(t,x) &= \int_{\mathbb{R}} f(t,x,v)\,\mathrm{d}v, \\
u(t,x) &= \frac{1}{\rho(t,x)} \int_{\mathbb{R}} v f(t,x,v)\,\mathrm{d}v, \\
\theta(t,x) &= \frac{1}{\rho(t,x)} \int_{\mathbb{R}} [v - u(t,x)]^2 f(t,x,v)\,\mathrm{d}v.
\end{aligned}$$

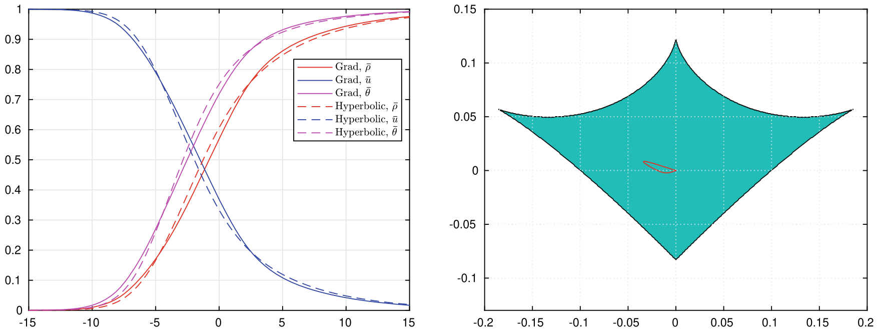
3.1 Shock Structure

The structure of plane shock waves is frequently used as a benchmark test in the gas kinetic theory. It shows that the physical shock, which appears to be a discontinuity in the Euler equations, is actually a smooth transition from one state to another. The computational domain is $(-\infty, +\infty)$ so that no boundary condition is involved, and the initial data are

$$f(0,x,v) = \begin{cases} \dfrac{\rho_l}{\sqrt{2\pi\theta_l}} \exp\left(-\dfrac{(v-u_l)^2}{2\theta_l}\right), & \text{if } x < 0, \\ \dfrac{\rho_r}{\sqrt{2\pi\theta_r}} \exp\left(-\dfrac{(v-u_r)^2}{2\theta_r}\right), & \text{if } x > 0, \end{cases} \tag{16}$$

where all the equilibrium variables are determined by the Mach number Ma:

$$\begin{aligned}
\rho_l &= 1, \qquad u_l = \sqrt{3}Ma, \qquad \theta_l = 1, \\
\rho_r &= \frac{2Ma^2}{Ma^2+1}, \\
u_r &= \frac{\sqrt{3}Ma}{\rho_r}, \\
\theta_r &= \frac{3Ma^2-1}{2\rho_r}.
\end{aligned}$$



(A) Grad vs HME

(B) Phase diagram of Grad's solution

Fig. 2 Left: The comparison of shock structures of two solutions with Mach number 1.4 and $M = 4$. Right: The green area is the hyperbolicity region (horizontal axis: $\hat{f}_{M-1}$, vertical axis: $\hat{f}_M$), and the red loop is the parametric curve $(\hat{f}_{M-1}, \hat{f}_M)$ with parameter x

We are interested in the steady-state of this problem. Since the parameter *Kn* only introduces a uniform spatial scaling, it does not affect the shock structure. Therefore we simply set it to be 1. Numerically, we set the computational domain to be $[-30, 30]$. The boundary condition is provided by the ghost-cell method, and the distribution functions on the ghost cells are set to be the two states defined in (16).

3.1.1 Case 1: *Ma* = 1.4 and *M* = 4

In this case, both Grad's system and the hyperbolic moment system work due to the relatively small Mach number. The numerical results are shown in Fig. 2. By convention, we plot the normalized density, velocity, and temperature defined by

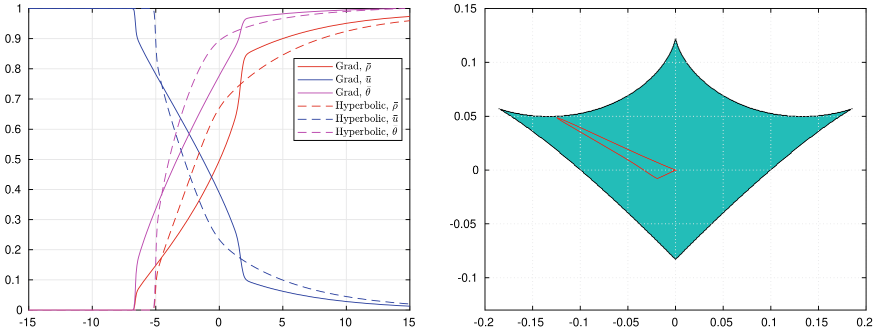
$$\bar{\rho}(x) = \frac{\rho(x) - \rho_l}{\rho_r - \rho_l}, \quad \bar{u}(x) = \frac{u(x) - u_r}{u_l - u_r}, \quad \bar{\theta}(x) = \frac{\theta(x) - \theta_l}{\theta_r - \theta_l},$$

so that the value of all variables are generally within the range [0, 1], unless the temperature overshoot is observed.

Figure 2b shows the hyperbolicity region of Grad's moment equations. It has been proven in [3] that for the one-dimensional physics, the hyperbolicity region can be characterized by the following two dimensionless quantities:

$$\hat{f}_{M-1} = \frac{f_{M-1}}{\rho\theta^{(M-1)/2}}, \qquad \hat{f}_M = \frac{f_M}{\rho\theta^{M/2}},$$

where f_M and f_{M-1} are the last two coefficients in the expansion (3). The red curve in Fig. 2b provides the trajectory of Grad's solution in this diagram. It can be seen that for such a small Mach number, the whole solution is well inside the hyperbolicity



(A) Grad vs HME

(B) Phase diagram of Grad's solution

Fig. 3 Left: The comparison of shock structures of two solutions with Mach number 2.0 and $M = 4$. Right: The green area is the hyperbolicity region (horizontal axis: $\hat{f}_{M-1}$, vertical axis: $\hat{f}_M$), and the red loop is the parametric curve $(\hat{f}_{M-1}, \hat{f}_M)$ with parameter x

region, so that the simulation of Grad's moment equations is stable. Figure 2a shows that both methods provide smooth shock structures, and the predictions for all the equilibrium variables are similar. This example confirms the applicability of both systems in weakly non-equilibrium regimes. Note that for one-dimensional physics, Grad's equations do not suffer form the loss of hyperbolicity near equilibrium.

3.1.2 Case 2: $Ma = 2.0$ and $M = 4$

Now we increase the Mach number to introduce stronger non-equilibrium. The same plots are provided in Fig. 3. In this example, despite the numerical diffusion, discontinuities can be identified without difficulty from the numerical solutions. These discontinuities, also known as subshocks, appear due to the insufficient characteristic speed in front of the shock wave, meaning that both systems are insufficient to describe the physics. To capture these discontinuities, 8000 grid cells are used in the spatial discretization. This example shows significantly different shock structures predicted by both methods. For Grad's moment equations, the subshock locates near $x = -7$, while for hyperbolic moment equations, the subshock appears near $x = -5$. The wave structures also differ a lot. By focusing on the high-density region, we find that the solution of hyperbolic moment equations is smoother, showing the possibly better description of the physics.

Here we remind the readers that the wave structure of hyperbolic moment equations may depend on the numerical method, due to its non-conservative nature. The locations and the strengths of the subshock may change when using the different shock conditions. However, we would like to argue that it is meaningless to justify any solution with subshocks for the hyperbolic moment equations, for it is unphysical and should not appear in the solution of the Boltzmann equation. In practice, the appearance of discontinuous solutions is an indication of the inadequate truncation of series, which inspires us to increase M to get more reliable solutions without subshocks.

Figure 3b shows that Grad's solution still locates within the hyperbolicity region, although the curve is already quite close to the boundary of the region. This example shows that even in its hyperbolicity region, Grad's moment method may lose its validity.

3.1.3 Case 3: $Ma = 2.0$ and $M = 6$

Now we try to increase M and carry out the simulation again for Mach number 2.0. The results are given in Fig. 4. With the hope that a larger M can provide a better solution, we actually see that Grad's moment equations lead to computational failure. The numerical solution before the computation breaks down is plotted in Fig. 4a. Figure 4b clearly shows that this is caused by the loss of hyperbolicity. We believe that this implies the non-existence of the solution.

On the contrary, the simulation of hyperbolic moment equations is still stable. As expected, it provides a smooth shock structure and improves the result predicted by $M = 4$.

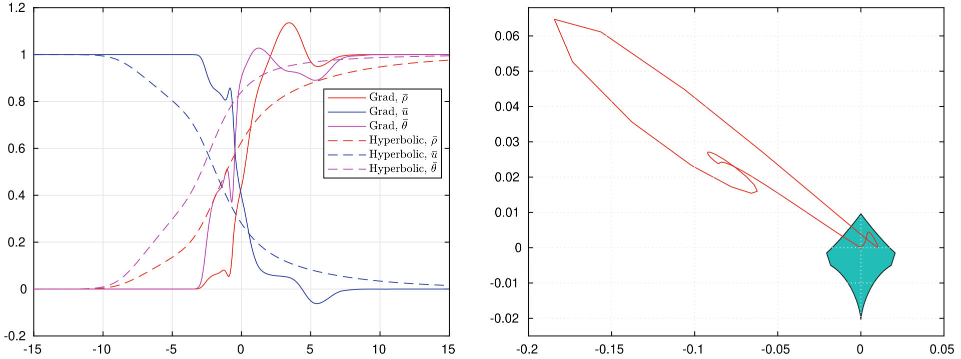
3.1.4 Case 4: $Ma = 1.7$ and $M = 6$

In this example, we decrease the Mach number so that the shock structure of Grad's equations can be found. Figure 5a shows that the results of both systems generally agree with each other, but it can be observed that hyperbolic moment equations provide smoother solutions than Grad's system, so that it is likely to be more accurate. Therefore, despite the higher nonlinearity of Grad's system, it does not necessarily help provide better solutions.

Interestingly, when looking at the phase diagram plotted in Fig. 5b, we see that Grad's solution has run out of the hyperbolicity region. It is to be further studied why the solution is still stable. Here we would like to conjecture that the collision term and the numerical diffusion help stabilize the numerical solution in the evolutionary process, and for the steady-state equations, solutions for non-hyperbolic equations may still exist. Nevertheless, all the above numerical tests show the superiority of hyperbolic moment equations for both accuracy and stability.

3.1.5 Case 5: $Ma = 2.0$ and $M = 10$

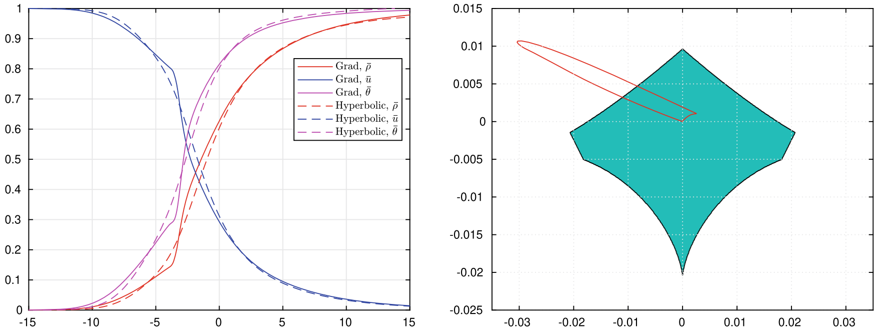
In this example, we would like to show the failure of both systems for a larger M. In Fig. 6, we plot the results at $t = 0.8$, where both numerical solutions contain negative temperatures. In [28], the reason for such a phenomenon has been explained, which lies in the divergence of the approximation (3) as M tends to infinity. It is



(A) Grad vs HME

(B) Phase diagram of Grad's solution

Fig. 4 Left: The shock structure of hyperbolic moment equations for Mach number 2.0 and $M = 6$, and Grad's solution before computational failure ($t = 1.0$). Right: The green area is the hyperbolicity region (horizontal axis: $\hat{f}_{M-1}$, vertical axis: $\hat{f}_M$), and the red loop is the parametric curve $(\hat{f}_{M-1}, \hat{f}_M)$ with parameter x

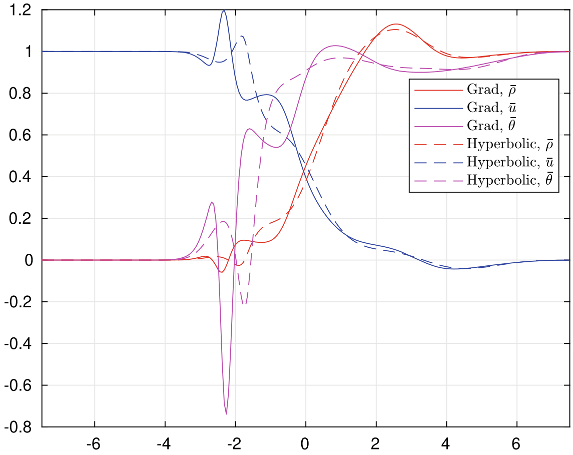


(A) Grad vs HME

(B) Phase diagram of Grad's solution

Fig. 5 Left: The comparison of shock structures of two solutions with Mach number 1.7 and $M = 6$. Right: The green area is the hyperbolicity region (horizontal axis: $\hat{f}_{M-1}$, vertical axis: $\hat{f}_M$), and the red loop is the parametric curve $(\hat{f}_{M-1}, \hat{f}_M)$ with parameter x

Fig. 6 The numerical solution at $t = 0.8$ for Mach number 2.0 and $M = 10$



rigorously shown in [13] that when $\theta_r > 2\theta_l$, for the solution of the steady-state BGK equation, the limit of $\mathcal{P}f$ (see (3)) as $M \to \infty$ does not exist. Here for $Ma = 2.0$, the temperature behind the shock wave is $\theta_r = 55/16 > 2 = 2\theta_l$. Thus for a large M, the divergence leads to a poor approximation of the distribution function, and it is reflected as a negative temperature in the numerical results. Such a divergence issue is independent of the subshock and the hyperbolicity, and should be regarded as a defect for both systems. The work on fixing the issue is ongoing.

3.2 Fourier Flow

In this test, we are interested in the performance of both methods with wall boundary conditions. The fluid we are concerned about is between two fully diffusive walls locating at $x = -1/2$ and $x = 1/2$. For the Boltzmann-BGK equation (15), the boundary condition is

$$\begin{aligned} f(t, -1/2, v) &= \frac{\rho_l}{\sqrt{2\pi\theta_l}} \exp\left(-\frac{v^2}{2\theta_l}\right), \qquad v > 0, \\ f(t, 1/2, v) &= \frac{\rho_r}{\sqrt{2\pi\theta_r}} \exp\left(-\frac{v^2}{2\theta_r}\right), \qquad v < 0, \end{aligned}$$

where $\theta_{l,r}$ stands for the temperature of the walls, and $\rho_{l,r}$ is chosen such that

$$\int_{\mathbb{R}} v f(t, \pm 1/2, v)\,\mathrm{d}v = 0.$$

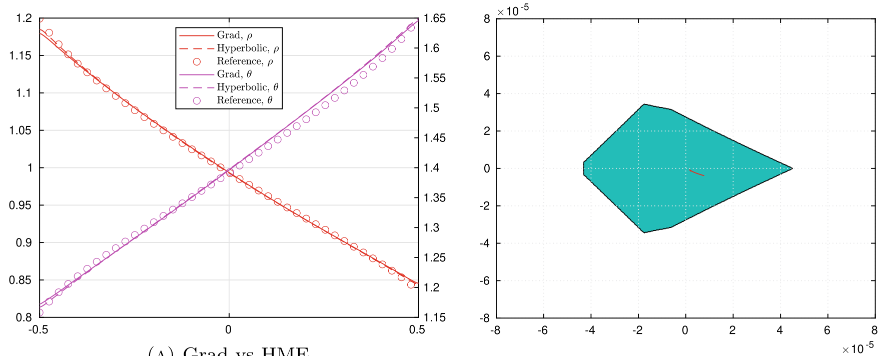
Following [24], the boundary conditions of moment equations can be derived by taking odd moments of the diffusive boundary condition. We choose the initial condition as

$$f(0, x, v) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) \tag{17}$$

for all x. Again we are concerned only about the steady-state of the solution.

In our numerical experiments, we choose $Kn = 0.3$, $\theta_l = 1$ and $M = 11$. Two test cases with $\theta_r = 1.9$ and $\theta_r = 2.7$ are considered. For the smaller temperature ratio $\theta_r = 1.9$, the numerical results are given in Fig. 7, where two solutions mostly agree with each other. The reference solution, computed using the discrete velocity model, is also provided in Fig. 7a. It can be seen that both models provide reasonable approximations to the reference solution. The good behavior of Grad's solutions can also be predicted by the phase diagram in Fig. 7b, from which one can observe that the whole solution locates in the central area of the hyperbolicity region.

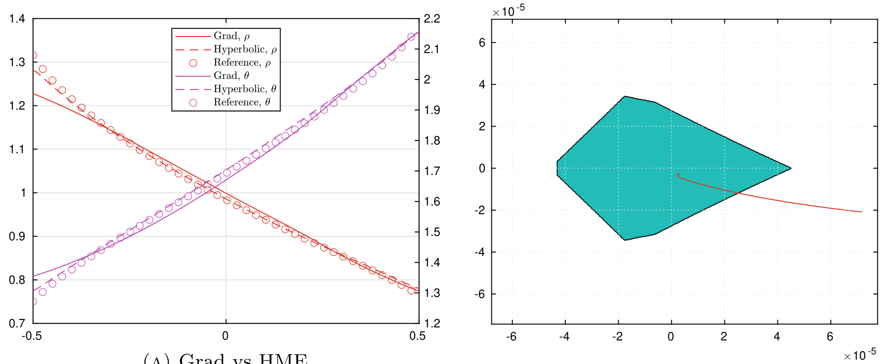
For $\theta_r = 2.7$, the results are plotted in Fig. 8. In this case, if we start the simulation of Grad's equations from the initial data (17), the computation will break down due to the loss of hyperbolicity in the evolutional process. Therefore, we first run the simulation for hyperbolic moment equations from the initial data (17) and evolve the solution to the steady-state. Afterward, this steady-state solution serves as the initial data of Grad's equations. Although the steady-state solution of Grad's equations can be found using this technique, the approximation looks poorer than hyperbolic moment equations. The phase diagram (Fig. 8b) shows that the solution near the left wall is outside the hyperbolicity region, so that the validity of boundary conditions on the left wall becomes unclear. In contrast, the hyperbolic moment equations still provide reliable approximation despite the high temperature ratio.



(A) Grad vs HME

(B) Phase diagram of Grad's solution

Fig. 7 Left: Steady Fourier flow for $\theta_r = 1.9$ (left vertical axis: ρ, right vertical axis: θ). Right: The green area is the hyperbolicity region (horizontal axis: $\hat{f}_{M-1}$, vertical axis: $\hat{f}_M$), and the red line is the parametric curve $(\hat{f}_{M-1}, \hat{f}_M)$ with parameter x



(A) Grad vs HME

(B) Phase diagram of Grad's solution

Fig. 8 Left: Steady Fourier flow for $\theta_r = 2.7$ (left vertical axis: ρ, right vertical axis: θ). Right: The green area is the hyperbolicity region (horizontal axis: $\hat{f}_{M-1}$, vertical axis: $\hat{f}_M$), and the red line is the parametric curve $(\hat{f}_{M-1}, \hat{f}_M)$ with parameter x

3.3 A Summary of Numerical Experiments

In all the above numerical experiments, we see that despite the loss of some nonlinearity, the hyperbolicity fix does not appear to lose accuracy in any of the numerical tests. In regimes with moderate non-equilibrium effects, Grad's equations may provide solutions outside the hyperbolicity region without numerical instability. In this situation, our experiments show that the hyperbolicity fix is likely to improve the accuracy of the model. It has also been demonstrated that other issues, such as subshocks and divergence, are not related to the hyperbolicity, and these issues have to be addressed independently.

4 Conclusion

The loss of hyperbolicity, as one of the major obstacles for the model reduction in gas kinetic theory, is almost cleared through the research works in recent years. With a handy framework introduced in Sect. 2, we can safely move our focus of model reduction to other properties such as the asymptotic limit, the stability, and the convergence issues. Our numerical experiments show that the hyperbolic regularization does not harm the accuracy of the model. It is our hope that such a framework can inspire more thoughts in the development of dimensionality reduction even beyond the kinetic theory.

References

1. Boltzmann, L.: Weitere studien über das wärmegleichgewicht unter gas-molekülen. Wiener Berichte **66**, 275–370 (1872)
2. Bourgault, Y., Broizat, D., Jabin, P.-E.: Convergence rate for the method of moments with linear closure relations. Kinetic Rel. Mod. **8**(1), 1–27 (2015)
3. Cai, Z., Fan, Y., Li, R.: Globally hyperbolic regularization of Grad's moment system in one dimensional space. Comm. Math. Sci. **11**(2), 547–571 (2013)
4. Cai, Z., Fan, Y., Li, R.: Globally hyperbolic regularization of Grad's moment system. Comm. Pure Appl. Math. **67**(3), 464–518 (2014)
5. Cai, Z., Fan, Y., Li, R.: On hyperbolicity of 13-moment system. Kin. Rel. Mod. **7**(3), 415–432 (2014)
6. Cai, Z., Fan, Y., Li, R.: A framework on moment model reduction for kinetic equation. SIAM J. Appl. Math. **75**(5), 2001–2023 (2015)
7. Cai, Z., Fan, Y., Li, R., Lu, T., Wang, Y.: Quantum hydrodynamic model by moment closure of Wigner equation. J. Math. Phys. **53**(10), 103503 (2012)
8. Cai, Z., Fan, Y., Li, R., Lu, T., Yao, W.: Quantum hydrodynamic model of density functional theory. J. Math. Chem. **51**(7), 1747–1771 (2013)
9. Cai, Z., Fan, Y., Li, R., Qiao, Z.: Dimension-reduced hyperbolic moment method for the Boltzmann equation with BGK-type collision. Commun. Comput. Phys. **15**(5), 1368–1406 (2014)
10. Cai, Z., Li, R., Qiao, Z.: Globally hyperbolic regularized moment method with applications to microflow simulation. Comput. Fluids **81**, 95–109 (2013)
11. Cai, Z., Li, R., Wang, Y.: Solving Vlasov equation using NRxx method. SIAM J. Sci. Comput. **35**(6), A2807–A2831 (2013)
12. Cai, Z., Torrilhon, M.: Numerical simulation of microflows using moment methods with linearized collision operator. J. Sci. Comput. **74**(1), 336–374 (2018)
13. Cai, Z., Torrilhon, M.: On the Holway-Weiss debate: Convergence of the Grad-moment-expansion in kinetic gas theory. Phys. Fluids **31**, 126105 (2019)
14. Cercignani, C.: Mathematical Methods in Kinetic Theory. Springer, New York (1969)
15. Chapman, S.: On the law of distribution of molecular velocities, and on the theory of viscosity and thermal conduction, in a non-uniform simple monatomic gas. Phil. Trans. R. Soc. A **216**(538–548), 279–348 (1916)
16. Di, Y., Fan, Y., Li, R.: 13-moment system with global hyperbolicity for quantum gas. J. Stat. Phys. **167**(5), 1280–1302 (2017)
17. Di, Y., Fan, Y., Li, R., Zheng, L.: Linear stability of hyperbolic moment models for Boltzmann equation. Num. Math. Theory Method and Appl. **10**(2), 255–277 (2017)

18. Enskog, D.: The numerical calculation of phenomena in fairly dense gases. Arkiv Mat. Astr. Fys. **16**(1), 1–60 (1921)
19. Fan, Y., Koellermeier, J., Li, J., Li, R., Torrilhon, M.: Model reduction of kinetic equations by operator projection. J. Stat. Phys. **162**(2), 457–486 (2016)
20. Fan, Y., Li, R.: Globally hyperbolic moment system by generalized Hermite expansion. Scientia Sinica Mathematica **45**(10), 1635–1676 (2015)
21. Fan, Y., Li, R., Zheng, L.: A nonlinear hyperbolic model for radiative transfer equation in slab geometry. arXiv preprint arXiv:1911.05472 (2019)
22. Fan, Y., Li, Ruo, Zheng, L.: A nonlinear moment model for radiative transfer equation in slab geometry. J. Comput. Phys. **404**, 109128 (2020)
23. Gombosi, T.I.: Gaskinetic Theory. Cambridge University Press (1994)
24. Grad, H.: On the kinetic theory of rarefied gases. Comm. Pure Appl. Math. **2**(4), 331–407 (1949)
25. Grad, H.: The profile of a steady plane shock wave. Comm. Pure Appl. Math. **5**(3), 257–300 (1952)
26. Grad, H.: Principles of the kinetic theory of gases. Handbuch der Physik **12**, 205–294 (1958)
27. Harris, S.: An introduction to the theory of the Boltzmann equation. Rinehart and Winston Inc., Holt (1971)
28. Holway, L.H.: Existence of kinetic theory solutions to the shock structure problem. Phys. Fluids **7**(6), 911–913 (1965)
29. Jeans, J.H.: An Introduction to The Kinetic Theory of Gases. Cambridge University Press, Cambridge (1967)
30. Junk, M.: Domain of definition of Levermore's five-moment system. J. Stat. Phys. **93**(5), 1143–1167 (1998)
31. Koellermeier, J., Schaerer, R., Torrilhon, M.: A framework for hyperbolic approximation of kinetic equations using quadrature-based projection methods. Kinet. Relat. Mod. **7**(3), 531–549 (2014)
32. Kuang, Yangyu, Tang, Huazhong: Globally hyperbolic moment model of arbitrary order for one-dimensional special relativistic Boltzmann equation. J. Stat. Phys. **167**(5), 1303–1353 (2017)
33. Levermore, C.: Moment closure hierarchies for kinetic theories. J. Stat. Phys. **83**(5–6), 1021–1065 (1996)
34. McDonald, J., Torrilhon, M.: Affordable robust moment closures for CFD based on the maximum-entropy hierarchy. J. Comput. Phys. **251**, 500–523 (2013)
35. Müller, I., Ruggeri, T.: Extended thermodynamics. Springer Tracts in Natural Philosophy, vol. 37. Springer, New York (1993)
36. Müller, I., Ruggeri, T.: Rational Extended Thermodynamics. Springer Tracts in Natural Philosophy, vol. 37, 2nd edn. Springer, New York (1998)
37. Peng, Y., Wasiolek, V.: Uniform global existence and parabolic limit for partially dissipative hyperbolic systems. J. Diff. Equ. (2016)
38. Yong, W.-A.: Singular perturbations of first-order hyperbolic systems with stiff source terms. J. Diff. Eq. **155**(1), 89–132 (1999)
39. Zhao, W., Yong, W., Luo, L.: Stability analysis of a class of globally hyperbolic moment system. Commun. Math. Sci. **15**(3), 609–633 (2016)

Cryptography and Digital Transformation

Kazue Sako

Abstract Cryptography is implemented using discrete mathematics with security defined in complexity theory. In this article, we review some cryptographic primitives for encryption, signing messages and interactive proofs. By combining cryptographic primitives, we can design and digitally implement various services with desired features in security, privacy and fairness. We will discuss some examples such as electronic voting and cryptocurrencies.

1 Digital Transformation

Research in mathematics and cryptography play a big role in shaping our digitalized society much better in coming years. There is an immense expectation that technology on Information and Communications, known as ICT, would transform our life to be more efficient, more productive and more functional. However, these are bright side of digital transformation. We also need to take care to transform ‘correctly’ so that we do not suffer from unexpected consequences.

One evident characteristic of ICT is that it makes us free from physical constraints. Digital data have little weight and thus we can make thousand copies and travel thousand miles at once. While this characteristic brings benefit, it also brings threats to our life. We need alternative ways to create ‘constraints’ to those who is willing to harm us, and one promising approach to creating such constraints is use of cryptography.

Cryptography started as a way to conceal information. We were able to design cryptographic algorithm that is computationally infeasible to recover the message without knowledge of a decryption key. There are rigorous mathematical proofs that guarantee that indeed this characteristic holds based on some hard problems, like NP problems or factorization. So this computational difficulty would serve as an alternative constraints in a digital world.

K. Sako (✉)
Waseda University, Tokyo, Japan
e-mail: kazuesako@aoni.waseda.jp

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_9

In this article, we provide two examples of use cryptography to implement secure digital systems. One is digitalization of voting system, and the other is digitalization of payment system called Bitcoin. Prior to these two examples we oversee some cryptographic primitives such as encryption schemes, digital signature schemes and interactive proofs.

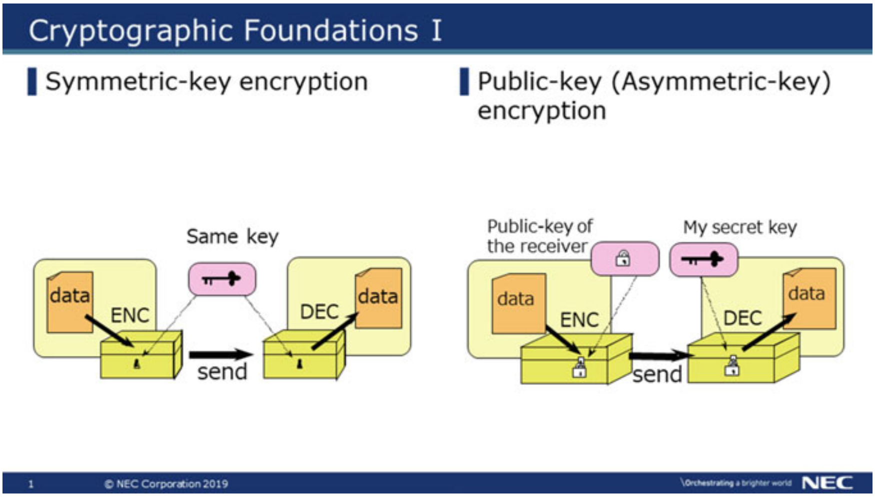
2 Cryptographic Foundations

In this section, we will introduce three fundamental notions in cryptography. They are Encryption Schemes, Digital Signature Schemes and Interactive Proofs.

2.1 Encryption Schemes

First, we begin by introducing two types of encryption schemes, depending on how we use keys. The first type, which is called Symmetric-key encryption schemes, uses the same key for both encryption and decryption. This type of encryption schemes existed since the age of Gaius Julius Caesar. The new type of encryption is called Publickey encryption schemes or Asymmetric-key encryption schemes, where we use different keys for encryption and decryption. Moreover, the key to encrypt data can be made public (Fig. 1).

Let us briefly discuss some mathematical model to define encryption schemes and its security. Encryption schemes, either symmetric or asymmetric, can be mod-



Cryptographic Foundations I

Symmetric-key encryption

Public-key (Asymmetric-key) encryption

Fig. 1 Two types of encryption schemes

eled in three non-deterministic functions, namely KeyGeneration, Encryption and Decryption, with a security parameter k. KeyGeneration, on input k, outputs a key pair EncKey and DecKey. (In case of Symmetric Key encryption schemes, EncKey = DecKey holds.) Encryption Function, given a message m from its domain and EncKey, outputs a ciphertext c.

$$c = \text{Encryption}(k, m, \text{EncKey})$$

Similarly, Decryption Function, given a ciphertext c from its domain and DecKey, outputs a message m'.

$$m' = \text{Decryption}(k, c, \text{DecKey})$$

A triplet of nondeterministic functions (KeyGeneration, Encryption, Decryption) is called Encryption scheme if and only if: For any k, for any output (EncKey, DecKey) of KeyGeneration on input k, and for any message in m,

$$m = \text{Decryption}(k, \text{Encryption}(k, m, \text{EncKey}), \text{DecKey})$$

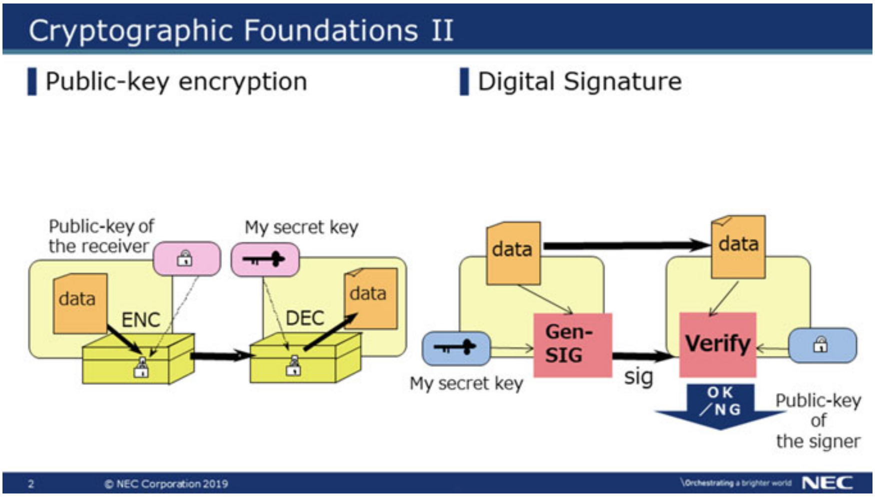
holds.

As seen in the definition, even an Encryption function that returns m as c is an Encryption Scheme. So we need to define what property we need to call an Encryption Scheme secure. Cryptographers had studied various ways to do this. A fundamental observation is: given any two messages m_1 and m_2, and given any ciphertext c_i of either m_1 or m_2, the encryption scheme is secure if no one can guess to which message a ciphertext c decrypts to with probability more than half. To be more rigorous, we need to define this in an asymptotic manner. That is, if we chose large enough k, the probability of guessing can be made larger than $1/2 + \epsilon$. We note that in Asymmetric Encryption Schemes, guessing is hard even if they know EncKey that was used to create c. There are various other security definitions for Encryption Schemes, be it strong or weak [1].

To prove security of some concrete Encryption Schemes, we assume existence of some one-way functions or some difficult problems like factorization.

2.2 Digital Signature Schemes

Another exciting tools related to Public Key Encryption Schemes are Digital Signature Schemes. If we can have two related keys PubKey and PrivKey, where one can publish PubKey without worrying about secrecy of PrivKey, we can construct a scheme that serves as Digital Signatures. A person would sign a message with PrivKey and outputs a signature sig. Anyone can verify whether or not the signature was generated using a key corresponding to PubKey, by performing Verification (Fig. 2).



Cryptographic Foundations II

Public-key encryption

Digital Signature

Fig. 2 Digital signature schemes

Similarly, Digital Signature Scheme is modeled by three nondeterministic functions (KeyGen, Gen-SIG, Verify). KeyGen, on input security parameter k, outputs a key pair PrivKey and PubKey. Gen-SIG Function, given a message m from its domain and PrivKey, outputs a signature sig.

$$\text{sig} = \text{Gen-SIG}(k, m, \text{PrivKey})$$

Verify Function, given a signature sig from its domain, the message m and PubKey, outputs either OK or NG.

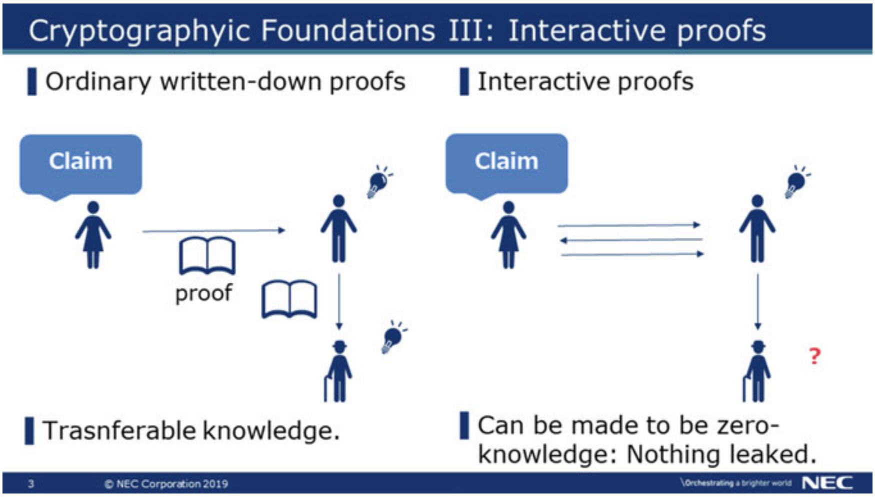
$$\text{OK/NG} = \text{Verify}(k, \text{sig}, m, \text{PubKey})$$

A triplet of nondeterministic functions (KeyGen, Gen-SIG, Verify) is called Signature scheme if and only if: For any k, for any output (PrivKey, PubKey) of KeyGeneration on input k, and for any message in m,

$$\text{OK} = \text{Verify}(k, \text{Gen-SIG}(k, m, \text{PrivKey}), m, \text{PubKey})$$

holds.

For security of signature schemes, we want to claim that it is only a person who knows PrivKey can generate sig corresponding to m that the Verify Function outputs OK. For this purpose, we claim a Signature Scheme is secure if there is an algorithm that can generate signatures that Verify outputs OK, then we can use the algorithm to ‘extract’ PrivKey. For sake of space, please refer to reference [1] for more mathematical definition for security of digital signature schemes.



Cryptographyic Foundations III: Interactive proofs

- Ordinary written-down proofs
- Interactive proofs

- Trasnferable knowledge.
- Can be made to be zero-knowledge: Nothing leaked.

3 © NEC Corporation 2019 \Orchestrating a brighter world NEC

Fig. 3 Interactive proofs

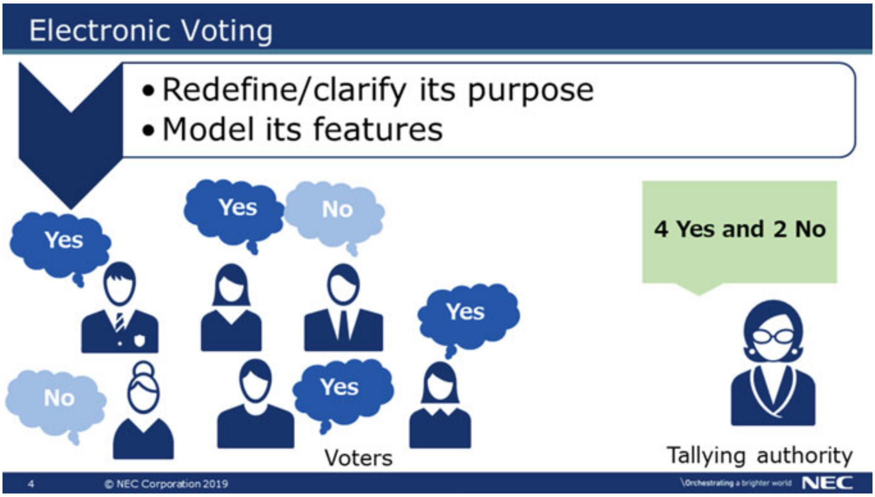
2.3 Interactive Proofs

The last primitive we will discuss in the article is Interactive Proofs. In Mathematics, when we say Proof, it is usually something that can be written down in the paper and those who have seen the Proof can verify the correctness of its claim. So the script of Proof is non-interactive. The Prover alone would generate the script of Proof by himself. Also the script of Proof is transferable, that any party who have seen the Proof can verify that the claim is correct.

Instead, there are protocols where Prover and Verifier talks interactively and at the end Verifier is persuaded that the Claim is correct. This is called Interactive Proofs (Fig. 3). This type of interactive proofs can provide further characteristic that the Verifier learn nothing from the interaction except that the Claim is correct. That is, Verifier learned no knowledge or zero knowledge in engaging the proof protocol. These types of protocols are called Zero Knowledge Interactive Proofs, which are frequently used in cryptographic protocols. Because the Verifier learned no new knowledge, he cannot prove to a third party that the Claim Prover proved is correct.

3 Digitalizing Voting

In this section we discuss how voting procedure can be securely digitalized using cryptography. Typically the process of designing cryptographic protocols consists of clarifying the purpose and modeling its feature, then design the protocol, and verify the designed protocol meets the previously set goal.


Electronic Voting

- Redefine/clarify its purpose
- Model its features

Fig. 4 Model of electronic voting

3.1 Requirements for Voting

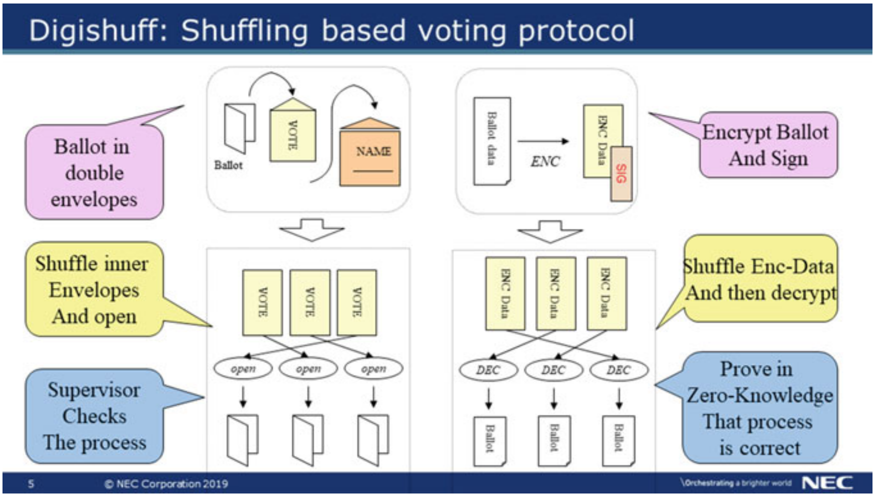
So let us clarify the purpose of the voting and its desired property. Here, we assume there is a list of legitimate voters with their respective public keys and a Tallying authority. Each legitimate voter cast either yes or no vote and the Tally authority wants to have a correct counting of the votes (Fig. 4). The three main requirements we need to meet are the following:

1. Only legitimate voters vote, and one vote per voter.
2. Tallying authority cannot announce faulty results.
3. No one can learn how each voter voted.

3.2 Designing Voting Protocol

It seems these three requirements are hard to achieve simultaneously. If we let all legitimate voters sign their vote, then the first requirement can be met. However, if the votes are signed with the voter's key, it means the votes are not anonymous thus conflicts the third requirement. If we make all votes anonymous, then we cannot verify if the votes are from legitimate voters or even if they are, they could have voted more than once. Moreover, we cannot verify if the Tallying Authority just neglected some of the anonymous votes cast in counting the tally.

There are several ideas to meet all three requirements that seems conflicting. In this subsection, we will discuss one of such ideas using shuffling [2].


Digishuff: Shuffling based voting protocol

Ballot in double envelopes

Shuffle inner Envelopes And open

Supervisor Checks The process

Encrypt Ballot And Sign

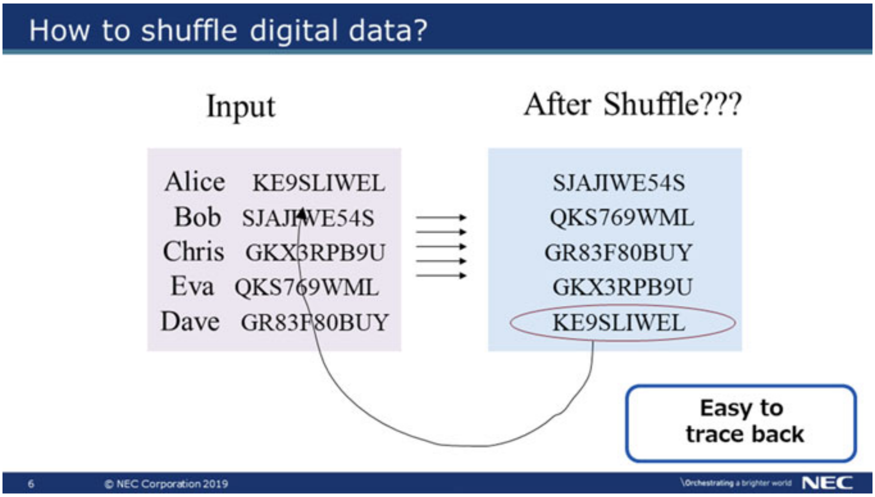
Shuffle Enc-Data And then decrypt

Prove in Zero-Knowledge That process is correct

Fig. 5 Overview of voting protocol using shuffling

The underlying idea came from how we meet those requirements using paper ballots in voting. In one providence, a voter fills in his paper ballot and put in a blank envelope. Then the voter puts this bank envelope in a larger envelope and signs with the voter's name. The voter hands this envelope to the Tallying Authority. The Tallying Authority can verify that the voter is a legitimate voter and has hand in one envelope, but because they are in an envelope the Authority cannot learn the vote. How about counting? On the day of counting the votes, all the outer envelopes are removed, but still in a blank inner envelope. All blank envelopes are thrown on the table and the envelopes will be shuffled manually so that no one learns which inner envelope came from which outer envelope. After adequate shuffling are performed, inner envelopes will be opened and count the ballots within. All the procedure will be supervised by an observer so that Tallying Authority cannot cheat while shuffling or opening the envelopes. So this trick may be able to use in digitalization (Fig. 5).

So we will encrypt the ballot using a public key of the system to mimic the blank inner envelope. As an outer envelope, the voters would sign on the encrypted ballot, and cast to the Tallying Authority. The Authority learns from the signature on the encrypted ballot that the ballot is from a legitimate voter and the same voter had not voted more than once, but the ballot itself cannot be seen as it is encrypted. Then the Authority removes the digital signature part and 'shuffles' the encrypted ballots. After the encrypted ballots has been well mixed, that is, it has been made difficult to match who submitted the encrypted ballot, the ballots will be decrypted to enable tallying. This way, we can ensure that we have only counted legitimate voter's vote once, and authority would not learn the vote of each voter as long as decrypting keys are kept safe. To ensure that the Authority performed correct Tallying, the Authority would provide Zero Knowledge Interactive Proofs to prove that it has



How to shuffle digital data?

Input		After Shuffle???
Alice	KE9SLIWEL	SJAJIWE54S
Bob	SJAJIWE54S	QKS769WML
Chris	GKX3RPB9U	GR83F80BUY
Eva	QKS769WML	GKX3RPB9U
Dave	GR83F80BUY	KE9SLIWEL

Easy to trace back

6 © NEC Corporation 2019 \Orchestrating a brighter world NEC

Fig. 6 Permutation is not shuffling

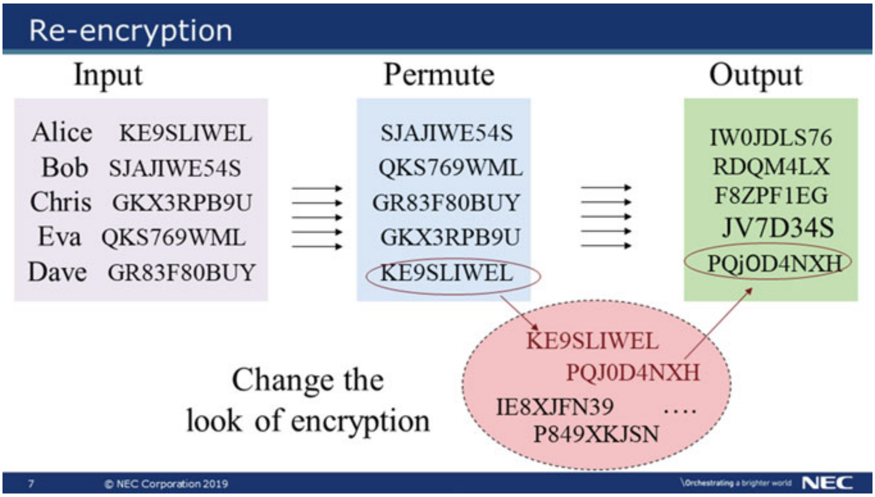
followed the procedure correctly and that the result of the tally is trustworthy. In the next subsection, we discuss in more detail how we ‘shuffle’ digital data.

3.3 Shuffling Encrypted Data Using Probabilistic Encryption

If ‘shuffling digital data’ was simply changing the location of some digital data, then even after shuffling it is easy to spot which digital data came from whom, by matching the bit patterns (Fig. 6).

So in digital shuffling, we need to change a look of digital data. For this purpose, we are going to use a public key encryption scheme that is probabilistic [3]. That is, the encryption function is non-deterministic, therefore there are many ciphertexts that decrypt to a same message. So changing ‘the look’ of encrypted digital data is to replace the encrypted data with another encrypted data that decrypts to the same message. Figure 7 illustrates such shuffling procedure. First a list of encrypted ballots are permutated. Then each encrypted ballot is replaced with another encrypted data without changing the content of the ballot. Looking at the input list and the output list, it is difficult to trace which ballot was shuffled to which position.

An example of a probabilistic encryption scheme that offer this characteristic is called ElGamal Encryption [4]. Here we provide an overview of the scheme. ElGamal Encryption is based on the assumption that given a prime p, an generator g of Zp and $y = g^a \bmod p$, it is difficult to compute a from (p, g, y) for randomly chosen y in Zp. This is called Discrete Logarithm Problem. So KeyGeneration function for ElGamal Encryption is generating p of length k (security parameter) g, and y for



Re-encryption

Input

Alice	KE9SLIWEL
Bob	SJAJIWE54S
Chris	GKX3RPB9U
Eva	QKS769WML
Dave	GR83F80BUY

Permute

SJAJIWE54S
QKS769WML
GR83F80BUY
GKX3RPB9U
KE9SLIWEL

Output

IW0JDLS76
RDQM4LX
F8ZPF1EG
JV7D34S
PQjOD4NXH

Change the look of encryption

KE9SLIWEL
PQJ0D4NXH
IE8XJFN39 ....
P849XKJSN

7 © NEC Corporation 2019 \Orchestrating a brighter world NEC

Fig. 7 Shuffling procedure

randomly chosen a. Public Key will be (p, g, y) and the exponent a will serve as secret key. Encryption function, on input message m in Zp and Public Key (p, g, y), generates a random number r, and outputs

$$(c_1, c_2) = (g^r \bmod p, m * y^r \bmod p)$$

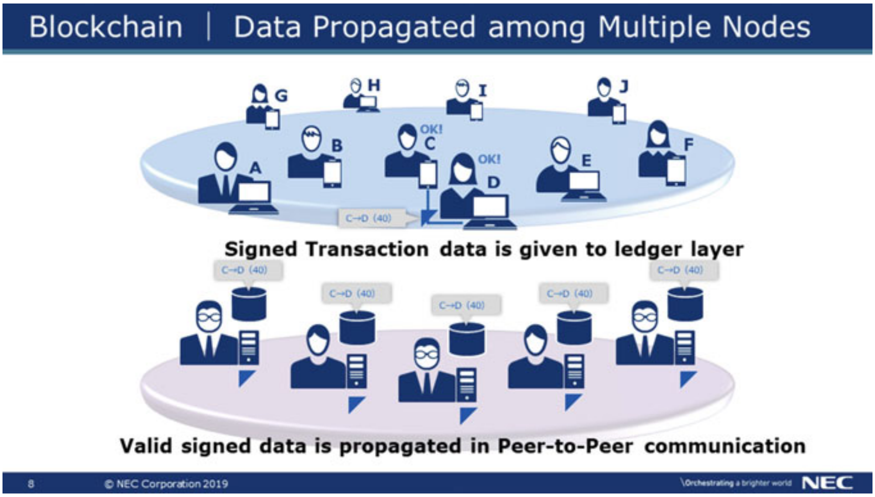
as a ciphertext of m. On input $(c1, c2)$ and secret key a, Decryption function performs $c2/(c1)^a \bmod p$ which should be equal to the message m if the ciphertext was correctly conveyed. In order to change the look of $(c1, c2)$,

$$(d_1, d_2) = (c_1 * g^s \bmod p, c_2 * y^s \bmod p)$$

for a randomly chosen s, would provide another different looking ciphertext that also decrypts to the message m. It is interesting to see that this transformation can be performed without the knowledge of the secret key.

4 Bitcoin Blockchain

Perhaps one of the most impressive digital transformation through cryptography was digitalizing ‘money’ called Bitcoin [5]. There are many prepaid electronic money systems today like PayPay, but it is restricted to one currency and there is an accountable organization who is operating the system. Satoshi Nakamoto designed a system where only the algorithms ensure the correctness of the money transfer and excluding the existence of a centralized authority. We provide below an overview of his design. We note some details are omitted for the sake of simplicity.



Blockchain | Data Propagated among Multiple Nodes

Signed Transaction data is given to ledger layer

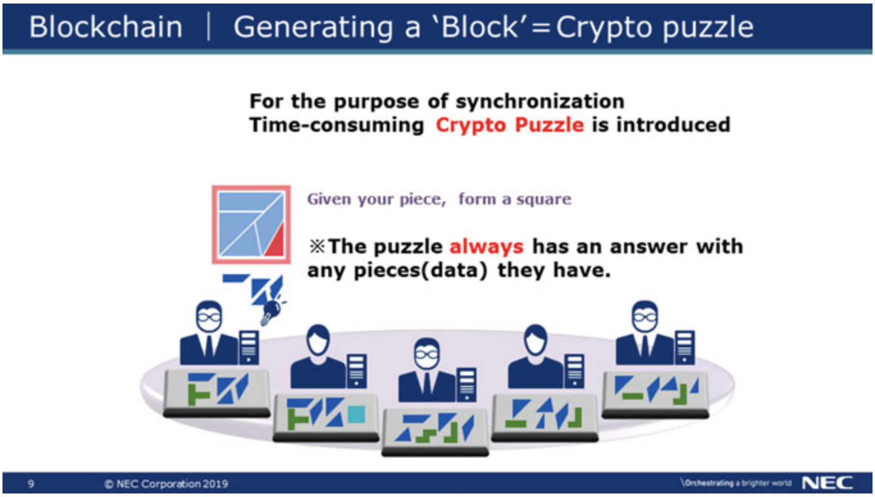
Valid signed data is propagated in Peer-to-Peer communication

8 © NEC Corporation 2019 \Orchestrating a brighter world NEC

Fig. 8 Data managers and transaction logs

4.1 Modeling Blockchain

Blockchain is a technology that is used to manage transaction data in Bitcoin. There are users of Bitcoin who issue transaction data, typically saying ‘sending x Bitcoin from my account yyy to the address zzz.’ The transaction is accepted if the message is indeed sent from the owner of the account yyy and indeed there are x Bitcoin left in the account. The log of transaction infers that after the transaction has been accepted, x Bitcoin should be decreased from the account yyy and added to the account zzz. Unlike previous systems where there is one organization keeping record of all the transactions, there are multiple voluntary ‘Data managers’ in Bitcoin known as Full Node, connected in Peer-to-peer fashion. When a user issued a transaction, Data managers check its correctness and propagates the transaction to other Data managers. The ideal goal is that all the Data managers keep these transaction log in a consistent way (Fig. 8). However, as transaction logs are created by various account holders internet-wide and that communication through Peer-to-peer network may not always be perfect, there is no guarantee that the list of logs are consistent among all the Data managers. So the big problem Satoshi had to solve was how to synchronize the transaction log among the Data managers while they are connected in asynchronous Peer-to-peer network.



Blockchain | Generating a 'Block'=Crypto puzzle

For the purpose of synchronization
Time-consuming Crypto Puzzle is introduced

Given your piece, form a square

※The puzzle always has an answer with any pieces(data) they have.

9 © NEC Corporation 2019 \Orchestrating a brighter world NEC

Fig. 9 Crypto puzzles for synchronization

4.2 Crypto Puzzle for Synchronization

A core idea behind synchronization is to restrict frequent distribution of transactions. If the distribution happens infrequently, for example once in every 10 min or so, that should provide enough time within Peer-to-peer network to share the same data. In order to achieve this, Bitcoin blockchain is designed so that a bulk of transaction log are bundled in a block, and the block cannot be distributed among Data Managers unless accompanied by a certain solved crypto puzzle related to the content of that block. This crypto puzzle is so designed that the puzzle for any block can be solved with high probability, but is time consuming. We note that while the puzzle is hard to solve, it is easy for other Data managers to verify that the solution is correct (Fig. 9).

In order to define crypto puzzle, we use a mathematical function called Hash Function. Hash Function deterministically maps an arbitrarily long input string to a fixed length integer of say 256 bits. The output is called a hashed value. With cryptographically secure hash function, it is computationally difficult to find two different input that maps to a same hashed value. There are known algorithms that is believed to achieve this property, such as SHA-256 [6].

Let us assume a Data Manager wants to add bulk of data $D_1, \ldots, D_n$, on top of the latest Block data Bn. The Puzzle is defined to find an string str that satisfies the following equation.

$$\text{Hash}(\text{Hash}(\text{Bn}) \ \| D_1 \| \ldots \| D_n \| \ \text{str}) < 2^{\text{Bn}(k)}$$

where || represents concatenation of strings and $\mathrm{Bn}(k)$ is an integer defined from the previous block Bn, which is called difficulty. A typical output of Hash function is an integer of length 256, so if $\mathrm{Bn}(k)$ is about 60, one need to try many possible str to check if it meets the equation. The difficulty is so designed that this try and error process would take 10 min on average to find the desired string str.

The list of Data $D_1, \ldots, D_n$, accompanied by the correct puzzle solution str, is the propagated as a new block within Data Managers. Other Data Managers who received the block verifies the correctness of the solution. If correct, they add this block on top of the previous blocks, as the chain of data store. Then they will try to solve the next puzzle based on the new block with other transaction log that has not yet been stored in the blockchain.

4.3 Incentives for Data Managers

We conclude the overview of Bitcon Blockchain by mentioning why the Data managers spend their computational effort to solve meaningless puzzle. The Data managers are awarded by Bitcoin if they solved the puzzle and followed by the future Blocks. Their incentives for receiving the award play a central role in maintaining consistent data among Data managers, and distract them from behaving maliciously.

5 Concluding Remarks

In this article we have discussed some of the examples of securely implementing current social activities in cyber world using cryptography. We have shown some of the cryptographic primitives are defined mathematically. The procedure to design secure protocols begin with clarifying the goal and requirements and then design to meet those criteria. Although these examples show that cryptography is a promising approach, we still lack in technology to model and evaluate mathematically overall system for digital transformation. The author sincerely hope that this article would encourage the researchers in mathematics, cryptography and information technology to get together and share their strengths for the goal of making our digital society more secure and fair place.

References

1. Goldreich, O.: Foundation of Cryptography. Cambridge University Press (2009)
2. Furukawa J, Mori K, Sako K (2010) An implementation of a mix-net based network voting scheme and its use in a private organization. In: Towards Trustworthy Elections, pp. 141–154 (2010)
3. Goldwasser, S., Micali, S.: Probabilistic encryption. J. Comput. Syst. Sci. **28**(2), 270–299 (1984)

4. El Gamal, Taher: A public key cryptosystem and a signature scheme based on discrete logarithms. IEEE Trans. Inform. Theory **31**(4), 469–472 (1985)
5. Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System (2009). https://bitcoin.org/bitcoin.pdf
6. NIST FIPS 180-4: Secure Hash Standard (SHS)

Efficient Algorithms for Tracking Moving Interfaces in Industrial Applications: Inkjet Plotters, Electrojetting, Industrial Foams, and Rotary Bell Painting

Maria Garzon, Robert I. Saye, and James A. Sethian

Abstract Moving interfaces are key components of many dynamic industrial processes, in which complex interface physics determine much of the underlying action and performance. Level set methods, and their descendents, have been valuable in providing robust mathematical formulations and numerical algorithms for tracking the dynamics of these evolving interfaces. In manufacturing applications, these methods have shed light on a variety of industrial processes, including the design of industrial inkjet plotters, the mechanics of electrojetting, shape and evolution in industrial foams, and rotary bell devices in automotive painting. In this review, we discuss some of those applications, illustrating shared algorithmic challenges, and show how to tailor these methods to meet those challenges.

Moving interfaces are key components of many dynamic industrial processes, whose dynamics are critical to the underlying physics. Examples include turbines, flames and combustion, plastic injection molding, microfluids, and pumping. In each of these examples, complex physics at the interface, such as between a fluid and a moving wall, or through a membrane or a transition region, determines much of the underlying action and performance (Fig. 1).

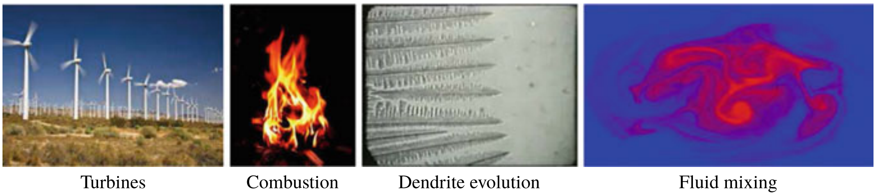
One approach to propagating interfaces is given by "level set methods". These algorithms to track interfaces in multiple dimensions, couple the driving physics with the interface in a natural way, and smoothly handle topological change due to merger and breaking. They accurately and robustly compute high order solutions

M. Garzon
Department of Applied Mathematics, University of Oviedo, Oviedo, Spain
e-mail: maria@uniovi.es

R. I. Saye
Mathematics Group, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
e-mail: rsaye@lbl.gov

J. A. Sethian (✉)
Department of Mathematics, University of California, Berkeley, California 94720, USA
e-mail: sethian@math.berkeley.edu

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_10



Turbines Combustion Dendrite evolution Fluid mixing

Fig. 1 Examples of industrial interfaces

to moving interface problems, and are easily discretized using standard techniques, such as finite difference, finite element, and discontinuous Galerkin methods.

The paper is a review of the application of these methods to some industrial problems, and draws from multiple sources [10–17, 26–29, 42–44] to discuss the design of industrial inkjet plotters, jetting and electrojetting devices, industrial foams, and rotary bell spray devices. Rather than extensively focus on the equations or the algorithms, we provide an overview of the approaches, with an emphasis on the results. References are provided for more in-depth discussions.

1 Modeling Interface Evolution Using Level Set Methods

Level set methods, introduced in [19], have been used in a large number of applications to track moving interfaces. They are based on both a general mathematical theory as well as a robust numerical methodology, which relies on exchanging the typical Lagrangian perspective on front propagation, in which the front is explicitly tracked, for an Eulerian view in which the moving interface is embedded as a particular level set of a higher dimensional function posed in a fixed coordinate system. The motion of the interface corresponds to solving the evolution of this higher-dimensional function according to a Hamilton-Jacobi-type initial value partial differential equation.

A brief summary is as follows. Consider a moving interface $\Gamma(t)$, parameterized by $N-1$ dimensions. We restrict ourselves to interfaces which are closed and simple, and separate the domain into an "inside" and an "outside". We recast the problem by implicitly defining the moving interface $\Gamma(t)$ propagating in $N-1$ dimensions as the zero level set of the solution to an evolving level set function $\phi(x,t)$, $\phi : \mathbb{R}^N \times t \to \mathbb{R}$, which satisfies a time-dependent partial differential equation. There are many ways to initialize this implicit function: one approach is to let $\phi(x, t = 0)$ be the signed distance from the interface $\Gamma(t = 0)$, linking the interface to the zero level set.

We assume that the underlying physics specifies a speed F normal to the interface at every point on the interface. Constructing this speed function typically involves solving complex physics both on and off the interface.

Thus, there are two embeddings. First, the interface itself is embedded and implicitly defined through a higher-dimensional function ϕ. Second, to move the other level sets, we embed the speed F in a higher-dimensional function, known in the literature as the "extension velocity" F_{ext}, which defaults to the given speed on the zero level set corresponding to the interface.

1.1 Equations of Motion

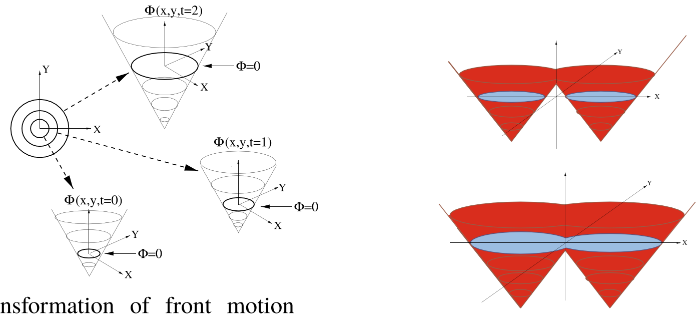
Here, we review the basic ideas behind the derivation and implementations of level set methods. We follow the derivation and discussion in [35, 36].

We wish to produce an Eulerian formulation for the motion of a hypersurface Γ representing the interface and propagating along its normal direction with speed F, where F can be a function of various arguments. Let $\pm d(x)$ be the signed distance from the point $x \in \mathbb{R}^N$ to the interface at time $t = 0$. Define a function $\phi(x, t = 0)$ by the equation

$$\phi(x, t = 0) = \pm d(x). \tag{1}$$

By requiring that the zero level set of the evolving ϕ (see Fig. 2, left) always match the propagating hypersurface, means that

$$\phi(x(t), t) = 0. \tag{2}$$



Transformation of front motion into initial value problem. An implicitly defined surface ϕ, whose ensuing motion satisfies Equation 3, and whose zero level set always matches the motion of the interface.

The level surface ϕ in red. Top: $\phi = 0$ corresponds to two separate initial fronts. Bottom: Later in time: the interface topology has changed, yielding a single curve as the zero level set.

Fig. 2 Left: Implicit embedding of level set function. Right: Topological change

By the chain rule, $\phi_t + \nabla\phi(x(t), t) \cdot x'(t) = 0$, Since $x'(t) \cdot \mathbf{n} = F_{ext}$, where $\mathbf{n} = \nabla\phi/|\nabla\phi|$ with extension velocity F_{ext}, this yields an evolution equation for ϕ, namely,

$$\phi_t + F_{ext}|\nabla\phi| = 0, \quad \text{given} \quad \phi(x, t = 0). \tag{3}$$

This is the level set equation introduced by Osher and Sethian [19]. Propagating fronts can develop shocks and rarefactions in the gradient, corresponding to corners and fans in the evolving interface, and numerical techniques designed for hyperbolic conservation laws can be exploited to construct schemes which produce the correct, physically reasonable entropy solution, see [32–34].

There are several advantages to this approach. First, the formulation works in any number of dimensions. Second, topological changes are handled without special attention: fronts split and merge. Third, geometric quantities along the interface can be calculated by taking advantage of the embedding and computing quantities in the fixed Eulerian setting. Fourth, this formulation naturally lends itself to numerical approximations, for example, through finite difference or finite element formulations on the fixed background mesh.

1.2 Computational Advances

Since its introduction, a large number of computational advances have been developed to make this approach efficient, accurate, and economical. These include

- The introduction of adaptive, “narrow band level set methods” [1] which confine computation to a thin band around the zero level set.
- Fast methods to construct extension velocities [2, 25].
- Incorporation of complex physics [3–5], transport of material quantities [6], and methods to handle multi-phase flows with a large number of distinct propagating regions coming together in complex junctions, triple points, etc. [26, 27].

A large number of reviews have been appeared over the years, containing these and many related ideas. We refer the interested reader to [20, 30, 35, 36, 38–40].

2 Industrial Printing

2.1 Physical Problem and Modeling Goals

Industrial inkjet printing involves ejecting ink housed in a well through a narrow nozzle, which is then deposited on a material. The ink in the bath is expelled by an electro-actuator mechanism at the bottom, which quickly propels ink through the nozzle. The shape of the nozzle, the force and timing of the actuator, and the properties of the ink are instrumental in determining the ultimate shape, delivery, and performance of the printing device.

This is a two-phase incompressible fluid flow problem, with the interface separating air and ink. Depending on the constituency of the ink, the flow can either be Newtonian or visco-elastic. Boundary conditions include both no-slip and no-flow at solid walls, and triple points where air-ink boundaries meets solid nozzle walls are subject to typical critical angle dynamics controlling slipping. While a common use for inkjet printers is in commercial home printing, over the past two decades a large number of sophisticated industrial applications have appeared, ranging from printing integrated circuits and the manufacture of display devices on through to construction of tissue scaffolding and layered manufacturing.

The goal of numerical simulation is to identify and optimize key aspects of the process, including

- Optimize the design of the nozzle and to control the actuator mechanism to aim, extend, and focus droplet delivery;
- Characterize wall wetting/non-wetting on the shape and separation of droplets;
- Determine and perhaps minimize the formation of secondary trailing droplets, which break off from the main ejected bubble as the fluid elongates, due to the effects of surface tension; and
- Understand how variations in viscosities and impurities affect droplet dynamics.

2.2 Equations of Motion and Computational Challenges

We solve for incompressible flow in a non-rectangular geometry, with no-slip and no-flow on walls, with air satisfying Newtonian flow and ink satisfying a visco-elastic Oldroyd-B model. The equations of motion [42–44], are given by

$$\begin{aligned} &(Ink)\quad \rho_1 \frac{D\boldsymbol{u}_1}{Dt} = -\nabla p_1 + \nabla \cdot (2\mu_1 \mathcal{D}_1) + \nabla \cdot \boldsymbol{\tau}_1 \ , \qquad \nabla \cdot \boldsymbol{u}_1 = 0 \ , \\ &\frac{D\boldsymbol{\tau}_1}{Dt} = \boldsymbol{\tau}_1 \cdot (\nabla \boldsymbol{u}_1) + (\nabla \boldsymbol{u}_1)^T \cdot \boldsymbol{\tau}_1 - \frac{1}{\lambda_1}\left(\boldsymbol{\tau}_1 - 2\mu_{p1}\mathcal{D}_1\right) \ . \end{aligned} \tag{4}$$

$$(Air)\quad \rho_2 \frac{D\boldsymbol{u}_2}{Dt} = -\nabla p_2 + \nabla \cdot (2\mu_2 \mathcal{D}_2) \ , \qquad \nabla \cdot \boldsymbol{u}_2 = 0 \ . \tag{5}$$

$$\mathcal{D}_i = \frac{1}{2}\left[\nabla \boldsymbol{u}_i + (\nabla \boldsymbol{u}_i)^T\right], \quad \boldsymbol{u}_i = u_i \boldsymbol{e}_r + v_i \boldsymbol{e}_z \ , \qquad i = 1, 2 \tag{6}$$

where, for the ink, $\boldsymbol{\tau}_1$ is the viscoelastic stress tensor, λ_1 is the viscoelastic relaxation time, μ_{p1} is the solute dynamic viscosity and subscript 2 refers to (Newtonian) air.

We use a level set method to track the air-ink interface, starting with the initial pressure disturbance in the reservoir: the fluid then moves through the nozzle and is then ejected into the ambient air, and then may separate into one or more droplets. We compute an approximate solution to the incompressible Navier-Stokes given above

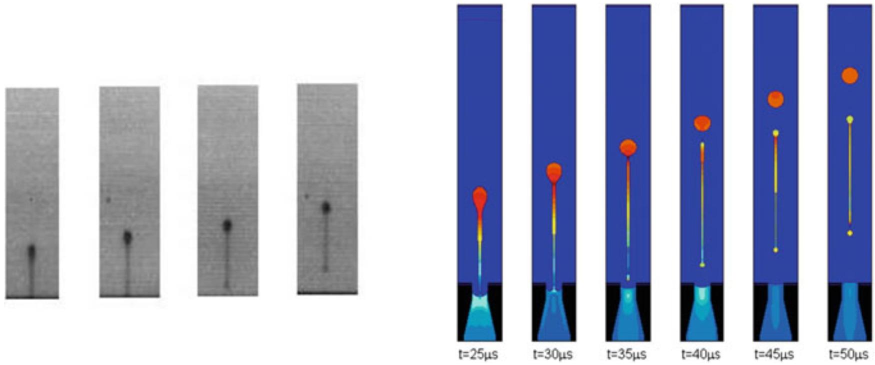


Fig. 3 Left: Experimental profiles, showing ejected ink and satellite formation; note the formation of the trailing satellite droplet as the initial bubble stretches, and changes topology. Right: simulation of full ejection cycle (taken from [43]). Inflow pressure from an equivalent circuit model which describes the cartridge, supply channel, vibration plate, PZT actuator, and applied voltage. Fluid is an Epson dye-based ink, with critical advancing $\theta_a = 70°$ and receding $\theta_r = 30°$ contact angle, and with $\rho_1 = 1070\,\text{kg/m}^3$, $\mu_1 = 3.34 \times 10^{-3}$ kg/m s, and $\sigma = 0.032\text{kg/s}^2$. The nozzle geometry has diameter 26 microns at opening and 65μm at bottom

in both phases simultaneously, with surface tension terms mollified to the right-hand-side as a forcing term. Thus, the solution accounts for both the ink velocity, the air-ink interface, and air currents induced in the air by the fluid ejection. We use a second order projection method [7–9] on a body-fitted logically rectangular mesh. Calculations are performed in both axi-symmetric two dimensions and full three-dimensional regimes. For details, see [42–44]. Figure 3 shows the results of both an experiment and simulation.

3 Droplet Formation and Electro-jetting

3.1 Physical Problem and Modeling Goals

A large number of industrial problems involve microjetting and droplet dynamics, in which small droplets both move through small structures and also transport key materials, for example, in such areas as deposition of evaporation substances, delivery of biological materials, and substance separation.

Part of the challenge in computing these problems stems from the critical role of surface tension and shear forces, which often drive topological change, breakage, and merger in the evolving droplets. Level set methods, because of their ability to handle these structural changes, are particularly well-suited for computing droplet dynamics. Here, we summarize work on microjetting dynamics first presented in [10], see also [11–17].

Consider the dynamics of a thin tube of fluid as it pinches off due to surface tensions effects at a narrowing neck of the fluid (see Fig. 5), where mean curvature drives the interface inward until it breaks into two separate lobes of fluid. The pinch-off dynamics reveal considerable intricacy: as the droplet breaks, rapidly moving capillary waves on the surface cause instabilities and oscillations in the fluid lobes.

3.2 Equations of Motion

Following the arguments in [11, 12], we model the fluid as incompressible and irrotational with a potential flow formulation. Euler's equation gives

$$\nabla \cdot \mathbf{u} = 0 \ \text{ in } \ \Omega(t) \tag{7}$$

$$\mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} = \frac{-\nabla p}{\rho} + \mathbf{bodyforces} \ \text{ on } \ \Gamma_t(\mathbf{s}). \tag{8}$$

Assuming irrotationality ($\nabla \times \mathbf{u} = 0$), the problem can then be written in terms of a fluid velocity potential $\mathbf{u} = \nabla\psi$, namely

$$\Delta\psi \quad = 0 \text{ in } \Omega(t) \tag{9}$$

$$\psi_t + \frac{1}{2}(\nabla\psi \cdot \nabla\psi) + \frac{p - p_a}{\rho} = 0 \ \text{ on } \Gamma_t(\mathbf{s}), \tag{10}$$

where p_a is the atmospheric pressure and ρ is the fluid density.

As shown in [11, 12], this can be reformulated as

$$\mathbf{u} = \nabla\psi \ \text{ in } \ \Omega(t), \quad \Delta\psi = 0 \ \text{ in } \ \Omega(t) \tag{11}$$

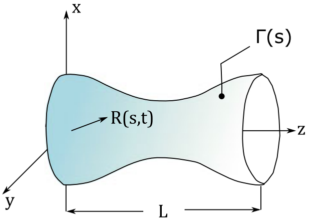
$$\frac{D\psi}{Dt} = \frac{1}{2}\big(\nabla\psi \cdot \nabla\psi\big) - \frac{\gamma}{\rho}\big(\frac{1}{R_1} + \frac{1}{R_2}\big) \ \text{ on } \ \Gamma_t(\mathbf{s}) \ , \tag{12}$$

where $\Omega(t)$ is the fluid tube, $\Gamma_t(s)$ is the boundary of the tube, R_1 and R_2 are the principle radii of curvature, and γ is the surface tension.

Although the potential ψ is only defined on the interface, our plan is to build an extension of both the potential and the interface to all of space, so that we can then employ the level set methodology. This embedded implicit formulation then allows calculation of the fluid interface motion through pinch off, and can compute dynamics of the split fluid lobes.

These embeddings produce a new set of equations, namely

$$
\begin{aligned}
&\mathbf{u} = \nabla\psi \ \text{ in } \ \Omega(t)\\
&\Delta\psi(r,z) = 0 \ \text{ in } \ \Omega(t)\\
&\phi_t + \mathbf{u}_{\text{ext}}\cdot\nabla\phi = 0 \ \text{ in } \ \Omega_D\\
&G_t + \mathbf{u}_{\text{ext}}\cdot\nabla G = f_{\text{ext}} \ \text{ in } \ \Omega_D
\end{aligned}
$$



For details about the derivation of these equations, see [10–12].

3.3 Computational Challenges

The computational challenges that stem from these equations of motion lie in part on the delicate, sharp singularity at pinch off. The curvature becomes very large, and as soon as pinch off occurs, the two pieces of the neck retract very quickly. Constructing correct extension values for the velocity and the potential requires care as well.

We solve these equations through a time-cycle. Given values for the embedded implicit potential and level set function on a fixed background mesh, we construct the zero level set corresponding to the interface, place boundary element nodes on that interface, and then employ a boundary element method to find the new potential and associated velocity field, suitably extended. These nodes are then discarded, and the discrete grid values for the level set function, potential, and velocity are updated.

3.4 Example Results

Extensive numerical experiments are given in [12, 14]: the self-similar behavior of some variables near pinch-off time is checked within the computations and the computed scaling exponents agree with experimental and theoretical reported values. Here we review those results. Figure 4 shows a snapshot after pinch-off, revealing capillary surface waves on the undulating surface. Figure 5 shows the fine-scale structure of droplet dynamics after pinch-off.

3.5 Charged Droplet Separation

The above situation becomes considerably more complicated when the droplets are electrically charged, in which the droplet motion is driven by a background electrical field. Applications include electrospray ionization, electrospinning to produce fibers by drawing charged threads of polymers, particle deposition for nanostructures, drug delivery systems, and electrostatic rotary bell painting.

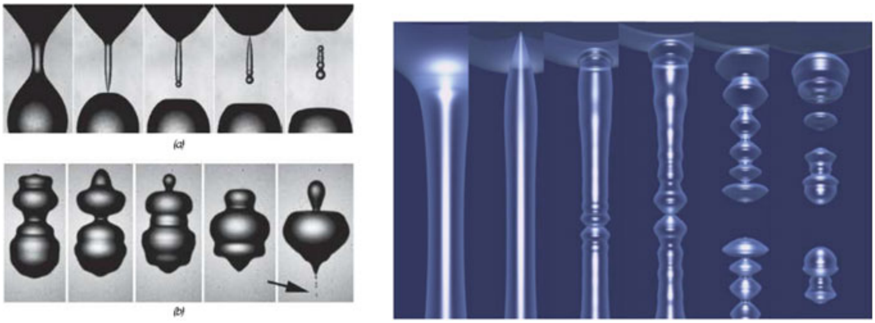


Fig. 4 Droplet dynamics. Left, experiment taken from [41]. Right, level set calculation of surface capillary waves, taken from [12]

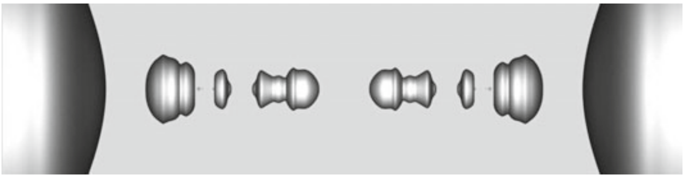


Fig. 5 Simulation: fine-scale structure of droplet dynamics after pinch-off [12]

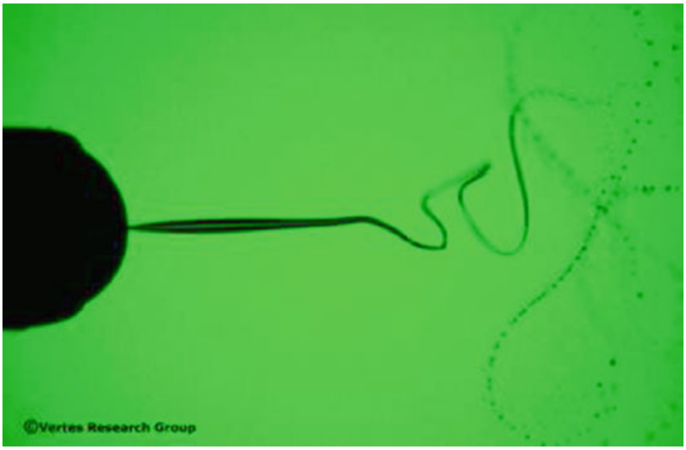
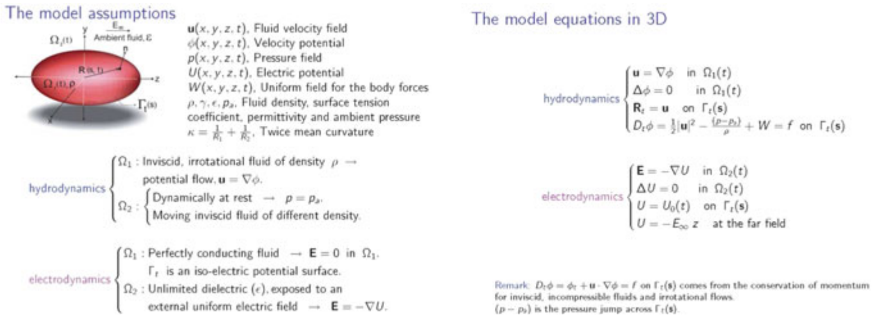


Fig. 6 Experimental profile of electrically charged droplet motion [18]

The fundamental mechanism relies on the motion of an electrically conductive liquid in an electric field. The shape of the droplets starts to deform under the action of the electric field, afterwards the competition between inertial, surface tension and electric forces drives the dynamics, see Fig. 6.



The model assumptions

$\mathbf{u}(x, y, z, t)$, Fluid velocity field
$\phi(x, y, z, t)$, Velocity potential
$p(x, y, z, t)$, Pressure field
$U(x, y, z, t)$, Electric potential
$W(x, y, z, t)$, Uniform field for the body forces
$\rho, \gamma, \epsilon, p_a$, Fluid density, surface tension coefficient, permittivity and ambient pressure
$\kappa = \frac{1}{R_1} + \frac{1}{R_2}$, Twice mean curvature

hydrodynamics $\begin{cases} \Omega_1 : \text{Inviscid, irrotational fluid of density } \rho \rightarrow \\ \quad \text{potential flow, } \mathbf{u} = \nabla\phi. \\ \Omega_2 : \begin{cases} \text{Dynamically at rest} \rightarrow p = p_a. \\ \text{Moving inviscid fluid of different density.} \end{cases} \end{cases}$

electrodynamics $\begin{cases} \Omega_1 : \text{Perfectly conducting fluid} \rightarrow \mathbf{E} = 0 \text{ in } \Omega_1. \\ \quad \Gamma_t \text{ is an iso-electric potential surface.} \\ \Omega_2 : \text{Unlimited dielectric } (\epsilon)\text{, exposed to an} \\ \quad \text{external uniform electric field} \rightarrow \mathbf{E} = -\nabla U. \end{cases}$

The model equations in 3D

hydrodynamics $\begin{cases} \mathbf{u} = \nabla\phi \quad \text{in } \Omega_1(t) \\ \Delta\phi = 0 \quad \text{in } \Omega_1(t) \\ \mathbf{R}_t = \mathbf{u} \quad \text{on } \Gamma_t(\mathbf{s}) \\ D_t\phi = \frac{1}{2}|\mathbf{u}|^2 - \frac{(p-p_a)}{\rho} + W = f \quad \text{on } \Gamma_t(\mathbf{s}) \end{cases}$

electrodynamics $\begin{cases} \mathbf{E} = -\nabla U \quad \text{in } \Omega_2(t) \\ \Delta U = 0 \quad \text{in } \Omega_2(t) \\ U = U_0(t) \quad \text{on } \Gamma_t(\mathbf{s}) \\ U = -E_\infty z \quad \text{at the far field} \end{cases}$

Remark: $D_t\phi = \phi_t + \mathbf{u} \cdot \nabla\phi = f$ on $\Gamma_t(\mathbf{s})$ comes from the conservation of momentum for inviscid, incompressible fluids and irrotational flows.
$(p - p_a)$ is the pressure jump across $\Gamma_t(\mathbf{s})$.

Fig. 7 Equations for electrically charged droplet motion. Note: In the shown equations, the velocity potential is labelled Φ but is labelled by Ψ in the main text

3.6 Equations of Motion and Computational Challenges

The equations of motion are the previous potential formulation for droplet hydrodynamic motion, plus electrodynamics. We assume a perfectly conducting fluid and an unlimited dieletric exposed to an external uniform force field. Model equations from [16] are shown in Fig. 7.

Algorithmic challenges include accurate and reliable computation of the electric field and handling sharp breakup and fast ejection.

3.7 Example Results

We show a numerical simulation [16] of a free charged droplet carrying a charge above the critical one, reproducing experimental results before and after jet emission. Figure 8 shows the focused droplet end from which charged tiny droplets are ejected.

4 Industrial Foams

4.1 Physical Problem and Modeling Goals

Many problems involve the interaction of multiply-connected regions moving together. These include the mechanics and architecture of liquid foams, such as polyurethane and colloidal mixtures, and of solid foams, such as wood and bone.

The industrial applications of these problems are manifold. Liquid foams are key ingredients in industrial manufacturing, used in fire retardants and in froth flotation for separating substances. Solidification of liquid foams results in solid foams, which have remarkably strong compressible strength because of their pore-like internal structure; and include lightweight bicycle helmets and automotive absorbers.

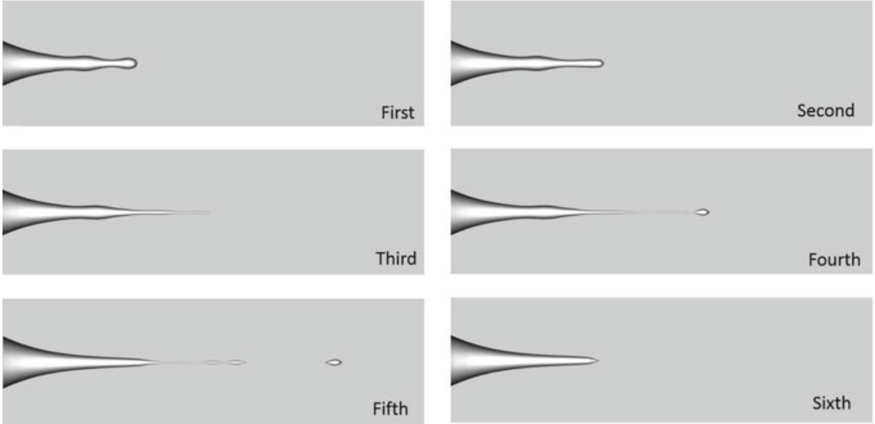


Fig. 8 Time evolution of electrically charged droplet motion, from [16]

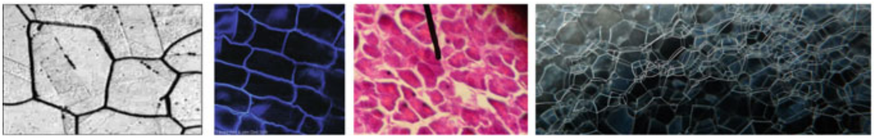

Grain metal boundaries Root structures Acinar cells Foamy fluids

Fig. 9 Examples of multiphase problems

In such problems, multiple domains share walls meeting at multiple junctions. Boundaries move under forces which depend on both local and global geometric properties, such as surface tension and volume constraints, as well long-range physical forces, including incompressible flow, membrane permeability, and elasticity.

Foam modeling is made challenging by the vast range of space and time scales involved [6]. Consider an open, half-empty bottle of beer. It may seem that nothing is happening in the collection of interconnected bubbles near the top, but currents in the lamellae separating the air pockets show slow but steady drainage. It can take tens to hundreds of seconds for the lamellae fluid to drain and then rupture, triggering an lamella explosion that retracts at hundreds of centimeters a second, after which the imbalanced configuration rights itself to a new stable structure in less than a second. Spatially, membranes are barely micrometers thick, while large gas pockets can span many millimeters or centimeters. All told, the biggest and smallest scales differ by roughly six orders of magnitude in space and time.

Another example comes from grain metal coarsening, in which surface energy, often associated with temperature changes, drives a system to larger structures. A third example comes from foam-foamed fiber networks, found in both industrial materials such as paper and biological materials, such as plant cells and tissues (Fig. 9).

In all of these engineering problems, understanding how such factors as pocket formation and distribution, tensile strengths, and foam architecture is a key part of producing mechanisms to optimize foam performance.

4.2 Computational Challenges

Producing good mathematical models and numerical algorithms that capture the motion of these interfaces is challenging, especially at junctions where multiple interfaces meet, and when topological connections change. Methods have been proposed, including front tracking, volume of fluid, variational, and level set methods. It has remained a challenge to robustly and accurately handle the wide range of possible motions of an evolving, highly complex, multiply-connected interface separating a large number of phases under time-resolved physics.

The problem is exacerbated by the nature of the mathematical components that contribute to the dynamics, including: velocities dependent on such factors as curvature, normal directions and anisotropy; the solution of complex PDEs with jump conditions, source terms, and prescribed values at the interface and internal boundary conditions; area and volume-dependent integrals over phases; thermal effects and diffusion within phases; and balance of forces at complex junctions.

From a numerical perspective, some of the challenges stem from the vast time and space scales involved. Using the same spatial resolution to resolve the physics along interfaces is often impractical in the bulk phases. Sharp resolution of the interface and front-driven physical quantities located on the interface is required as input to the bulk PDEs. Accurately resolving interface junctures is critical in order to provide reliable values for the balances of forces at junctions.

All told, these lead to formidable numerical modeling challenges.

4.3 Voronoi Implicit Interface Methods

Voronoi Implicit Interfaces Methods (VIIM), introduced in [26], provide an accurate, robust, and reliable way to track multiphase physics and problems with a large number of collected, interacting phases. They work in any number of space dimensions, represent the complete phase structure by a single function value plus indicator at each discretized element of the computational domain, couple easily to complex physics, and handle topological change, merger, breakage, and phase extinction in a natural manner. The underlying equations of motion that represent the evolving interface and complex physics may be approximated in either a finite difference or finite element framework. These equations couple level set methods for an evolving initial value Hamilton-Jacobi-type partial differential equation to a computational geometry-based Eikonal equation to produce a faithful phase representation. Here, we provide a brief review of the methods. For details, see [26].

The starting point is to consider a collection of non-overlapping phases which divide up the domain. The "interface" consists of places where these phases meet. In two dimensions, the simplest example is a single curve separating two phases. More complex structures might have multiple closed curves, each surrounding a separate phase, which meet in triple points or higher-order junctions. In three dimensions, the situation is far more complex.

The Voronoi Implicit Interface Method begins by characterizing the entire system through an implicit representation. For each point x in the plane, define $\phi(x)$ as the distance to the closest interface. Additionally, define $\chi(x)$ as an integer-valued function which indicates the phase. By construction, the interface representing all possible boundaries is given as the zero level set $\{\phi(x) = 0\}$ of this unsigned distance function, and the indicator function reveals the type of phase.

Thus, for example, if $\phi(x) = 5$ and $\chi(x) = 4$, then we know that the point x is located in phase 4, and the closest interface point is located a distance 5 away.

Starting with this unsigned distance function representation, we execute a two-step process. With interface speed F in the normal direction:

- Advance ϕ through k time steps using the standard level set methodology. That is, produce ϕ^{n+1} from ϕ^n by solving a discrete approximation to

$$\phi_t + F|\nabla\phi| = 0.$$

- Use the ϵ level sets of this time-advanced solution to reconstruct a new unsigned distance function. This is done by first computing the Voronoi interface from the ϵ level sets: this corresponds to the set of all points equidistant from at least two of the ϵ level sets from different phases, and closer to any of the non-equidistant phases. This Voronoi interface is then used to rebuild the unsigned distance function.

These two steps give the method its name: "Implicit Interface" because of the level set step for the time evolution, and "Voronoi" because of the reconstruction step used to rebuild the unsigned distance function and characteristic indicator function.

There are several things to note:

- The method works because of a comparison principle which, for a large fraction of physically reasonable flows built through the use of extension velocities (see [2]), keeps the zero level set trapped between the neighboring ϵ level sets. These ϵ level sets may be updated for a short period of time without suffering from the influence of the non-smooth ridge along the zero level set.
- The Voronoi reconstruction can be accomplished without explicit construction through two applications of fast Eikonal solvers [25, 37].
- Regions spontaneously disappear (appear) if they become small (large) enough so that an ϵ-level set does not exist (can be constructed).
- Careful numerical algorithms can be devised to allow for any non-negative value for ϵ, including $\epsilon = 0^+$.

For details, see [26, 27].

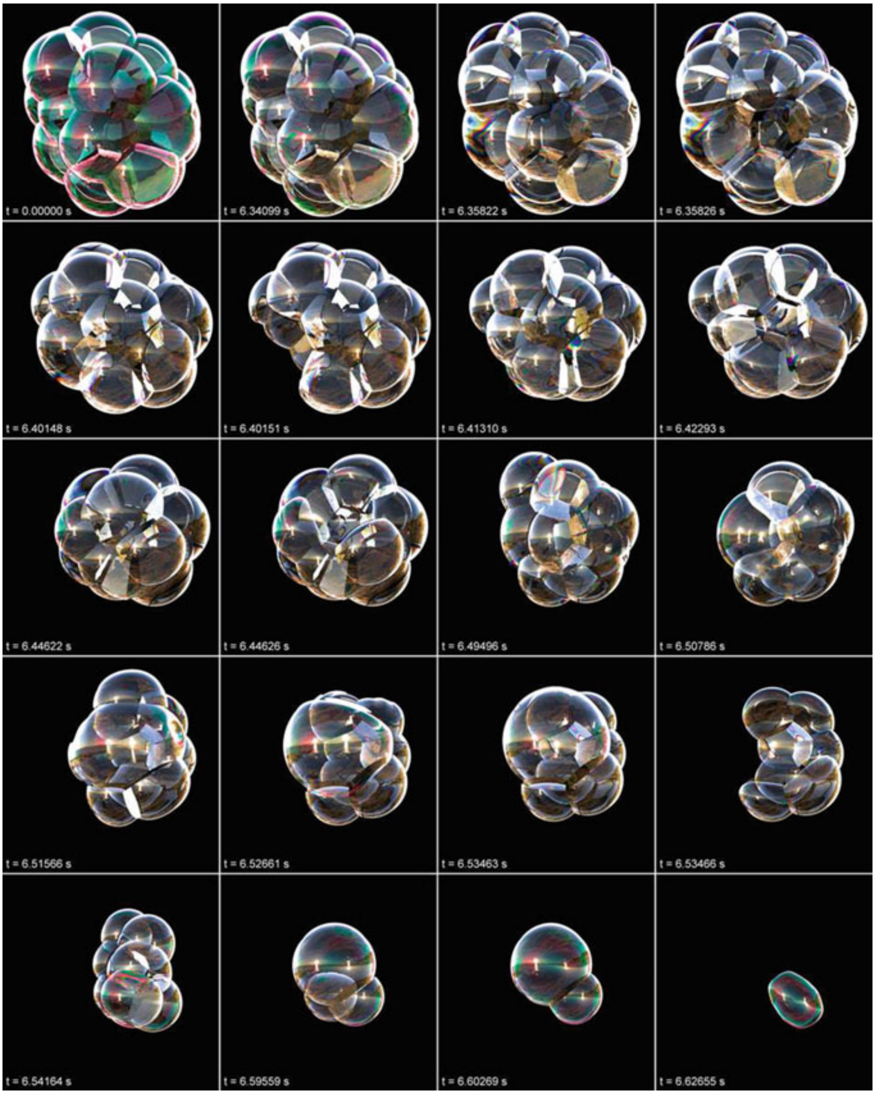


Fig. 10 Collapse of a foam cluster, visualized with thin-film interference taken from [29]

4.4 Application of VIIM to Foam Dynamics

Here, we review some current work applying VIIM to tracking the evolution of liquid foams. The vast time and space scales mean that one cannot compute over all scales simultaneously. Instead, we use a scale-separation model which allows us to divide the foam physics into three distinct stages.

We characterize the foam structures as represented by thin, interconnected membranes (lamellae) each surrounding pockets of air, and containing fluid. Membranes can share common walls, and fluid in each lamella drains toward common, shared Plateau borders that form a network of triple junctions and quadruple points. This drainage is slow, and once a membrane becomes too thin, it ruptures, causing the large air pockets to be out of macroscopic balance, which then readjust according to the equations of incompressible flow driven by interfacial forces along the lamellae.

These events can be thought of as taking places over different scales. The macroscopic air-fluid incompressible flow phase takes place over the whole domain, and evolves to an equilibrium relatively quickly. The lamellae drainage phase is slow, but takes place only over the very thin membrane walls. Rupture occurs very quickly.

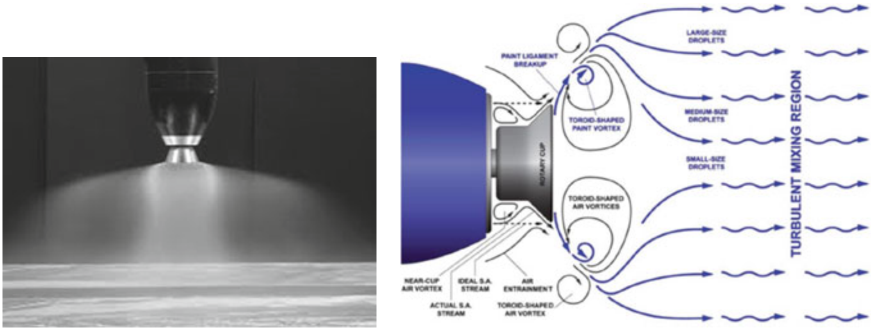
In [28, 29], these three phases were used to develop a mathematical model and numerical simulation framework for foam evolution. During the macroscopic phase, a second order projection method is used to solve the incompressible Navier-Stokes equations on a rectangular mesh, with the interface smoothing its influence to the right-hand side through a mollified surface tension term. The individual lamellae are advanced under the incompressible flow by the Voronoi Implicit Interface Method, with the internal liquid transported by the method of characteristics. When the motion is almost gone, the model enters a different phase and assumes that the multi-phase configuration has essentially reached equilibrium; a fourth order PDE is then solved for thin film drainage, approximated through a discretized finite element triangulation. The final phase results from membrane rupture, idealized as an instantaneous disappearance of a lamella when a user-chosen minimal thickness is reached, which then redistributes the lamella liquid mass and sends the configuration into macroscopic disequilibrium.

4.5 *Example Results*

An example of the complete dynamics developed in the multi-scale foam model is shown in Fig. 10, which shows the time evolution of a bubble cluster, starting from 26 separate bubbles and ending up in a single bubble. The bubble colors are computed from thin film interference determined by the computed fluid thickness in the lamellae.

5 Rotary Bell Painting in the Automotive Industry

In manufacturing settings, paints are frequently applied by an electrostatic rotary bell atomizer. Paint flows to a cup rotating at 10,000–70,000 rpm and is driven by centrifugal forces to form thin sheets and tendrils at the cup edge, where it then tears apart into dispersed droplets. Vortical structures generated by shaping air currents are key to shearing these sheets and transporting paint droplets. Advantages of this manufacturing process include the ability to paint at high volume and to achieve uniform consistency in the paint application.



Schematic of paint flow and air currents [21] in rotary bell atomizing applications

Understanding the generation, size distribution, delivery, and adhesion of these paint droplets is a problem of considerable importance. For example, (a) much of the energy involved in automotive assembly is associated with the paint process; (b) a significant amount of paint does not attach to the cars and ends up as pollutants; and (c) 10–20% of automobiles need to be repainted due to aberrations in the process.

The goal of computational modeling of the rotary bell delivery system includes

- Optimizing the atomization process for higher paint flow rates to obtain more uniform and consistent atomization in the 30,000 to 60,000 rpm range.
- Studying the atomization process as a function of paint fluid properties (such as density, viscosity, and surface tension) and physical properties, such as inflow rates, bell rotation speeds and shaping air currents.
- Analyzing film dynamics, particularly in the immediate atomization zone adjacent to the cup edge, including the dynamics of filament formation and droplet size and distribution and their trajectories.

5.1 Computational Challenges

The computational challenges posed by the painting delivery mechanism are formidable. The range of physical parameters is substantial. The droplet size ranges from 5 to 100 μm, the films are 10–50 μm thick, while the rotary bell diameter is on the order of centimeters. The cup rotates at 200 m/s, droplets breakup over microseconds, whereas droplet statistics requires milliseconds. As such, modeling requires tracking droplets across a wide range of length scales, paint fluid mechanics is subject to high centrifugal and Coriolis forces, and the impact of highly vortical air structures on film sheeting requires careful resolution.

From a computational point of view, these translate into daunting challenges:

- Interfaces are very contorted and complex.
- Very thin sheets of paint roll off, and then break into droplets.
- Fluid dynamics is highly three-dimensional with gas eddies playing a key role.
- Droplets are tiny, and break off and subsequently merge in highly complex ways.

- Mass conservation is important: tracking and accurately accounting for small droplets is critical, since all the paint ultimately breaks into such small objects.

These translate into several modeling/mathematical/algorithmic/numerical challenges which must be tackled in order to build a workable approach, including:

- *High-order accurate fluid solvers and sharp interface physics:* The standard level set approach to tracking two- or multi-phase fluid problems is to solve both the evolving level set equation and the Navier-Stokes equations on a background fixed mesh, smearing forces jump conditions, and discontinuities across the air/fluid interfaces through mollified delta functions into forcing terms on the background mesh. Because the droplets are so small, and because the viscosity/density jumps are so large, this approach is too inaccurate. Instead, we need to employ incompressible Navier-Stokes solvers that allow us to represent these forces sharply, by using implicitly defined meshes that adapt to the moving geometry of the liquid-gas interface.
- *Develop hybrid interface solvers coupled to high order fluid solvers.* Coupling these high-order fluid solvers to the interface dynamics requires building accurate methods to allow information transfer between the background Cartesian level set mesh and the unstructured interface-fitted mesh.
- *Non-Newtonian fluids:* Another complex challenge stems from the fact that paint is in fact non-Newtonian. One must carefully design and embed experimental shear stress models inside numerical calculations.
- *Mesh adaptivity:* In order to capture the shaping air currents and spinning bell, which occupy large length scales, as well as the smallest scales of droplets and thin films, we need to employ aggressive adaptive mesh refinement strategies.
- *Multi-core high performance computing:*This is an involved calculation, requiring small time steps, many mesh elements, and highly accurate elliptic solvers. Attention must be paid to parallel implementations on sophisticated computing architectures.

5.2 Level Set Methods and High-Order Multiphase Flow

The central problem in applying level set methods is that the equations of motion need to include jump conditions at the air-paint interface, e.g., droplet boundaries. The usual level set approach of “smearing” forces to a background mesh in order to provide source terms to the incompressible Navier-Stokes equations is problematic. The droplets can be so small, and the density/viscosity jumps so large and sharp, that this mollified approach does not provide the required accuracy.

Instead, we make use of an algorithmic technology building on implicitly-defined meshes [22–24]. There are several ideas at work in this approach:

- First, two-phase incompressible flow is solved using a discontinuous Galerkin (DG) approach, with a level set method used to track paint-air interfaces.

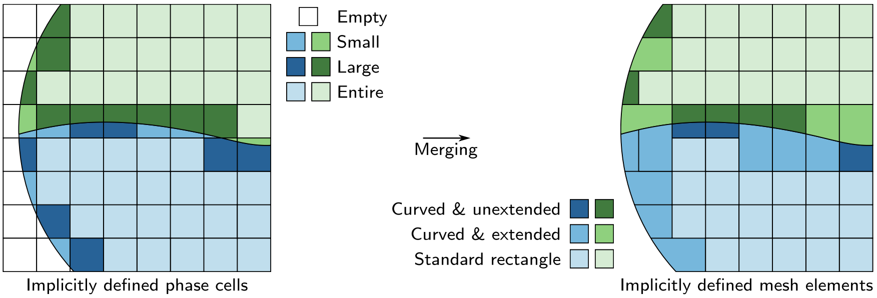


Fig. 11 Implicitly defined meshes using multi-phase cell merging. Left: Phase cells, defined by the intersection of each phase (blue and green) with the cells of a background Cartesian/quadtree grid, are classified according to whether they fall entirely within one phase, entirely outside the domain, or according to whether they have a small or large volume fraction. Right: Small cells are merged with neighboring cells in the same phase to form a finite element mesh composed of standard rectangular elements and elements with curved, implicitly defined boundaries. Figures adapted from [23, 24]

- The level set method is solved using finite differences on a fixed background mesh in a time-evolving narrow-banded data structure.
- The zero level set corresponding to the paint-air boundaries, which cuts through the cells of a background octree grid, is used to drive a cell-merging procedure which creates an implicitly-defined mesh, whose element shapes exactly coincide with the curved geometry of the interface; see Fig. 11.
- This mesh is used to accurately incorporate the now body-aligned interface jump conditions in the DG solver.

Adaptivity: The next issue stems from the fact that there is a wide range of physical space scales involved in the process. The paint comes off the bell as a very thin film, and then breaks into small bubbles; as such, computing on a uniform mesh is impractical. Instead, we employ adaptively refined meshes wherein the mesh resolution adapts to such triggers as: (a) the distance to liquid-gas interface; (b) amount of curvature of interface; (c) the thickness of droplets, tendrils, films; and (d) the proximity to bell cup. See, for example, Fig. 12.

High performance computing: The above calculations are complex and the time step, spatial resolution, and physics make it impossible to model the entire bell. With a numerical framework targeting high performance computing facilities, using massively parallel MPI and OpenMP techniques, we can conduct high-resolution in-depth studies of rotary bell atomization on small wedges, about 5 degrees in angle, using tens of thousands of cores. In Fig. 13 we present one result from a large family of parameter studies. For further details, see [31].

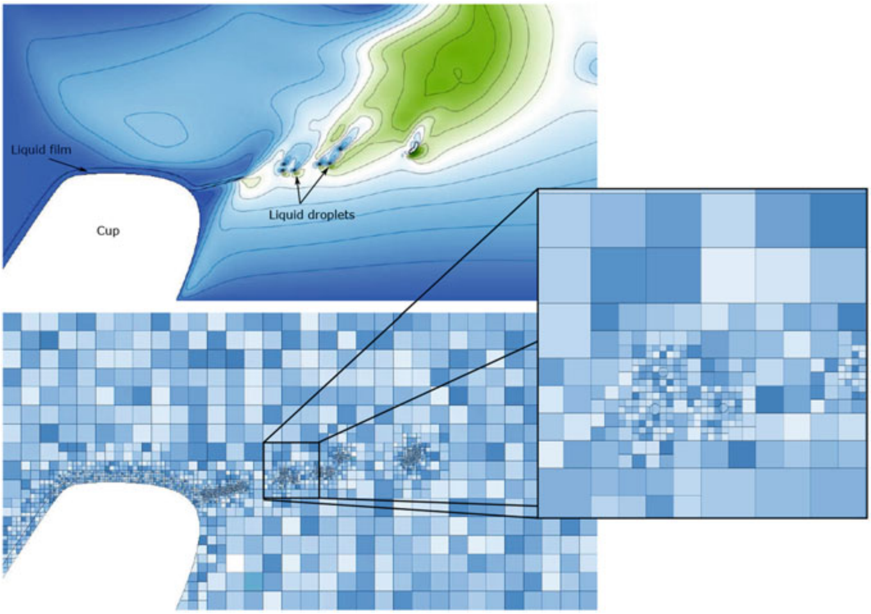
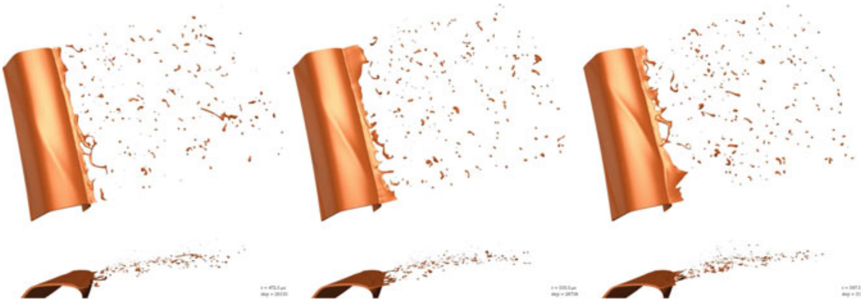


Fig. 12 Adaptively refined meshing in the rotary bell atomizing problem



$t = 472.5\,\mu\mathrm{s}$ $t = 535.0\,\mu\mathrm{s}$ $t = 597.5\,\mu\mathrm{s}$

Fig. 13 Three-dimensional model results of rotary bell atomization for time- and spatially-varying inflow film thickness, high mesh resolution, and shaping air currents simulating nozzle inlets. In each of the nine panels, two viewpoints at the same time frame are given: a top-down perspective and a side-on view to show the vertical drifting of the shedding droplets, being pushed upwards by the shaping air currents. The liquid surface is colored copper, with the bell cup situated beneath

6 Conclusions and Summary

We have tried to review a few examples in which the interface dynamics are a profound contributor to the efficiency of the industrial processes, and have focused on the application of level set methods for interface tracking to these problems. We have considered only a few contributions and works, and refer the interested reader to the referenced review articles.

Acknowledgements This work was supported in part by the Applied Mathematics Program of the U.S. DOE Office of Advanced Scientific Computing Research under contract number DE-AC02-05CH11231.

References

1. Adalsteinsson, D., Sethian, J.A.: A Fast Level Set Method for Propagating Interfaces. J. Comp. Phys. **118**(2), 269–277 (1995)
2. Adalsteinsson, D., Sethian, J.A.: The fast construction of extension velocities in level set methods. J. Comp. Phys. 148:2–22
3. Adalsteinsson, D., Sethian, J.A.: A unified level set approach to etching, deposition and lithography I: algorithms and two-dimensional simulations. J. Comp. Phys. **120**(1), 128–144 (1995)
4. Adalsteinsson, D., Sethian, J.A.: A unified level set approach to etching, deposition and lithography II: three-dimensional simulations. J. Comp. Phys. **122**(2), 348–366 (1995)
5. Adalsteinsson, D., Sethian, J.A.: A unified level set approach to etching, deposition and lithography III: complex simulations and multiple effects. J. Comp. Phys. **138**(1), 193–223 (1997)
6. Adalsteinsson, D., Sethian, J.A.: Transport and diffusion of material quantities on propagating interfaces via level set methods. J. Comp. Phys **185**(1), 271–288 (2002)
7. Almgren, A., Bell, J.B., Szymczak, W.G.: A numerical method for the incompressible Navier-Stokes equations based on an approximate projection. SIAM J. Sci. Comput. **17**(2), 358–369 (1996)
8. Bell, J.B., Colella, P., Glaz, H.M.: A second-order projection method for the incompressible Navier-Stokes equations. J. Comp. Phys. **85**, 257–283 (1989)
9. Chorin, A.J.: Numerical solution of the Navier-Stokes equations. Math. Comp. **22**, 745 (1968)
10. Garzon, M., Bobillo-Ares, Sethian, J.A.: Some Free Boundary Problems in Potential Flow Regime using a Level Set Method. Recent Advances in Fluid Mechanics. Nova Publishers (2008)
11. Garzon, M, Gray, L.G., Sethian, J.A.: Wave breaking over sloping beaches using a coupled boundary integral-level set method. Interfaces Free Boundaries **7**(3), 229–239 (2008)
12. Garzon, M., Gray, L.G., Sethian, J.A.: Numerical simulation of non-viscous liquid pinch-off using a coupled level set-boundary integral method. J. Comput. Phys. **228**(17), 6079–6106 (2009)
13. Garzon, M., Gray, L.G., Sethian, J.A.: Axisymmetric boundary integral formulation for a two-fluid system. Int. J. Numer. Meth. Fluids **69**, 1124–1134 (2012)
14. Garzon, M., Gray, L.G., Sethian, J.A.: Simulation of the droplet-to-bubble transition in a two-fluid system. Phys. Rev. E **83**, 4 (2011)
15. Garzon, M., Gray, L.G., Sethian, J.A.: Droplet and bubble pinch-off computations using level sets. J. Comput. Appl. Math. **236**(12), 3034–3041 (2012)
16. Garzon, M., Gray, L.G., Sethian, J.A.: Numerical simulations of electrostatically driven jets from non-viscous droplets. Phys. Rev. E **89**, 033011 (2014)

17. Garzon, M., Johansson, A., Sethian, J.A.: A three-dimensional coupled Nitsche and level set method for electrohydrodynamic potential flows in moving domains. J. Comput. Phys. **309**, 1–386 (2016)
18. Nemes, P., Margineau, I., Vertes, A.: Spraying mode effect on droplet formation and ion chemistry on electrosprays. Anal. Chem. **79**, 3105–3116 (2007)
19. Osher, S., Sethian, J.A.: Fronts propagating with curvature-dependent dpeed: algorithms based on Hamilton-Jacobi formulations. J. Comp. Phys. **79**, 12–49 (1988)
20. Osher, S., Fedkiw, R.: Level Set Methods and Dynamic Implicit Surfaces. Springer, Berlin (2002)
21. Salazar, A.: Computational modeling of relevant automotive rotary spray painting process. In: Toda, K., Salazar, A., Saito, K. (eds.) Automotive Painting Technology: A Monozukuri-Hitozukuri Perspective (2012)
22. Saye, R.I.: High-order quadrature methods for implicitly defined surfaces and volumes in hyperrectangles. SIAM J. Sci. Comput. **37**(2), A993–A1019 (2015)
23. Saye, R.I.: Implicit mesh discontinuous galerkin methods and interfacial gauge methods for high-order accurate interface dynamics, with applications to surface tension dynamics, rigid body fluid-structure interaction, and free surface flow: Part I. J. Comput. Phys. **344**, 647–682 (2017)
24. Saye, R.I.: Implicit mesh discontinuous galerkin methods and interfacial gauge methods for high-order accurate interface dynamics, with applications to surface tension dynamics, rigid body fluid-structure interaction, and free surface flow: Part II. J. Comput. Phys. **344**, 683–723 (2017)
25. Saye, R.I.: High-order methods for computing distances to implicitly defined surfaces. Commun. Appl. Math. Comput. Sci. **9**, 107–141 (2014)
26. Saye, R.I., Sethian, J.A.: The Voronoi implicit interface method for computing multiphase physics. In: Proceedings of the National Academy of Sciences, 21 Nov 2011
27. Saye, R.I., Sethian, J.A.: Analysis and applications of the Voronoi implicit interface method. J. Comput. Phys. **231**(18), 6051–6085 (2012)
28. Saye, R.I., Sethian, J.A.: Multi-scale modelling of membrane rearrangement, drainage, and rupture in evolving foams. Sci. Mag. **340**(6133), 720–724 (2013)
29. Saye, R.I., Sethian, J.A.: Multiscale modelling of evolving foams. J. Comput. Phys. **315**, 273–301 (2016)
30. Saye, R.I., Sethian, J.A.: A review of level set methods to model interfaces moving under complex physics: recent challenges and advances. Handbook Numerical Anal. **21**(2020), 509–554 (2020)
31. Saye, R.I., Sethian, J.A.: Numerical simulation of rotary bell dynamics in automotive painting (Work in progress) (2020)
32. Sethian, J.A.: An analysis of flame propagation. Ph.D. dissertation, Department of Mathematics, University of California, Berkeley, CA (1982)
33. Sethian, J.A.: Curvature and the evolution of fronts. Comm. in Math. Phys. **101**, 487–499 (1985)
34. Sethian, J.A.: Numerical methods for propagating fronts. In: Concus, P., Finn, R. (eds.) Variational Methods for Free Surface Interfaces. Springer, New York (1987)
35. Sethian, J.A.: Level Set Methods and Fast Marching Methods. Cambridge University Press, Cambridge (1996)
36. Sethian, J.A.: Level Set Methods and Fast Marching Methods. Cambridge University Press, Cambridge (1999)
37. Sethian, J.A.: A fast marching level set method for monotonically advancing fronts. Proc. Nat. Acad. Sci. **93**(4), 1591–1595 (1996)
38. Sethian, J.A.: Evolution, implementation, and application of level set and fast marching methods for advancing fronts. J. Comp. Phys. **169**, 503–555 (2001)
39. Sethian, J.A., Adalsteinsson, D.: An overview of level set methods for etching, deposition, and lithography development. IEEE Trans. Semiconductor Dev. **10**(1), 167–184 (1997)

40. Sethian, J.A., Smereka, P.: Level set methods for fluid interfaces. Annual review of fluid mechanics **35**, 341–372 (2003)
41. Thorodssen, S.T.: Micro-dropslets and micro-bubbles. Imaging motion at small scales. Nus. Eng. Res. News **22**(1) (2007)
42. Yu, J.-D., Sakai, S., Sethian, J.A.: A coupled level set projection method applied to ink jet simulation. Interfaces Free Boundaries **193**(1), 275–305 (2003)
43. Yu, J.-D., Sakai, S., Sethian, J.A.: A coupled quadrilateral grid level set projection method applied to ink jet simulation. J. Compu. Phys. **206**(1), 227–251 (2005)
44. Yu, J.-D., Sakai, S., Sethian, J.A.: Two-phase viscoelastic jetting. J. Comp. Phys. **220**(2), 568–585 (2007)

Numerical Study for Blood Flows in Thoracic Aorta

Hiroshi Suito, Koki Otera, Viet Q. H. Huynh, Kenji Takizawa, Naohiro Horio, and Takuya Ueda

Abstract Numerical simulations for blood flows related to cardiovascular diseases are presented. Differences in vessel morphologies produce different flow characteristics, stress distributions, and ultimately different outcomes. Some examples illustrating the effects of curvature and torsion on blood flows are presented both for simplified and patient-specific simulations. The goal of this study is to understand relationships between geometrical characteristics of blood vessels and blood flow behaviors.

1 Introduction

In aging societies, cardiovascular conditions such as aortic aneurysms and aortic dissections persist as life-threatening diseases. Moreover, congenital diseases such as hypoplastic left heart syndrome constitute an important issue for our society. In recent years, patient-specific simulations have become common in the biomedical engineering field. Several mathematical viewpoints are expected to be added and to play important roles in this context. For instance, geometrical characterization of blood vessels, which vary widely among individuals, provides useful information to medical sciences. Differences in blood vessel morphology give rise to different flow

H. Suito (✉) · V. Q. H. Huynh
Advanced Institute for Materials Research, Tohoku University, 2-1-1 Katahira,Aobaku, Sendai 980-8577, Japan
e-mail: hiroshi.suito@tohoku.ac.jp

K. Otera
Graduate School of Environmental and Life Sciences, Okayama University, Okayama, Japan

K. Takizawa
Faculty of Science and Engineering, Waseda University, Shinjuku City, Japan

N. Horio
Department of Cardiovascular Surgery, Okayama University Hospital, Okayama, Japan

T. Ueda
Department of Diagnostic Radiology, Tohoku University Hospital, Sendai, Japan

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_11

characteristics, which cause different stress distributions and outcomes. Therefore, characterization of these vessels' respective morphologies represents an important clinical question. Our objective in this study is to understand possible mechanisms connecting geometrical characteristics and stress distributions through flow behaviors. The studies presented in this paper are parts of a CREST [1] framework supported by the Japan Science and Technology Agency in a strategic area for promoting collaboration between mathematical science and other scientific fields.

2 Numerical Methods and Results

2.1 Governing Equations

We adopted incompressible Navier–Stokes equations as governing equations.

$$\begin{cases} \dfrac{\partial u_i}{\partial t} + u_j \dfrac{\partial u_i}{\partial x_j} = -\dfrac{1}{\rho}\dfrac{\partial p}{\partial x_i} + \nu \dfrac{\partial}{\partial x_j}\left(\dfrac{\partial u_i}{\partial x_j} + \dfrac{\partial u_j}{\partial x_i}\right), \\ \dfrac{\partial u_j}{\partial x_j} = 0 \end{cases} \quad \text{in } \Omega \times (0, T)\,. \qquad (1)$$

In those equations, t, u_i $(i = 1, 2, 3)$, p, ρ, and ν respectively represent time, velocity, pressure, density, and the kinematic viscosity of blood. We assumed that blood can be regarded as a Newtonian fluid in large arteries. Several numerical results with different numerical methods are presented in the following subsections. Finite difference method is used in Sect. 2.2, applied for blood flows in a thoracic aorta and for flows in simple spiral tubes to examine torsion effects. Then, finite element method is applied in Sect. 2.3 where fluid structure interaction (FSI) is considered and some flow mechanisms in a configuration after Norwood surgery are examined.

2.2 Finite Difference Approximation

2.2.1 Visualization of Flows in a Thoracic Aorta

Effects of curvature on flows in curved tubes have been discussed extensively in earlier studies [2–4]. When a tube has curvature, centrifugal force acts in the opposite direction, depending on the axial component of the velocity. Subsequently, secondary flow occurs on the cross-section and forms a set of twin vortices called Dean's vortices, thereby playing an important role in blood flow through the aortic arch where a strong curvature exists.

Figure 1 presents streamlines that can be visualized based on numerical results obtained through an earlier study [5]. We assumed a blood vessel as a rigid body and applied finite-difference method on a centerline-fitted curvilinear coordinate system, where the centerlines and cross-sections were extracted from patient-specific CT scans of patients with aortic aneurysms. Incompressible Navier–Stokes equations were solved numerically with a boundary condition for the inflow velocity profile given by a phase-contrast MRI measurement.

Figure 1a presents streamlines through the whole thoracic aorta at peak systolic phase. Circulation in the aneurysm is apparent. Figure 1b shows the Dean's vortices on the aortic arch superimposed to the main axial flow. In Fig. 1c, a spiral flow is apparent in the descending aorta.

Helicity, $\boldsymbol{u} \cdot (\nabla \times \boldsymbol{u})$, represents swirling flow regions of opposite signs. Figure 2a depicts helicity isosurfaces of a positive and a negative values, which shows Dean's vortices generated at the aortic arch and subsequently flowing down to the descending aorta. In Fig. 2b, an isosurface of the second largest eigenvalue λ_2 of $S^2 + \Omega^2$, where S and Ω respectively represent symmetric and antisymmetric parts of the velocity gradient tensor, also shows a swirling flow region [6]. Enstrophy, $|\nabla \times \boldsymbol{u}|^2$, exhibits the strength of vorticity in Fig. 2c. In Fig. 2b, c, colors of isosurfaces show λ_2 values.

2.2.2 Effects of Torsion in Simple Spiral Tubes

We also examined the effects of torsion using a pulsating flow in simple spiral tubes, as shown in [5]. Torsion of a three-dimensional curve is defined through the Frenet–Serret formula shown below.

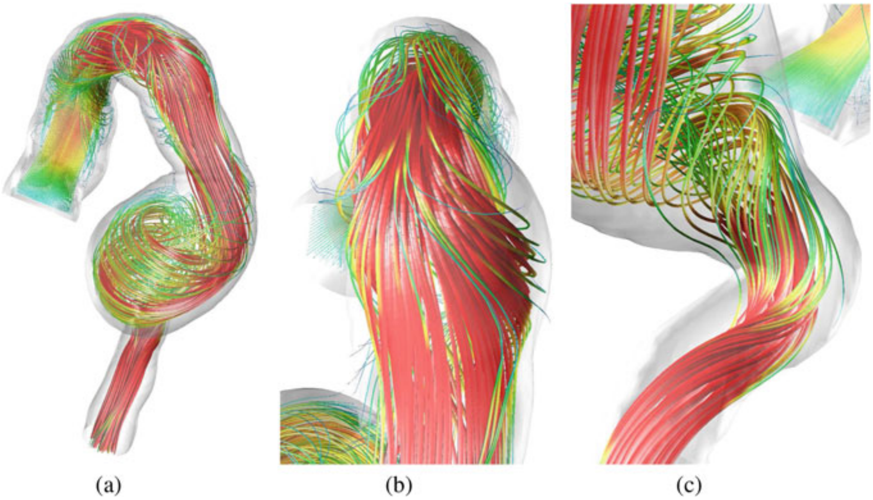


Fig. 1 Instantaneous streamlines

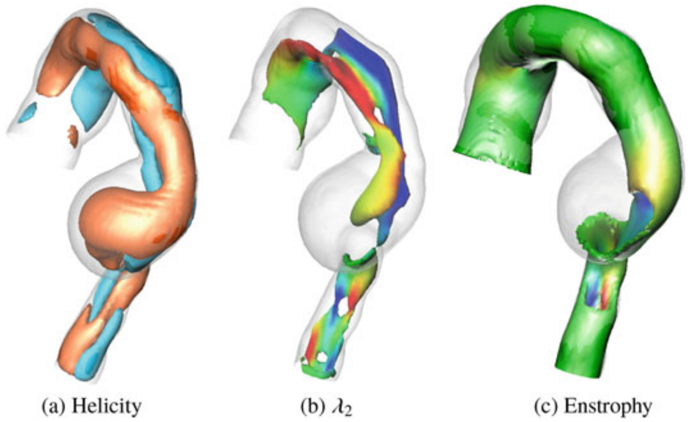


Fig. 2 Several fluid dynamics quantities

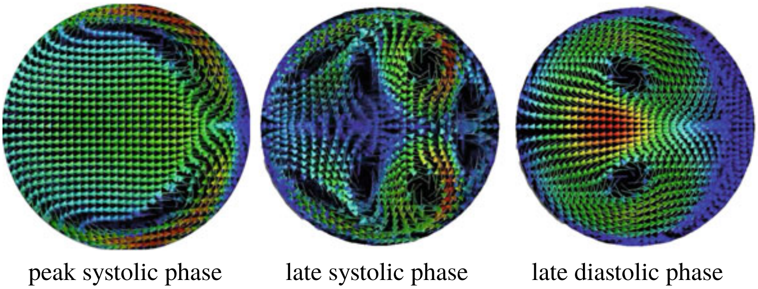


Fig. 3 Secondary flows in a zero-torsion tube

$$\frac{d}{ds}\begin{pmatrix} \boldsymbol{t} \\ \boldsymbol{n} \\ \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} 0 & \chi & 0 \\ -\chi & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{t} \\ \boldsymbol{n} \\ \boldsymbol{b} \end{pmatrix}. \tag{2}$$

Therein, χ and τ respectively represent curvature and torsion, where $\boldsymbol{t}$, $\boldsymbol{n}$, and $\boldsymbol{b}$ respectively denote the tangential, normal, and bi-normal vectors.

Figures 3 and 4 portray secondary flows, which are obtainable by subtracting the main axial flow from the total flow velocities at peak systolic, late systolic, and late diastolic phases, respectively, for zero-torsion and nonzero-torsion cases. When the torsion is zero, the secondary flow is invariably symmetric. However, when the torsion is not zero, merging phenomena occur; one large vortex persists in a diastolic phase. Such difference brings about differences in torque exerted on vessel walls.

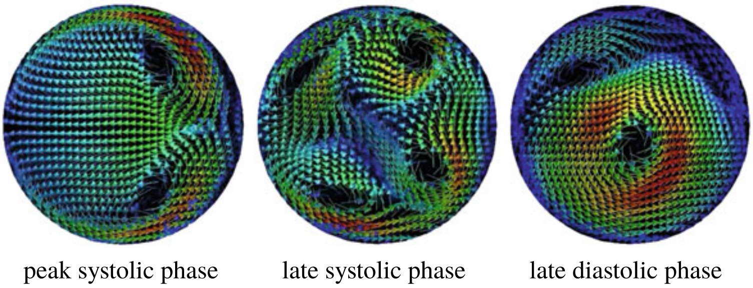


Fig. 4 Secondary flows in a nonzero-torsion tube

2.3 Finite Element Approximation

2.3.1 Torsion Effects on Flows in the Thoracic Aorta

Next we consider fluid–structure interaction (FSI) to examine torsion effects using patient-specific morphologies [7]. Here, FSI analysis is handled with the Sequentially-Coupled Arterial FSI (SCAFSI) technique [8] because the class of an FSI problem here has temporally–periodic FSI dynamics. Fluid mechanics equations are solved using Space–Time Variational Multiscale (ST-VMS) method [9–11]. First, we carry out structural mechanics computation to assess arterial deformation under an observed blood pressure profile in a cardiac cycle. Then we apply fluid mechanics computation over a mesh that moves to follow the lumen as the artery deforms. These steps are iterated where the stress obtained in fluid mechanics computation is used for the next structural mechanics computation. To assess torsion effects, the torsion-free model geometry is generated by projecting the original centerline to its averaged plane of curvature, as presented in Fig. 5.

Figure 6 presents secondary flows. On the left-hand side (projected shape), symmetric Dean's vortices are apparent, although they are not visible on the right-hand side (original shape), similarly to the simple spiral tubes in Fig. 4.

Next we compare the wall shear stresses (WSS) patterns corresponding to the projected and the original geometries to examine the influence of torsion. Figure 7 presents WSS at peak systolic phase. In the projected torsion-free shape, a high WSS region is apparent at the aortic arch, which results from the strong Dean's twin vortices, although it is not apparent in the original shape with torsion there.

2.3.2 Flow Mechanism in Morphology After Norwood Surgery

This subsection presents examples of patient-specific blood flow simulations at an anastomosis site after Norwood surgery for hypoplastic left heart syndrome. Our target is the geometry surrounding an anastomosis site of the aortic arch and pulmonary artery after Norwood surgery, which is one step taken during surgeries for

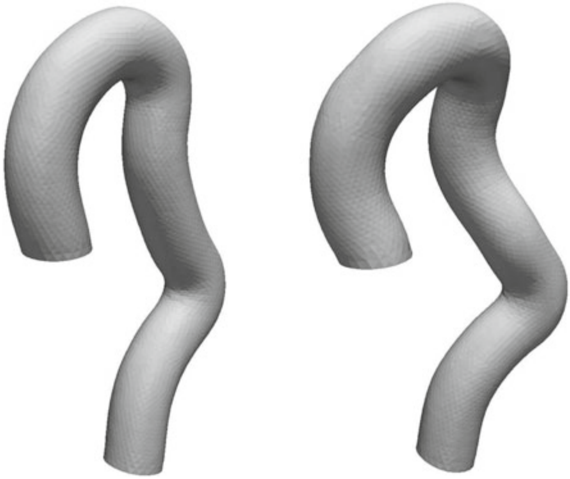
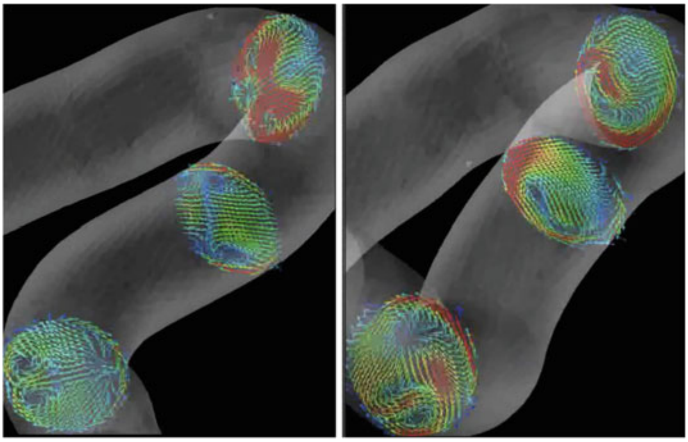


Fig. 5 Projected and original shapes



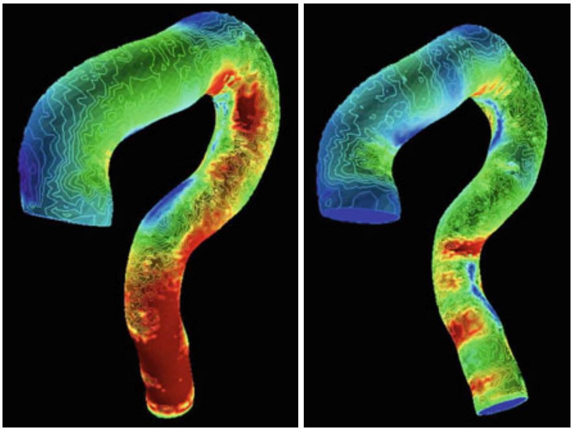
(a) Projected shape (b) Original shape

Fig. 6 Secondary flows in projected and original shapes

hypoplastic left heart syndrome. The target geometry was extracted from a CT scan with boundary conditions obtained from ultrasound measurements. Here, we again adopt the rigid body assumption, i.e., not considering fluid–structure interactions. The SUPG/PSPG stabilized finite element formulation is used, which is solved on P1/P1 elements.

Figure 8a portrays instantaneous streamlines at the peak systolic phase, whereas Fig. 8b depicts the energy-dissipation distribution. Energy dissipation is a clinically important quantity because it imposes a load on the heart directly [12]. In Fig. 8b, high energy dissipation is apparent at the anastomosis site, which can be understood

Fig. 7 Wall shear stresses at peak systolic phase


(a) Projected shape (b) Original shape

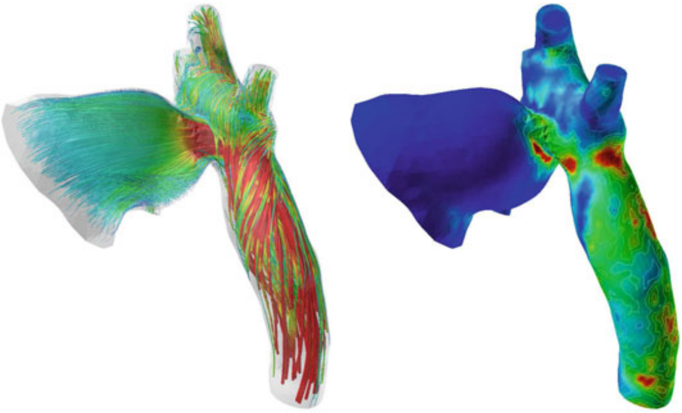

(a) Streamlines (b) Energy dissipation

Fig. 8 Streamlines at an anastomosis site after Norwood surgery

straightforwardly because the velocity is extremely high there. Although high energy dissipation is also apparent in the descending aorta, it cannot be qualified straightforwardly. This dissipation apparently derives from spiral flow there, which is generated at the aortic arch immediately after blood passes out of the thin anastomosis channel, as shown in Fig. 9. Here, a relation can be found between morphology and energy dissipation patterns through flow structures.

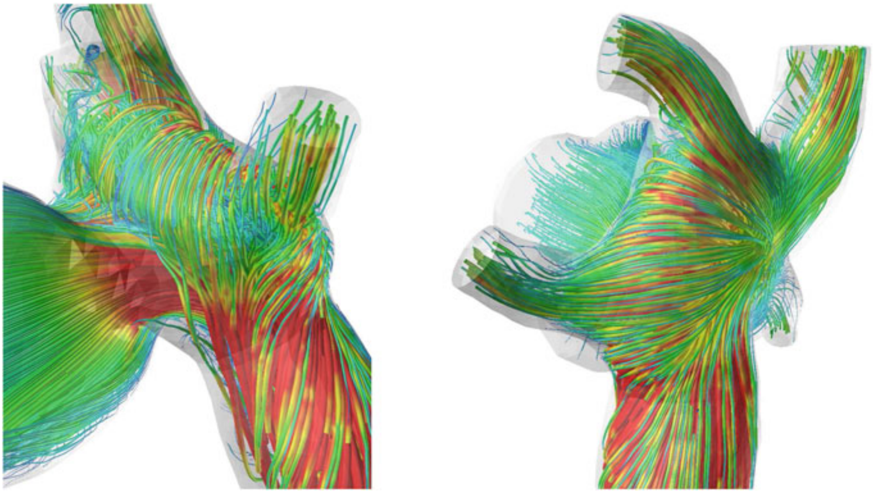


Fig. 9 Front and back views of streamlines

3 Conclusions

We have presented some relations between geometrical characteristics of blood vessels and flow behaviors. Those relations are expected to explain how and why vessel morphologies affect WSS distributions and energy dissipations. As described in Sect. 2.2, vessel curvature induces Dean's vortices as a secondary flow by centrifugal force, thereby creating strong WSS there. Moreover, Dean's vortices show different behaviors depending on the existence of torsion. In the example from a Norwood surgery morphology, an energy dissipation pattern on the descending aorta can be explained through flow structures. As a next step, predictions based on geometrical characteristics of blood vessels are expected to contribute to better risk assessments and surgery planning through mathematical modellings and numerical simulations.

Acknowledgements This work was supported by JST CREST Grant Number JPMJCR15D1, Japan.

References

1. Japan Science and Technology Agency website: https://www.jst.go.jp/kisoken/crest/en/
2. Germano, M.: On the effect of torsion on a helical pipe flow. J. Fluid Mech. **125**, 1–8 (1982)
3. Berger, S.A., Talbot, L., Yao, L.-S.: Flow in curved pipes. Ann. Rev. Fluid Mech. **15**, 461–512 (1983)
4. Lee, K.E., Parker, K.H., Caro, C.G., Sherwin, S.J.: The spectral/HP element modeling of steady flow in non-planar double bends. Int. J. Num. Meth. Fluids **57**, 519–529 (2008)

5. Suito, H., Ueda, T., Sze, D.: Numerical simulation of blood flow in the thoracic aorta using a centerline-fitted finite difference approach. Jap. J. Ind. Appl. Math. **30**(3), 701–710 (2013)
6. Jeong, J., Hussain, F.: On the identification of a vortex. J. Fluid Mech. **285**, 69–94 (1995)
7. Suito, H., Takizawa, K., Huynh, V.Q.H., Sze, D., Ueda, T.: FSI analysis of the blood flow and geometrical characteristics in the thoracic aorta. Comput. Mech. **54**(4), 1035–1045 (2014)
8. Tezduyar, T.E., Takizawa, K., Moorman, C., Wright, S., Christopher, J.: Multiscale sequentially-coupled arterial FSI technique. Comput. Mech. **46**, 17–29 (2010)
9. Takizawa, K., Tezduyar, T.E.: Multiscale space-time fluid-structure interaction techniques. Comput. Mech. **48**, 247–267 (2011)
10. Takizawa, K., Tezduyar, T.E., Buscher, A., Asada, S.: Space-time interface-tracking with topology change (ST-TC). Comput. Mech. **54**(4), 955–971 (2014)
11. Takizawa, K., Tezduyar, T.E., Busche, A.: r and S. Asada, space-time fluid mechanics computation of heart valve models. Comput. Mech. **54**(4), 973–986 (2014)
12. Ueda, T., Suito, H., Ota, H., Takase, K.: Computational fluid dynamics modeling in aortic diseases. Cardiovasc. Imag. Asia **2**(2), 58–64 (2018)

An Iterative Thresholding Method for Topology Optimization for the Navier–Stokes Flow

Haitao Leng, Dong Wang, Huangxin Chen, and Xiao-Ping Wang

Abstract We develop an efficient iterative thresholding method for topology optimization for the Navier–Stokes flow. The method is proposed to minimize an objective energy functional which consists of the potential power in the fluid and a fluid-solid interface perimeter penalization. The perimeter is approximated by a nonlocal energy, subject to a fluid volume constraint and the incompressible Navier–Stokes equation. The method is an iterative scheme which alternates two steps: (1) solving a system containing the Brinkman equation and an adjoint system, and (2) convolution and thresholding. Various numerical experiments in both two and three dimensions are given to show the performance of the proposed method.

1 Introduction

Topology optimization was originally developed for the optimal design in structural mechanics ([3, 4, 6]). Nowadays it has attracted much attention due to its wide application in the fields of industry problems such as optimization of transport vehicles, biomechanical structure, etc. So far, the density method [5, 31] has been well devel-

H. Leng
School of Mathematical Sciences, South China Normal University, Guangzhou 510631, Guangdong, China
e-mail: htleng@m.scnu.edu.cn

D. Wang
School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, Guangdong, China
e-mail: wangdong@cuhk.edu.cn

H. Chen
School of Mathematical Sciences and Fujian Provincial Key Laboratory on Mathematical Modeling and High Performance Scientific Computing, Xiamen University, Fujian 361005, China
e-mail: chx@xmu.edu.cn

X.-P. Wang (✉)
Department of Mathematics, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China
e-mail: mawang@ust.hk

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_12

oped for implementation of topology optimization. It was originally developed for the design of stiffness and compliant mechanism [32, 33] and has been applied in various physical problems such as acoustics, electromagnetics, fluid flow, and thermal problems [7, 11, 15, 24, 34]. In fluid mechanics, the concept of density method was first developed by Borrvall and Petersson [7] for topology optimization for the Stokes flow. Then it was extended to the Darcy-Stokes flow [21, 43], the Navier–Stokes flow [12, 18, 20, 27, 36, 47], the non-Newtonian flow [30], the turbulent flow [13], and more complicated fluidic devices [1, 25, 26]. Approaches using the topological sensitivity analysis (providing an asymptotic expansion of a shape function with respect to the size of a small inclusion inserted inside the domain) can also be used for shape optimization for Stokes flows [22] and Navier–Stokes flows [2]. Generally, the discrete optimization problem for the topology optimization was solved by the method of moving asymptotes (MMA) [35], level set based methods [8, 36, 47] and phase field based methods [18].

The threshold dynamics method developed by Merriman, Bence and Osher (MBO) [23] is an efficient method for approximating the mean curvature flow. In this method, the interface is implicitly represented by the characteristic functions of the domains. It alternates two simple steps: convolution between the characteristic functions and a heat kernel and point-wise thresholding. Recently, Esedoglu and Otto generalized the original MBO method to multiphase problems with arbitrary surface tensions [17]. The method has attracted much attention and it has been extended to many other applications, such as image processing [16, 37, 39], wetting dynamics [38, 44, 45], and target-valued problems [28, 29, 40–42].

In this paper we extend the iterative thresholding method developed in [9] to topology optimization for the Navier–Stokes flow. The porous medium approach based on the density method is utilized in the algorithm, and a Darcy term is introduced into the Navier–Stokes equation to "interpolate" between the Navier–Stokes equation in the fluid region and the Darcy flow through a porous medium (a weakened solid region with low permeability) (i.e., Brinkman equation). Then the total energy consists of the potential power in the fluid, the perimeter regularization, and a Darcy term. The perimeter term is computed based on the convolution between the heat kernel and the characteristic functions of regions. There are two steps per iteration in the proposed algorithm. The first step is to solve the Brinkman equation and an adjoint system, which can both be efficiently solved using the mixed finite element method. The second step is to update the fluid-solid regions by a simple convolution and thresholding step. The convolution can be efficiently computed on a uniform grid by the fast Fourier transform (FFT) with the computational complexity $O(N \log N)$. A variety of numerical experiments in both two and three dimensions are shown to verify the efficiency of the proposed algorithm. In addition, numerical results indicate that the total energy decays.

The paper is organized as follows. In Sect. 2, we introduce the mathematical model, the approximation to the model, and the derivation of the iterative thresholding method. The numerical implementation is discussed in Sect. 3. We verify the performance through extensive numerical experiments in Sect. 4. We draw some conclusions in Sect. 5.

2 Derivation of the Method

2.1 *The Mathematical Model*

In this section, we consider the mathematical model for topology optimization for the Navier–Stokes flow. Denote $\Omega \in \mathbb{R}^d$ $(d = 2, 3)$ as the computational domain which is fixed throughout optimization and assume that Ω is a bounded Lipschitz domain with an outer unit normal $\mathbf{n}$ such that $\mathbb{R}^d \setminus \overline{\Omega}$ is connected. Furthermore, we denote $\Omega_0 \subset \Omega$ as the domain of the fluid which is a Caccioppoli set whose boundary is measurable and has a (at least locally) finite measure and $\Omega \setminus \Omega_0$ as the domain of solid. Our goal is to determine an optimal shape of Ω_0 that minimizes the following objective functional consisting of the total potential power and a perimeter regularization term,

$$\min_{(\Omega_0,\mathbf{u})} J_0(\Omega_0, \mathbf{u}) = \int_{\Omega} \frac{\mu}{2} |\nabla \mathbf{u}|^2 d\mathbf{x} + \gamma |\Gamma| \tag{1}$$

subject to

$$\nabla \cdot \mathbf{u} = 0, \quad \text{in} \;\; \Omega_0, \tag{2a}$$

$$(\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p - \nabla \cdot (\mu \nabla \mathbf{u}) = 0, \quad \text{in} \;\; \Omega_0, \tag{2b}$$

$$\mathbf{u} = 0, \quad \text{in } \Omega \setminus \overline{\Omega}_0 \quad \text{and on } \partial\Omega_0, \tag{2c}$$

$$\mathbf{u}|_{\partial\Omega} = \mathbf{u}_D, \quad \text{on} \;\; \partial\Omega, \tag{2d}$$

$$|\Omega_0| = \; \beta|\Omega| \text{ with a fixed parameter } \beta \in (0, 1). \tag{2e}$$

Here, $\mathbf{u} : \Omega \to \mathbb{R}^d$, μ is the viscosity of the fluid, p is the pressure, $\mathbf{u}_D : \partial\Omega \to \mathbb{R}^d$ is a given function, $|\Gamma|$ is the perimeter of the boundary (i.e., $\Gamma = \partial\Omega_0$), and $\gamma > 0$ is a weighting parameter.

2.2 *The Relaxation and Approximation of the Problem*

Since the goal is to minimize the objective functional (1) subject to several constraints (2) with respect to the fluid-solid interface, it is necessary to have a proper representation of the fluid-solid interface. Motivated by [9, 17, 37, 44], in this paper, we use the characteristic function χ_1 of the fluid domain (i.e., Ω_0) to implicitly represent the fluid-solid interface, i.e.,

$$\chi_1(\mathbf{x}) := \begin{cases} 1, & \textit{if } \mathbf{x} \in \Omega_0, \\ 0, & \textit{otherwise}. \end{cases}$$

$\chi_2(\mathbf{x}) = 1 - \chi_1(\mathbf{x})$ is denoted as the characteristic function of $\Omega \setminus \Omega_0$. Then, the interface Γ is implicitly represented by χ_1 and χ_2. Under this representation, $|\Gamma|$ can be approximated by

$$|\Gamma| \approx \sqrt{\frac{\pi}{\tau}} \int\limits_{\Omega} \chi_1 G_\tau * \chi_2 d\mathbf{x} = \sqrt{\frac{\pi}{\tau}} \int\limits_{\Omega} \chi_1 G_\tau * (1 - \chi_1) d\mathbf{x}, \tag{3}$$

where $G_\tau(\mathbf{x}) = \dfrac{1}{(4\pi\tau)^{\frac{d}{2}}} \exp\left(-\dfrac{|\mathbf{x}|^2}{4\tau}\right)$ $(d = 2, 3)$ is the Gaussian kernel and $*$ denotes the convolution [17].

Similar to [9], to avoid solving the Navier–Stokes equation in a changing domain at each iteration, the porous medium approach [18] is utilized to "interpolate" between the Navier–Stokes equation in the fluid region (i.e., $\{\mathbf{x}|\ \chi_1(\mathbf{x}) = 1\}$) and $\mathbf{u} = 0$ in the solid region (i.e., $\{\mathbf{x}|\ \chi_2(\mathbf{x}) = 1\}$) by introducing an additional penalization term, $\alpha(\mathbf{x})\mathbf{u}$, as follows:

$$\nabla \cdot \mathbf{u} = 0, \quad \text{in } \Omega, \tag{4a}$$

$$(\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p - \nabla \cdot (\mu \nabla \mathbf{u}) + \alpha(\mathbf{x})\mathbf{u} = 0, \quad \text{in } \Omega, \tag{4b}$$

$$\mathbf{u}|_{\partial\Omega} = \mathbf{u}_D, \quad \text{on } \partial\Omega, \tag{4c}$$

$$\int\limits_{\Omega} \chi_1 d\mathbf{x} = \beta|\Omega|. \tag{4d}$$

Accordingly, the original objective functional (1) can be approximated by adding a Darcy penalty term as follows:

$$J^\tau(\chi, \mathbf{u}) = \int\limits_{\Omega} \left(\frac{\mu}{2}|\nabla \mathbf{u}|^2 + \frac{\alpha(\mathbf{x})}{2}|\mathbf{u}|^2\right) d\mathbf{x} + \gamma \sqrt{\frac{\pi}{\tau}} \int\limits_{\Omega} \chi G_\tau * (1 - \chi) d\mathbf{x} \tag{5}$$

where χ denotes the characteristic function of the solid domain, i.e., $\chi = \chi_2$.

Now, we discuss the computation of α in the current representation of the interface (i.e., using characteristic functions). Theoretically, α should be large enough in the solid domain to penalize the condition $\mathbf{u} = 0$ and close to 0 in the fluid domain to make $\mathbf{u}$ satisfy the Navier–Stokes equation. For numerical considerations, we relax α to a smooth function which undergoes rapid changes through the interface. We use the 0.5 level set of $\varphi = G_\tau * \chi$ to approximate the position of the interface Γ and such φ is a smooth function between [0, 1] and admits a change from 0 to 1 in an $O(\sqrt{\tau})$ transition region. Thus, we compute α by

$$\alpha(\mathbf{x}) = \bar{\alpha}\varphi = \bar{\alpha} G_\tau * \chi \tag{6}$$

where $\bar{\alpha}$ is a sufficiently large constant, and thus by the porous medium approach we can solve the system (4) in a fixed domain Ω.

Finally, using (6), we arrive in the following formulation of the problem:

$$\min_{\chi,\mathbf{u}} J^{\tau}(\chi,\mathbf{u}) = \int_{\Omega} \left(\frac{\mu}{2}|\nabla \mathbf{u}|^2 + \frac{\bar{\alpha}}{2}(G_\tau * \chi)|\mathbf{u}|^2 + \gamma\sqrt{\frac{\pi}{\tau}}\chi G_\tau * (1-\chi) \right) d\mathbf{x} \quad (7)$$

subject to

$$\chi \in \mathcal{B} := \{\chi \in BV(\Omega) \mid \chi(x) = \{0,1\},\ a.e.,\ \text{and} \int_{\Omega} (1-\chi) d\mathbf{x} = \beta|\Omega|\} \quad (8a)$$

$$\nabla \cdot \mathbf{u} = 0, \ \text{ in } \ \Omega, \quad (8b)$$

$$(\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p - \nabla\cdot(\mu\nabla\mathbf{u}) + (\bar{\alpha}G_\tau * \chi)\mathbf{u} = 0, \ \text{ in } \ \Omega, \quad (8c)$$

$$\mathbf{u}|_{\partial\Omega} = \mathbf{u}_D, \ \text{ on } \ \partial\Omega. \quad (8d)$$

2.3 Derivation of the Method

In this section, we will derive an iterative scheme to find the approximate solution for (7) and (8). Denote

$$\mathbf{U} := \{\mathbf{u} \in H^1(\Omega) | \nabla\cdot\mathbf{u} = 0, \mathbf{u}|_{\partial\Omega} = \mathbf{u}_D\} \quad \text{and} \quad \mathbf{V} := \{\mathbf{v} \in H_0^1(\Omega) | \nabla\cdot\mathbf{v} = 0\}.$$

To derive the first order necessary optimality conditions for a solution $(\chi_\tau, \mathbf{u}_\tau)$ of (7) and (8), we introduce the Lagrangian $\mathcal{E} : \mathcal{B} \times \mathbf{U} \times \mathbf{V} \to \mathbb{R}$ by

$$\mathcal{E}^{\tau}(\chi,\mathbf{u},\tilde{\mathbf{u}}) := J^{\tau}(\chi,\mathbf{u}) + \int_{\Omega} (\mathbf{u}\cdot\nabla)\mathbf{u}\cdot\tilde{\mathbf{u}} + \mu\nabla\mathbf{u}\cdot\nabla\tilde{\mathbf{u}} + (\bar{\alpha}G_\tau * \chi)\mathbf{u}\cdot\tilde{\mathbf{u}} d\mathbf{x}$$

where the pressure term is not shown because $\nabla\cdot\tilde{\mathbf{u}} = 0$. The variational inequality is formally derived by

$$\left\langle \frac{\delta\mathcal{E}^{\tau}}{\delta\chi}(\chi_\tau, \mathbf{u}_\tau, \tilde{\mathbf{u}}_\tau), \chi - \chi_\tau \right\rangle \geq 0, \quad \forall\, \chi \in \mathcal{B} \quad (9)$$

and the adjoint equation can be deduced by

$$\left\langle \frac{\delta\mathcal{E}^{\tau}}{\delta\mathbf{u}}(\chi_\tau, \mathbf{u}_\tau, \tilde{\mathbf{u}}_\tau), \mathbf{v} \right\rangle = 0, \quad \forall\, \mathbf{v} \in \mathbf{V} \quad (10)$$

where $\langle\cdot,\cdot\rangle$ denotes the L^2-inner product.

To be specific, assume $(\chi_\tau, \mathbf{u}_\tau) \in \mathcal{B} \times \mathbf{U}$ is a minimizer of (7) and (8), the following inequality is fulfilled:

$$\left\langle \frac{\bar{\alpha}}{2} G_\tau * |\mathbf{u}_\tau|^2 + \gamma \sqrt{\frac{\pi}{\tau}} G_\tau * (1 - 2\chi_\tau) + \bar{\alpha} G_\tau * (\mathbf{u}_\tau \cdot \tilde{\mathbf{u}}_\tau), \chi - \chi_\tau \right\rangle \geq 0, \ \forall \chi \in \mathcal{B} \tag{11}$$

where $\tilde{\mathbf{u}}_\tau$ is the solution to the following adjoint system at $(\mathbf{u}_\tau, \chi_\tau)$:

$$-(\mathbf{u}_\tau \cdot \nabla)\mathbf{u}_\tau - (\mathbf{u}_\tau \cdot \nabla)\tilde{\mathbf{u}} + (\nabla \mathbf{u}_\tau)^T \tilde{\mathbf{u}} + \nabla \tilde{p} - \nabla \cdot (\mu \nabla \tilde{\mathbf{u}}) + (\bar{\alpha} G_\tau * \chi_\tau)\tilde{\mathbf{u}} = 0, \tag{12a}$$

$$\nabla \cdot \tilde{\mathbf{u}} = 0, \tag{12b}$$

$$\tilde{\mathbf{u}}|_{\partial\Omega} = 0. \tag{12c}$$

Here, $\tilde{p}$ is the pressure associated to the adjoint system.

Based on the first order necessary optimality condition, to solve (7) and (8), we use an iterative scheme to decrease the value of the objective functional with $\mathbf{u}$ satisfying (8) and $\tilde{\mathbf{u}}$ satisfying (12). Without loss of generality, assume the k-th iteration χ^k is given, we compute $(\mathbf{u}^k, \tilde{\mathbf{u}}^k)$ via solving the following system

$$\begin{cases} \nabla \cdot \mathbf{u} = 0, \\ \nabla \cdot \tilde{\mathbf{u}} = 0, \\ (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p - \nabla \cdot (\mu \nabla \mathbf{u}) + (\bar{\alpha} G_\tau * \chi^k)\mathbf{u} = \mathbf{f}, \\ -(\mathbf{u} \cdot \nabla)\mathbf{u} - (\mathbf{u} \cdot \nabla)\tilde{\mathbf{u}} + (\nabla \mathbf{u})^T \tilde{\mathbf{u}} + \nabla \tilde{p} - \nabla \cdot (\mu \nabla \tilde{\mathbf{u}}) + (\bar{\alpha} G_\tau * \chi^k)\tilde{\mathbf{u}} = 0, \\ \mathbf{u}|_{\partial\Omega} = \mathbf{u}_D, \\ \tilde{\mathbf{u}}|_{\partial\Omega} = 0. \end{cases} \tag{13}$$

After $(\mathbf{u}^k, \tilde{\mathbf{u}}^k)$ are solved from (13), χ^{k+1} is updated through

$$\chi^{k+1} = \arg\min_{\chi \in \mathcal{B}} \mathcal{E}^\tau(\chi, \mathbf{u}^k, \tilde{\mathbf{u}}^k). \tag{14}$$

Write the objective functional $\mathcal{E}^\tau(\chi, \mathbf{u}^k, \tilde{\mathbf{u}}^k)$ into $\tilde{\mathcal{E}}^{\tau,k}(\chi)$:

$$\begin{aligned} \tilde{\mathcal{E}}^{\tau,k}(\chi) := \mathcal{E}^\tau(\chi, \mathbf{u}^k, \tilde{\mathbf{u}}^k) = & \int_\Omega \frac{\bar{\alpha}}{2} \chi G_\tau * |\mathbf{u}^k|^2 d\mathbf{x} + \gamma \sqrt{\frac{\pi}{\tau}} \int_\Omega \chi G_\tau * (1 - \chi) d\mathbf{x} \\ & + \int_\Omega \bar{\alpha} \chi G_\tau * (\mathbf{u}^k \cdot \tilde{\mathbf{u}}^k) d\mathbf{x} + \mathcal{N}(\mathbf{u}^k, \tilde{\mathbf{u}}^k), \end{aligned}$$

where $\mathcal{N}(\mathbf{u}^k, \tilde{\mathbf{u}}^k)$ contains all other terms in $\mathcal{E}^\tau(\chi, \mathbf{u}^k, \tilde{\mathbf{u}}^k)$ which are independent of χ. The only problem now is to minimize $\tilde{\mathcal{E}}^{\tau,k}(\chi)$ on $\mathcal{B}$, i.e., finding χ^{k+1} such that

$$\chi^{k+1} = \arg\min_{\chi\in\mathcal{B}} \tilde{\mathcal{E}}^{\tau,k}(\chi). \tag{16}$$

We first relax (16) to a problem defined on a convex admissible set by finding r^{k+1} such that

$$r^{k+1} = \arg\min_{r\in\mathcal{H}} \tilde{\mathcal{E}}^{\tau,k}(r), \tag{17}$$

where $\mathcal{H}$ is the convex hull of $\mathcal{B}$:

$$\mathcal{H} := \{r \in BV(\Omega) \mid r(x) \in [0, 1]\ a.e., \text{ and} \int_\Omega r d\mathbf{x} = V_0\}.$$

The following lemma holds similarly as that in [9] and we refer the details of a similar proof to [9]. Thus, we can solve the relaxed problem (17) instead of (16).

Lemma 2.1 *Let* $\mathbf{u} \in H^1_{\mathbf{u}_D}(\Omega, \mathbb{R}^d)$ *be a given function and* $r = (r_1, r_2)$. *Then we have*

$$\arg\min_{r\in\mathcal{H}} \tilde{\mathcal{E}}^{\tau,k}(r) = \arg\min_{r\in\mathcal{B}} \tilde{\mathcal{E}}^{\tau,k}(r).$$

Next we show that (17) can be solved by a thresholding step. Because $\tilde{\mathcal{E}}^{\tau,k}(r)$ is quadratic and concave in r, we first linearize the energy $\tilde{\mathcal{E}}^{\tau,k}(r)$ at r^k by

$$\tilde{\mathcal{E}}^{\tau,k}(r) \approx \tilde{\mathcal{E}}^{\tau,k}(r^k) + \mathcal{L}^{\tau,k}_{r^k}(r - r^k),$$

where

$$\begin{aligned}\mathcal{L}^{\tau,k}_{r^k}(r) &= \int_\Omega \left(\gamma\sqrt{\frac{\pi}{\tau}} r G_\tau * (1 - 2r^k) + r\frac{\bar{\alpha}}{2} G_\tau * |\mathbf{u}^k|^2 + r\bar{\alpha} G_\tau * (\mathbf{u}^k \cdot \tilde{\mathbf{u}}^k)\right) d\mathbf{x} \\ &= \int_\Omega r\phi d\mathbf{x}\end{aligned}$$

where $\phi = \gamma\sqrt{\frac{\pi}{\tau}} G_\tau * (1 - 2r^k) + \frac{\bar{\alpha}}{2} G_\tau * |\mathbf{u}^k|^2 + \bar{\alpha} G_\tau * (\mathbf{u}^k \cdot \tilde{\mathbf{u}}^k)$. Then (17) can be approximately solved by

$$\chi^{k+1} = \arg\min_{r\in\mathcal{H}} \mathcal{L}^{\tau,k}_{r^k}(r) = \arg\min_{r\in\mathcal{H}} \int_\Omega r\phi d\mathbf{x}. \tag{18}$$

Then we have the following lemma as in [9] and one can also refer the details of proof to [9].

Lemma 2.2 *Let* $\phi = \gamma\sqrt{\frac{\pi}{\tau}}G_\tau * (1-2\chi^k) + \frac{\bar{\alpha}}{2}G_\tau * |\mathbf{u}^k|^2 + \bar{\alpha}G_\tau * (\mathbf{u}^k \cdot \tilde{\mathbf{u}}^k)$ *and*

$$D_2^{k+1} = \{\mathbf{x} \in \Omega |\ \phi < \delta\}$$

for some δ *such that* $|D_2^{k+1}| = (1-\beta)|\Omega|$. *Then with* $\chi^{k+1} = \chi_{D_2^{k+1}}$, *we have*

$$\mathcal{L}^{\tau,k}_{\chi^k}(\chi^{k+1}) \le \mathcal{L}^{\tau,k}_{\chi^k}(\chi^k) \quad \textit{for all}\ \tau > 0.$$

The above lemma shows that (18) can be solved by

$$\begin{cases} \chi^{k+1}(\mathbf{x}) = 1, & \text{if } \ \phi(\mathbf{x}) < \delta, \\ \chi^{k+1}(\mathbf{x}) = 0, & \text{otherwise}, \end{cases}$$

where δ is chosen as a constant such that $\int_\Omega \chi^{k+1} d\mathbf{x} = (1-\beta)|\Omega|$.

To determine the value of δ, one can treat $\int_\Omega \chi^{k+1} d\mathbf{x} - (1-\beta)|\Omega|$ as a function of δ (i.e., $f(\delta) = \int_\Omega \chi^{k+1} d\mathbf{x} - (1-\beta)|\Omega|$) and use an iteration method (e.g., bisection method or Newton's method) to find the root of $f(\delta) = 0$. For the uniform discretization of Ω, a more efficient method is the quick-sort technique proposed in [44]. Assume we have a uniform discretization of Ω with grid size h, we can approximate $\int_\Omega \chi^{k+1} d\mathbf{x}$ by mh^d where m is the number of grid points where $\chi^{k+1} = 1$. Assume $(1-\beta)|\Omega|$ is approximated by Mh^d, we then sort the values of ϕ in an ascending order and simply set $\chi^{k+1} = 1$ on the first M points.

Now, we arrive at Algorithm 1.

Remark 2.1 We remark here that it's obvious that the Step 2 in Algorithm 1 decreases the energy which can be proved similar as we did in [9], i.e.,

$$J^\tau(\chi^{k+1}, \mathbf{u}^k) \le J^\tau(\chi^k, \mathbf{u}^k).$$

In the Step 1, we don't have

$$J^\tau(\chi^k, \mathbf{u}^k) \le J^\tau(\chi^k, \mathbf{u}^{k-1})$$

because this step can be interpreted as a projection step. It could increase the value of the energy. However, in the numerical experiments in Sect. 4, we checked the energy curves for all examples as displayed. All of them indicate that the algorithm has the energy decaying property.

Remark 2.2 In the implementation, the stopping criteria is $\chi^{k+1} = \chi^k$ on each grid point. It is easy to see that the stationary solution (obtained from Algorithm 1) satisfies the first order necessary optimality condition (8), (9), and (10).

Algorithm 1 An iterative thresholding method for topology optimization for the Navier-Stokes flow

Input: Discretize Ω uniformly into a grid $\mathcal{T}_h$ with grid size h and set $M = (1-\beta)|\Omega|/h^d$. Set $\tau > 0, \bar{\alpha} > 0, k = 0$, a tolerance parameter $tol > 0$ and give the initial guess $\chi^0 \in \mathcal{B}$.

Iterative solution:

Step 1. Given χ^k, update u and ũ. Solve the following system

$$\begin{cases} \nabla \cdot \mathbf{u} = 0, \\ \nabla \cdot \tilde{\mathbf{u}} = 0, \\ (\mathbf{u} \cdot \nabla)\mathbf{u} + \nabla p - \nabla \cdot (\mu \nabla \mathbf{u}) + (\bar{\alpha} G_\tau * \chi^k)\mathbf{u} = \mathbf{f}, \\ -(\mathbf{u} \cdot \nabla)\mathbf{u} - (\mathbf{u} \cdot \nabla)\tilde{\mathbf{u}} + (\nabla \mathbf{u})^T \tilde{\mathbf{u}} + \nabla \tilde{p} - \nabla \cdot (\mu \nabla \tilde{\mathbf{u}}) + (\bar{\alpha} G_\tau * \chi^k)\tilde{\mathbf{u}} = 0, \\ \mathbf{u}|_{\partial\Omega} = \mathbf{u}_D, \\ \tilde{\mathbf{u}}|_{\partial\Omega} = 0. \end{cases}$$

to obtain $\mathbf{u}^k$ and $\tilde{\mathbf{u}}^k$.

Step 2. Update χ. Evaluate

$$\phi = \gamma \sqrt{\frac{\pi}{\tau}} G_\tau * (1 - 2\chi^k) + \frac{\bar{\alpha}}{2} G_\tau * |\mathbf{u}^k|^2 + \bar{\alpha} G_\tau * (\mathbf{u}^k \cdot \tilde{\mathbf{u}}^k),$$

sort the values of ϕ in an ascending order, and set $\chi^{k+1} = 1$ on the first M points.

Step 3. Compute $e_\chi^k = \|\chi^{k+1} - \chi^k\|_2$. If $e_\chi^k \le tol$, stop the iteration and go to the output step. Otherwise, let $k + 1 \to k$ and continue the iteration.

Output: $(\chi, \mathbf{u})$ that approximately solves (7) subject to (8)(a-d).

3 Numerical Implementation

Now we illustrate the implementation of Algorithm 1 and we focus on Step 1. The Navier-Stokes equations with a Dacry term penalty and the adjoint problem (13) are solved by the mixed finite element method, and the standard Taylor-Hood finite element space is used for discretization. Let $\mathcal{T}_h$ be a uniform grid of the domain Ω, and $\mathcal{N}_h$ is the set of all vertices of $\mathcal{T}_h$. For a given $\overline{\chi}_h \in \mathcal{B}_h$ where $\mathcal{B}_h$ is the discrete version of $\mathcal{B}$ defined on $\mathcal{N}_h$. We introduce the Taylor-Hood finite element space

$$\begin{aligned} \boldsymbol{V}_h &:= \{\mathbf{v} \in H^1(\Omega, \mathbb{R}^d) \mid \mathbf{v}|_K \in [P_2(K)]^d, \ K \in \mathcal{T}_h\}, \\ Q_h &:= \{q \in L^2(\Omega, \mathbb{R}) \mid \int\limits_\Omega q \, d\mathbf{x} = 0, \ q|_K \in P_1(K), \ K \in \mathcal{T}_h\}. \end{aligned}$$

Let $\boldsymbol{V}_h^D := \{\mathbf{v} \in \boldsymbol{V}_h \mid \mathbf{v}|_{\partial\Omega} = \mathbf{u}_D^h\}$, where $\mathbf{u}_D^h$ is the a suitable approximation of the Dirichlet boundary condition $\mathbf{u}_D$ on the boundary edges/faces of $\mathcal{T}_h$. For the solution of (13), find $(\mathbf{u}_h, p_h) \in \boldsymbol{V}_h^D \times Q_h$ such that

$$\begin{aligned} ((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \boldsymbol{v}_h) - (p_h, \nabla \cdot \mathbf{v}_h) + (\mu \nabla \mathbf{u}_h, \nabla \mathbf{v}_h) + (\alpha(\overline{\chi}_h)\mathbf{u}_h, \mathbf{v}_h) = 0, &\quad \forall\, \mathbf{v}_h \in \boldsymbol{V}_h^0, \\ (\nabla \cdot \mathbf{u}_h, q_h) = 0, &\quad \forall\, q_h \in Q_h. \end{aligned}$$

and $(\tilde{\mathbf{u}}_h, \tilde{p}_h) \in \boldsymbol{V}_h^0 \times Q_h$ such that

$$\begin{aligned} -((\mathbf{u}_h \cdot \nabla)\tilde{\mathbf{u}}_h, \boldsymbol{v}_h) + ((\nabla \mathbf{u}_h)^T \tilde{\mathbf{u}}_h, \mathbf{v}_h) - (\tilde{p}_h, \nabla \cdot \mathbf{v}_h) + (\mu \nabla \tilde{\mathbf{u}}_h, \nabla \mathbf{v}_h) + (\alpha(\overline{\chi}_h)\tilde{\mathbf{u}}_h, \mathbf{v}_h) \\ = ((\mathbf{u}_h \cdot \nabla)\mathbf{u}_h, \boldsymbol{v}_h), \qquad \forall\, \mathbf{v}_h \in \boldsymbol{V}_h^0, \\ (\nabla \cdot \tilde{\mathbf{u}}_h, q_h) = 0, \qquad \forall\, q_h \in Q_h, \end{aligned}$$

where $\boldsymbol{V}_h^0 = \boldsymbol{V}_h \cap H_0^1(\Omega)$. All above systems are solved by standard Newton's iteration and each iteration is solved by the generalized minimal residual method (GMRES).

We also note that the above bilinear form can be straightforwardly extended to the problem both with Dirichlet boundary Γ_D and Neumann boundary Γ_N, where $\Gamma_D \cap \Gamma_N = \emptyset$, $\Gamma_D \cup \Gamma_N = \partial\Omega$, and $(\mu \nabla \mathbf{u} - p\boldsymbol{I}) \cdot \boldsymbol{n}|_{\Gamma_N} = \boldsymbol{g}$.

When $\mathbf{u}_h$ and $\tilde{\mathbf{u}}_h$ are obtained, we can use the FFT to compute ϕ^h on each node of $\mathcal{N}_h$ as follows:

$$\phi^h = \gamma \sqrt{\frac{\pi}{\tau}} G_\tau * (1 - 2\overline{\chi}^h) + \frac{\bar{\alpha}}{2} G_\tau * (|\mathbf{u}_h|^2 + 2\mathbf{u}_h \cdot \tilde{\mathbf{u}}_h)$$

Following Algorithm 1, we can now use ϕ^h to update the indicator function χ_h by the strategy presented in Algorithm 1.

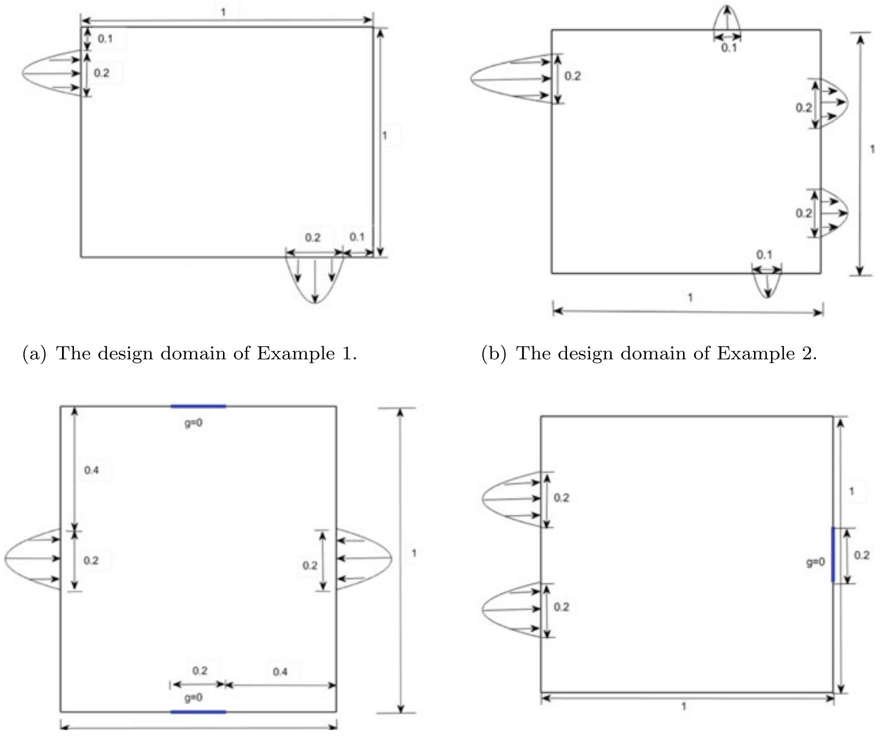
Remark 3.1 Similar to the adaptive in time strategy used in [9], we can modify Algorithm 1 into an adaptive algorithm by adjusting τ during the iterations. We set a threshold value τ_t and a given tolerance e_t, if $e_\chi^k \le e_t$, let $\tau_{\text{new}} = \eta\tau$ with $\eta \in (0, 1)$ and update $\tau := \tau_{\text{new}}$ in the next iteration unless $\tau \le \tau_t$. Otherwise, τ will not be updated, and the iteration will continue with the same τ.

4 Numerical Experiments

In this section, we perform extensive numerical examples to demonstrate the efficiency of our new algorithm with an adaptive strategy for the choice of τ. We choose $\eta = 0.5$ in the update of τ. If no confusion is possible, we still denote by τ as its initialization in the following. Also, we denote the Reynolds number by $Re = \frac{1}{\mu}$.

4.1 Two Dimensional Results

In this section, we test the performance of the proposed algorithm on two dimensional problems on several different design domains as displayed in Fig. 1. For most examples in this section, we assume that the Dirichlet boundary condition with a parabolic profile and the magnitude of the velocity are set as $|\mathbf{u}_D| = \overline{g}(1 - 4(\frac{t-a}{l})^2)$



(a) The design domain of Example 1.

(b) The design domain of Example 2.

(c) The design domain of Example 3.

(d) The design domain of Example 4.

Fig. 1 Design domains of two dimensional examples

with $t \in [a - \frac{l}{2}, a + \frac{l}{2}]$, where l is the length of the section of the boundary at which the inflow/outflow velocity is imposed, and $\bar{g}$ is the prescribed velocity at the midpoint a of the flow profile. The directions of the inflow/outflow velocity are illustrated separately in the design domain in each example.

Example 1 In this example, we consider the design of a bend, which has been tested by the level set method in [10, 14, 19]. The design domain is presented in Fig. 1a. Let $\bar{g}$ be 1 both in inlet and outlet, and we set the fluid fraction as $\beta = 0.08\pi$. Here, we use our algorithm to obtain the optimal design result on a 128×128 grid. We assume the initial distribution $\chi = 0$ in the whole domain, and set the parameter $\bar{\alpha} = 1.5\mu \times 10^4$ through this example.

The boundary conditions in this example are slightly different with [10, 19], but are same as that in [14]. Based on the 128×128 grid, firstly, we test the example for different Reynolds numbers, in which the other parameters are set as $\tau = 0.001$ and $\gamma = 0.0001$. The optimal design results together with the velocity field and the energy decaying curve are displayed in Fig. 2 for the cases of $\text{Re} = 10$, 100 and 1000,

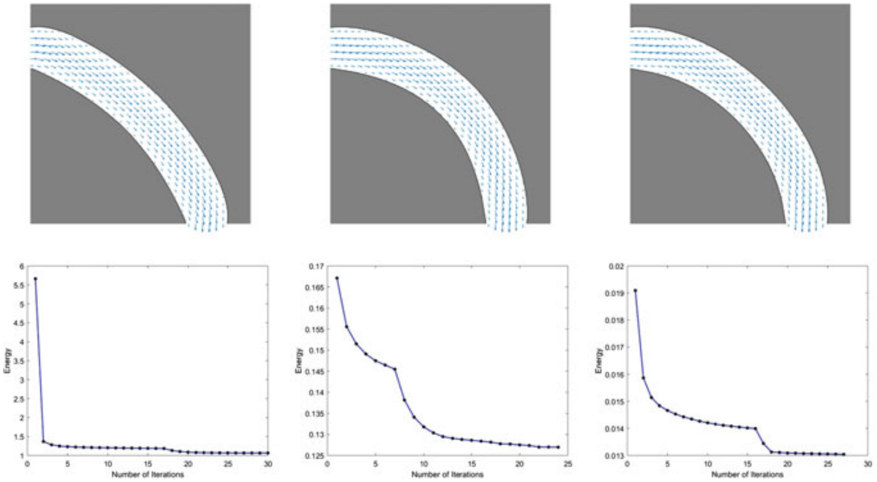


Fig. 2 (Example 1) Left to right: Optimal results and the corresponding energy decaying curve for the cases of Re $= 10, 100$, and 1000. The parameters are set as $\tau = 0.001$, $\gamma = 0.0001$ and $\bar{\alpha} = 1.5\mu \times 10^4$

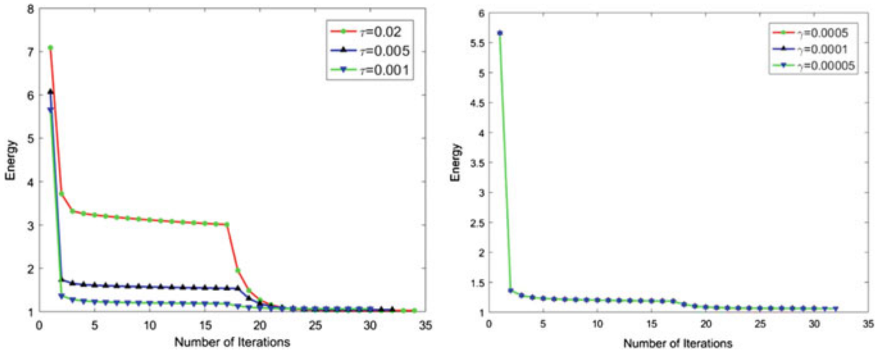


Fig. 3 (Example 1) Plots of energy curves for $\bar{\alpha} = 1.5\mu \times 10^4$ and Re $= 10$. Left: For fixed $\gamma = 0.0001$, energy curves for the cases of $\tau = 0.02, 0.005, 0.001$. Right: For fixed $\tau = 0.001$, energy curves for the cases of $\gamma = 0.0005, 0.0001, 0.00005$

separately. It was mentioned in [46] that the radius of curvature of the fluid domain is decreased as the Reynolds number is increased. This phenomenon can also be observed in Fig. 2, and the optimal results are consistent with those obtained by the level set methods in [10, 14, 19].

Furthermore, we numerically check the sensitivity of τ and γ on the energy decaying properties. In Fig. 3, we displayed the energy decaying curves for different choices of τ and γ with fixed Re $= 10$. We observe that the energy converges to almost the same value. In addition, the final design results we obtained are also identical to the left one in Fig. 2.

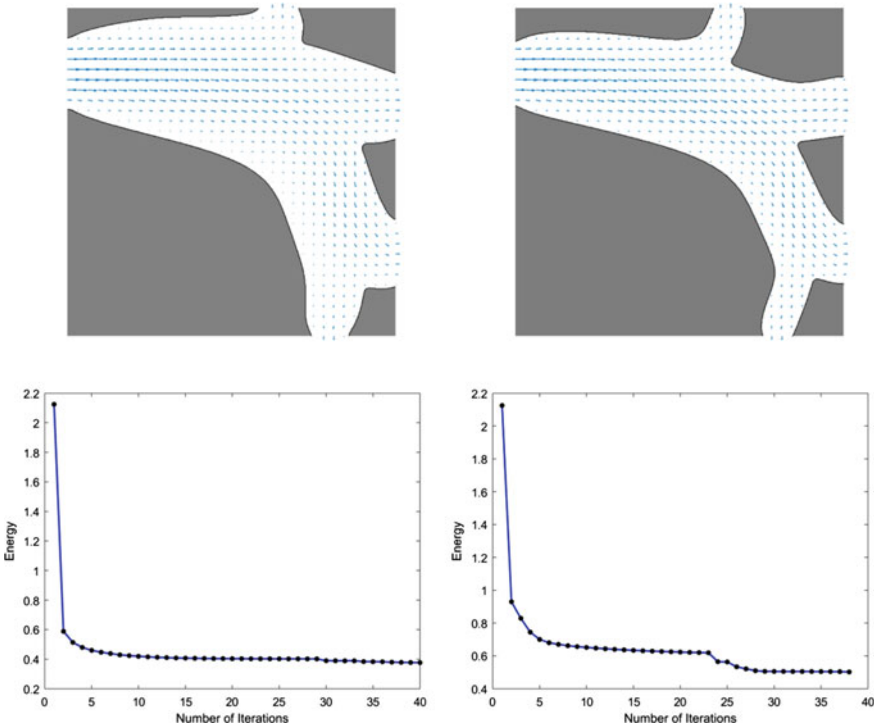


Fig. 4 (Example 2) Left to right: Optimal results and energy curves for $\beta = 0.5$ and $\beta = 0.4$

Example 2 We test the example presented in Fig. 1b which has one parabolic inlet and four parabolic outlets. We assume $\bar{g} = 3$, $l = 0.2$ and $a = 0.8$ on the inlet boundary $x = 0$. For the four outlets, we let $(\bar{g}, l, a) = (1, 0.1, 0.8)$, $(1, 0.1, 0.65)$, $(1, 0.2, 0.7)$ and $(1, 0.2, 0.25)$ on $y = 0$, $y = 1$, $x = 1$ and $x = 1$, respectively. This example has been tested by the phase field method in [18] with the same boundary conditions. Here, we use our algorithm to obtain the final optimal result on a 256×256 grid. Throughout this example, we set $\tau = 0.001$, $\gamma = 0.01$, $\bar{\alpha} = 1.5\mu \times 10^4$ and $\mathrm{Re} = 10$.

For the initial distribution $\chi = 1 - \chi_{\{(x,y):x\in(0,1),y\in(\frac{1}{6},\frac{5}{6})\}}$, we test this example for different fluid fractions β. For the left graph of Fig. 4 with $\beta = 0.5$, we obtain the optimal result after 40 iterations. For the the right graph of of Fig. 4 with $\beta = 0.4$, the optimal result is obtained after 38 iterations. We find that the final result in Fig. 4 has a treelike structure which is consistent with that obtained using the phase field method in [18]. The energy decaying curves for different fluid fractions β are also displayed in Fig. 4.

Example 3 In this example, we consider the minimization of the power dissipation in a four terminal device. We set $\bar{g} = 1$ for the two inflows and homogeneous Neumann boundaries on parts of the top and bottom boundaries with centers [0.5, 0] and

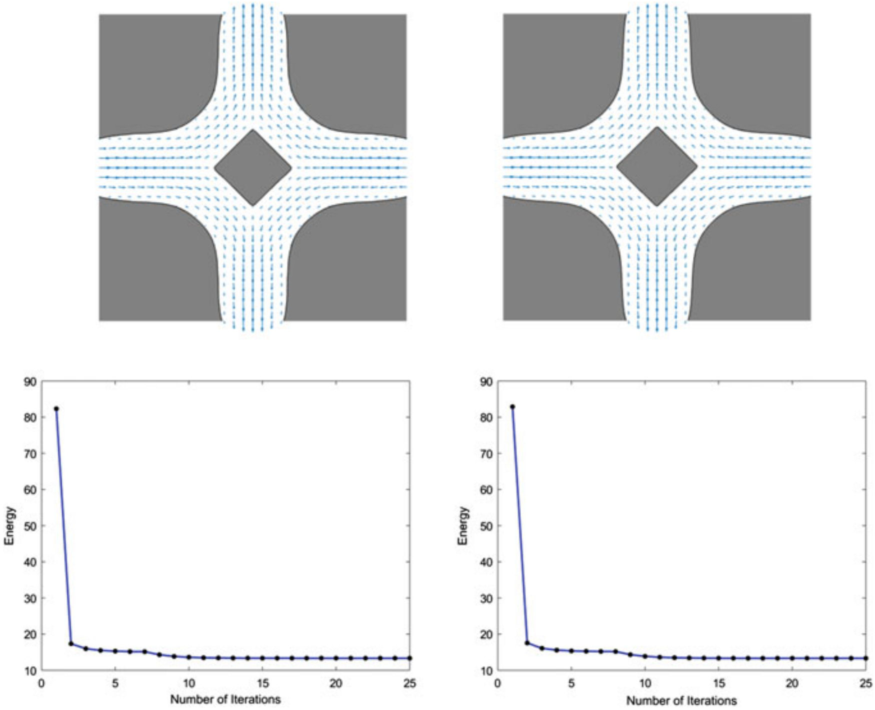


Fig. 5 (Example 3) Left to right: Optimal results and energy curves on a 128 × 128 grid and 256 × 256 grid. The parameters are set as $\tau = 0.001$, $\gamma = 0.0001$, $\bar{\alpha} = 2.5\mu \times 10^4$ and $\mathrm{Re} = 1$

[0.5, 1] (see Fig. 1c). The fluid fraction is defined as $\beta = 0.4$. Here, we utilize our algorithm to achieve the optimal configurations on 128 × 128 and 256 × 256 grids.

We test the case for $\tau = 0.001$, $\gamma = 0.0001$, $\bar{\alpha} = 2.5\mu \times 10^4$ and $\mathrm{Re} = 1$ on 128 × 128 and 256 × 256 grids. The initial distribution is set as $\chi = 1 - \chi_{\{(x,y):x\in(0,1),y\in(\frac{1}{3},\frac{2}{3})\}}$. In Fig. 5, we observe that the final optimal configuration is consistent with the result obtained using the level set method in [10]. And the final results for different grids are almost the same, which indicates that our algorithm is independent on grid for this example. Furthermore, the energy decaying property can be observed in Fig. 5.

Example 4 In this example, we consider a three terminal device on the design domain as displayed in Fig. 1d. We set $\bar{g} = 1$ on the two inflows and the homogeneous Neumann boundary condition on the outflow. The fluid fraction is set as $\beta = 0.3$ and we test this example on a 128 × 128 grid for $\tau = 0.0005$, $\gamma = 0.0002$ and $\bar{\alpha} = 1.5\mu \times 10^4$.

In this example, we study the relation of optimal configurations on different choices of Reynolds numbers. Based on the initial $\chi = 1 - \chi_{\{(x,y):x\in(0,1),y\in(\frac{1}{5},\frac{4}{5})\}}$,

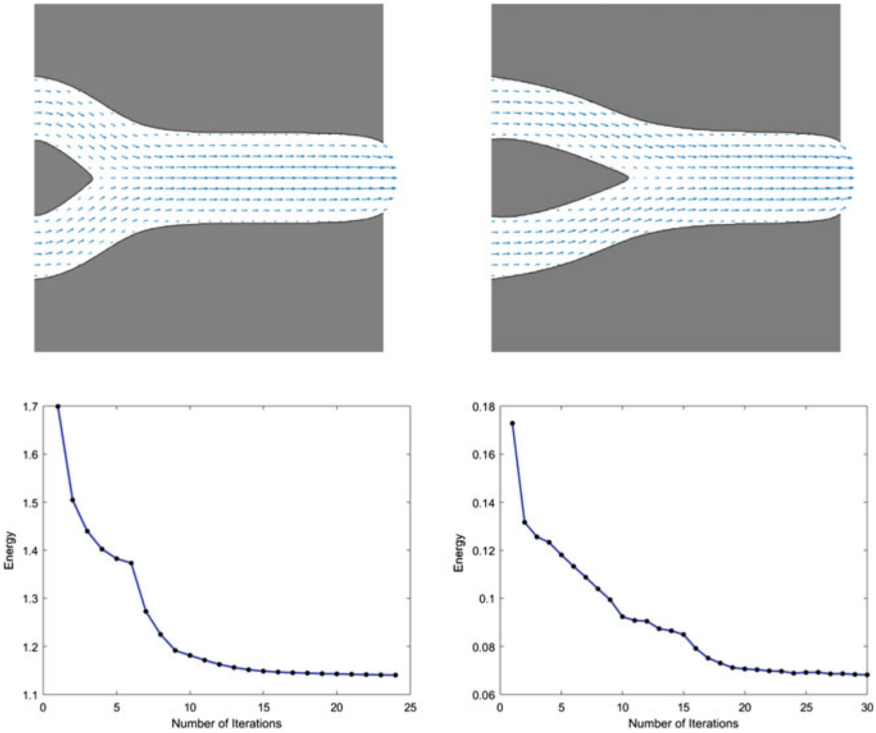


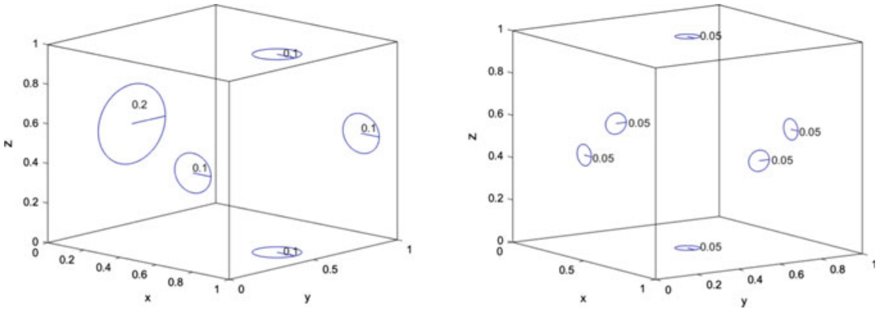
Fig. 6 (Example 4) Left to right: Optimal configurations and energy decaying curves for Re = 20 and 500

the final optimal design results with the velocity fields for Re = 20, and 500 are displayed in Fig. 6. We observe that the configuration gradually separates from each other as the Reynolds number increases. The energy decaying curves are also displayed and the iteration converges in about 20 steps for Re = 20 and 25 steps for Re = 500, respectively.

4.2 Three Dimensional Results

In this section, we show the performance of the algorithm on several three dimensional problems for different design domains in Fig. 7. In the following examples, the magnitude of the velocity for the Dirichlet boundary condition on a slice is set as

$$|\mathbf{u}_D| = \bar{g}\Big(1 - \frac{(s_1 - a)^2 + (s_2 - b)^2}{l^2}\Big),$$



(a) The design domain of Example 5.

(b) The design domain of Example 6.

Fig. 7 Design domains of three dimensional examples

where $\bar{g}$ is the prescribed velocity at the center (a, b) of a circle in which the inflow/outflow velocity is imposed, l is the radius of the circle, (s_1, s_2) are Cartesian coordinates on the slice.

Example 5 In the example, we consider the multi-outlet problem in Fig. 7a. For the inflow, we set $\bar{g} = 1$, $l = 0.2$, and $(a, b) = (\frac{1}{2}, \frac{1}{2})$ on $x = 0$ plane. For the outflow, we set $l = 0.1$, $\bar{g} = 1$, and $(a, b) = (0.8, 0.5)$, $(0.8, 0.5)$, $(0.8, 0.5)$, and $(0.8, 0.5)$ on $y = 0$, $y = 1$, $z = 0$, and $z = 1$ planes respectively. Throughout this example, we choose the initial distribution with fluid domain in a region of $\{(x, y, z) : x \in (0, 1), y \in (0, 1), z \in (\frac{1}{3}, \frac{2}{3})\}$, and set $\beta = 0.2$, $\alpha = 2.5\mu \times 10^4$ and $\mathrm{Re} = 20$.

We first test the case for $\tau = 0.005$ and $\gamma = 0.0001$ on $32 \times 32 \times 32$ and $85 \times 85 \times 85$ grids. The optimal results in the left graphs of Fig. 8 are consistent with those obtained using the level set method in [10]. In addition, from the energy decaying curves in Fig. 8, we observe that the iteration converges in about 20 steps and 30 steps on coarse and fine grids respectively. In Fig. 9, we displayed the slices on $32 \times 32 \times 32$ grids on $z = 0.5$ and $y = 0.5$ planes.

Next, we compute the result for different τ and γ on the $32 \times 32 \times 32$ grid. The energy curves for $\gamma = 0.0001$ and $\tau = 0.01, 0.005, 0.001$ are displayed in the left graph of Fig. 10, and the energy curves for $\tau = 0.005$ and $\gamma = 0.001, 0.0005, 0.0001$ are displayed in the right graph of Fig. 10. We observe that the energy converges to almost the same value for different γ and τ.

Example 6 Here, we consider an example with two inlets and four outlets. The design domain is defined in Fig. 7a. For the two inflows, let $\bar{g} = 2$, $l = 0.05$ and $(a, b) = (0.5, 0.5)$ on $x = 0$ and $x = 1$ planes respectively. For the four outflows, we set $\bar{g} = 1$, $l = 0.05$ and $(a, b) = (0.5, 0.5)$ on $y = 0$, $y = 1$, $z = 0$ and $z = 1$ planes respectively. In the example, we use our algorithm to obtain the final optimal result for $\tau = 0.001$, $\gamma = 0.0001$, $\bar{\alpha} = 2.5\mathrm{\mu} \times 10^4$ and $\mathrm{Re} = 1$. The initial distribution of fluid region is set as $\{(x, y, z) : x \in (0, 1), y \in (0, 1), z \in (\frac{1}{6}, \frac{5}{6})\}$.

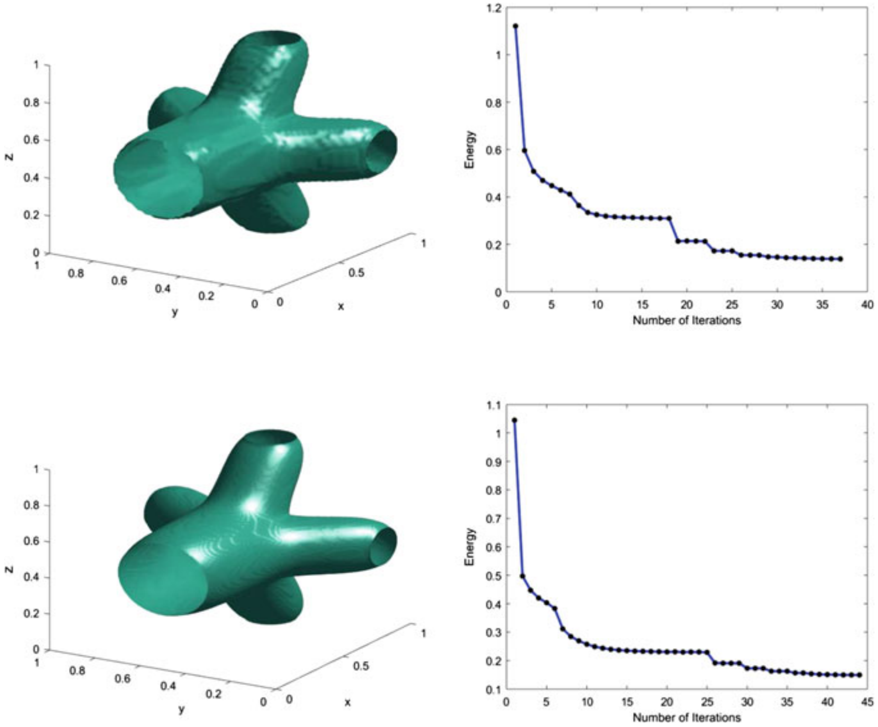


Fig. 8 (Example 5) Left to right: Optimal configurations on different grids (top: 32 × 32 × 32, bottom: 85 × 85 × 85) and energy curves. The parameters are set as $\tau = 0.005$, $\gamma = 0.0001$, $\bar{\alpha} = 2.5\mu \times 10^4$ and Re $= 20$

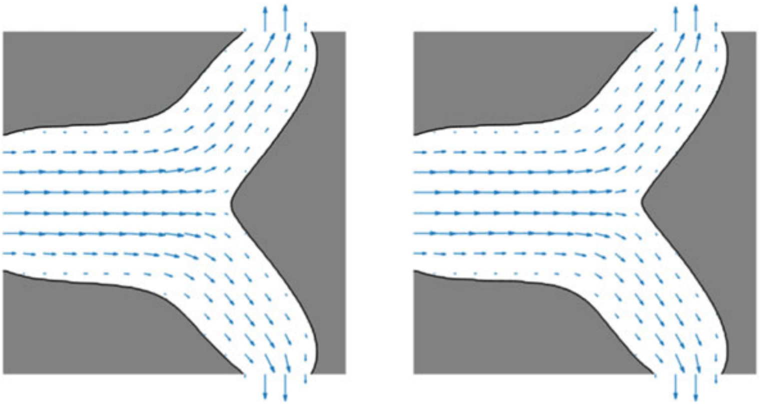


Fig. 9 (Example 5) The slices on the 85 × 85 × 85 grid for $\tau = 0.005$, $\gamma = 0.0001$, $\bar{\alpha} = 2.5\mu \times 10^4$ and Re $= 20$. Left: The slice on $z = 0.5$ plane. Right: The slice on $y = 0.5$ plane

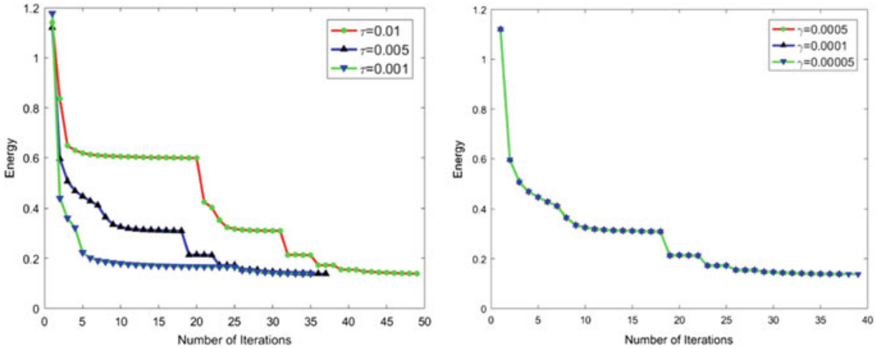


Fig. 10 (Example 5) Plots of energy curves for $\bar{\alpha} = 2.5\mu \times 10^4$ and $\mathrm{Re} = 20$. Left: For fixed $\gamma = 0.0001$, energy curves for the cases of $\tau = 0.01, 0.005, 0.001$. Right: For fixed $\tau = 0.005$, energy curves for the cases of $\gamma = 0.0005, 0.0001, 0.00005$

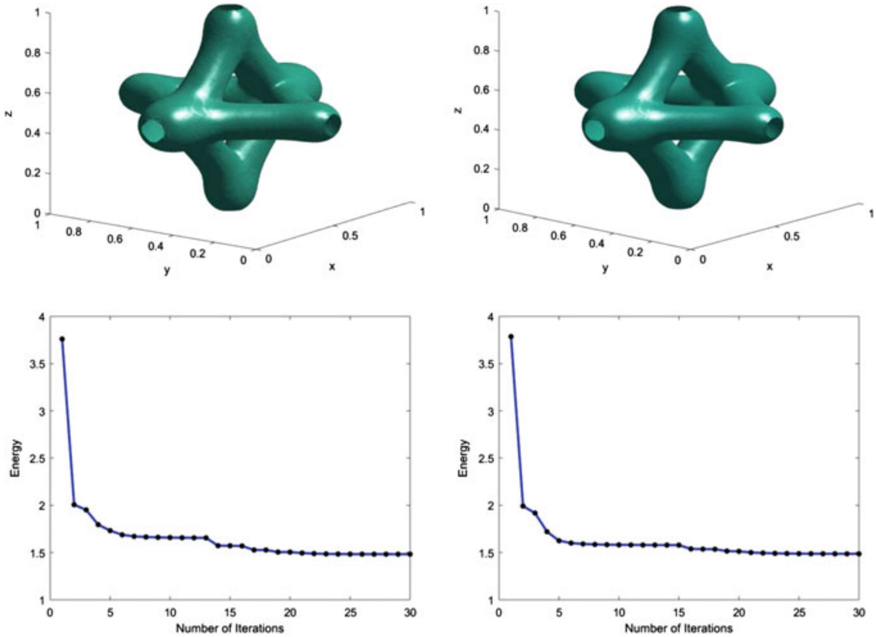


Fig. 11 (Example 6) Left to right: Optimal configurations on the different grids (top: $64 \times 64 \times 64$, bottom: $90 \times 90 \times 90$) and energy decaying curves. The fluid fraction is $\beta = 0.18$

For the fluid fraction $\beta = 0.1$, we design optimal configurations on $64 \times 64 \times 64$ and $90 \times 90 \times 90$ grids. The final results for the coarse and fine grids with corresponding energy decaying curves are displayed in Fig. 11. We observe that the interface is smoother on the fine mesh and the iteration converges in 25 and 30 steps for the coarse and fine grids respectively.

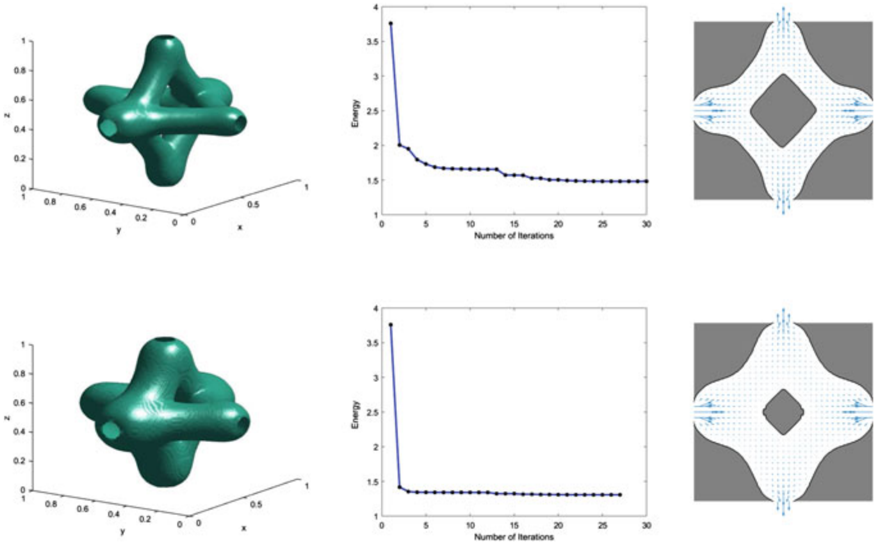


Fig. 12 (Example 6) Left to right: Optimal configurations for different β (top: $\beta = 0.1$, bottom: $\beta = 0.18$), energy decaying curves, and slices on $y = 0.5$ plane

Based on the $64 \times 64 \times 64$ grid, we check the dependency of the results on the choice of β. In Fig. 12, we displayed the results, energy decaying curves, and slices on the $y = 0.5$ plane for the optimal shape obtained by $\beta = 0.1$ and 0.18. The iteration converges in about 25 steps and 20 steps for $\beta = 0.1$ and 0.18. From Fig. 12, we can observe that the solid domain in the center shrinks as β increases.

5 Conclusion

In this paper, we present an efficient threshold dynamics method for topology optimization for Navier–Stokes flow. This is an extension of our previous work [9] to the case of fluids in Navier–Stokes flow. We aim to minimize a total energy functional that consists of the potential power and the perimeter approximated by nonlocal energy. Different from the algorithm in [9], during the iterations of the algorithm, we need to solve not only the Brinkman equation but also an adjoint problem by the mixed finite element method. Then the indicator functions of fluid-solid regions are updated by a thresholding step which is based on the convolutions evaluated by the FFT. A simple adaptive time strategy is used to accelerate the convergence of the algorithm. Some numerical examples are presented to verify the efficiency of the new algorithm, and the total energy decaying property of the proposed algorithm can be observed numerically. The proposed algorithm is simple and easy to implement. For the numerical experiments that we have performed, the proposed algorithm always finds an optimal shape and the numerical results are relatively insensitive to the initial guesses and parameters.

Acknowledgements This research was supported in part by the Hong Kong Research Grants Council (GRF grants 16324416, 16303318 and 16305819). The work of H. Leng was supported by the NSF of China (Grant No. 12001209). The work of D. Wang was supported by the University Development Fund of the Chinese University of Hong Kong, Shenzhen (UDF 01001803). The work of H. Chen was supported by the NSF of China (Grant No. 11771363, 91630204, 51661135011), the Fundamental Research Funds for the Central Universities (Grant No. 20720180003).

References

1. Andreasen, C.S., Gersborg, A.R., Sigmund, O.: Topology optimization of microfluidic mixers. Int. J. Numer. Methods Fluids **61**, 498–513 (2009)
2. Amstutz, S.: The topological asymptotic for the Navier-Stokes equations, ESAIM: control. Optim. Calculus Variations **11**(3), 401–425 (2005)
3. Bendsøe, M.P., Kikuchi, N.: Generating optimal topologies in structural design using a homogenization method. Comput. Methods Appl. Mech. Eng. **71**, 197–224 (1988)
4. Bendsøe, M.P., Sigmund, O.: Optimization of Structural Topology, Shape, and Material. Springer, Berlin (1995)
5. Bendsøe, M.P., Sigmund, O.: Material interpolations in topology optimization. Arch. Appl. Mech. **69**, 635–654 (1999)
6. Bendsøe, M.P., Sigmund, O.: Topology Optimization: Theory, Methods and Applications. Springer, Berlin (2003)
7. Borrvall, T., Petersson, J.: Topology optimization of fluids in Stokes flow. Int. J. Numer. Methods Fluids **41**, 77–107 (2003)
8. Challis, V.J., Guest, J.K.: Level set topology optimization of fluids in Stokes flow. Int. J. Numer. Meth. Engrg. **79**, 1284–1308 (2009)
9. Chen, H., Leng, H., Wang, D., Wang, X.P.: An efficient threshold dynamics method for topology optimization for fluids, arXiv preprint, arXiv:1812.09437 (2018)
10. Dai, X., Zhang, C., Zhang, Y., Gulliksson, M.: Topology optimization of steady Navier-Stokes flow via a piecewise constant level set method. Struct. Multidisc. Optim. **57**, 2193–2203 (2018)
11. Dbouk, T.: A review about the engineering design of optimal heat transfer systems using topology optimization. Appl. Thermal Eng. **112**, 841–854 (2017)
12. Deng, Y., Liu, Z., Zhang, P., Liu, Y., Wu, Y.: Topology optimization of unsteady incompressible Navier-Stokes flows. J. Comput. Phys. **230**, 6688–6708 (2011)
13. Dilgen, C., Sumer, B.D., David, R.F., Sigmund, O., Boyan, S.L.: Topology optimization of turbulent flows. Computer Methods Appl. Mech. Eng. **331**, 363–393 (2018)
14. Duan, X., Ma, Y., Zhang, R.: Shape-topology optimization for Navier-Stokes problem using variational level set method. J. Comput. Appl. Math. **222**, 487–499 (2008)
15. Duhring, M.B., Jensen, J.S., Sigmund, O.: Acoustic design by topology optimization. J. Sound Vibr. **317**, 557–575 (2008)
16. Esedoglu, S., Tsai, Y.H.R.: Threshold dynamics for the piecewise constant Mumford-Shah functional. J. Comput. Phys. **211**, 367–384 (2006)
17. Esedoglu, S., Otto, F.: Threshold dynamics for networks with arbitrary surface tensions. Comm. Pure Appl. Math. **68**, 808–864 (2015)
18. Garcke, H., Hecht, C., Hinze, M., Kahle, C.: Numerical approximation of phase field-based shape and topology optimization for fluids. SIAM J. Sci. Comput. **37**, A1846–A1871 (2015)
19. Gersborg-Hansen, A., Sigmund, O., Haber, R.B.: Topology optimization of channel flow problems. Struct. Multidisc. Optim. **30**, 181–192 (2005)

20. Gersborg-Hansen, A., Sigmund, O., Haber, R.: Topology optimization of channel flow problems. Struct. Multidiscip. Optim. **30**, 181–192 (2005)
21. Guest, J.K., Prévost, J.H.: Topology optimization of creeping fluid flows using a Darcy-Stokes finite element. Int. J. Numer. Meth. Engrg. **66**, 461–484 (2006)
22. Guillaume, P., Idris, K.S.: Topological sensitivity and shape optimization for the Stokes equations. SIAM J. Control Optim. **43**(1), 1–31 (2004)
23. Merriman, B., Bence, J.K., Osher, S.: Diffusion generated motion by mean curvature. UCLA CAM Report 92-18 (1992)
24. Van Oevelen, T., Baelmans, M.: Numerical topology optimization of heat sinks. In: Proceedings of the 15th International Heat Transfer Conference, pp. 5985–5999 (2014)
25. Okkels, F., Olesen, L.H., Bruus, H.: Application of topology optimization in the design of micro and nanofluidic systems. NSTI-Nanotech, pp. 575–578 (2005)
26. Okkels, F., Bruus, H.: Scaling behavior of optimally structured catalytic microfluidic reactors. Phys. Rev. E **75**, 1–4 (2007)
27. Olesen, L.H., Okkels, F., Bruus, H.: A high-level programming-language implementation of topology optimization applied to steady-state Navier-Stokes flow. Int. J. Numer. Meth. Eng. **65**, 975–1001 (2006)
28. Osting, B., Wang, D.: A diffusion generated method for orthogonal matrix-valued fields. Math. Comp. **89**, 515–550 (2020)
29. Osting, B., Wang, D.: Diffusion generated methods for denoising target-valued images. AIMS Inverse Probl. Imag. **14**(2), 205–232 (2020)
30. Pingen, G., Maute, K.: Optimal design for non-Newtonian flows using a topology optimization approach. Comput. Math. Appl. **59**, 2340–2350 (2010)
31. Rozvany, G.I.N.: Aims, scope, methods, history and unified terminology of computer-aided topology optimization in structural mechanics. Struct. Multidisc. Optim. **21**, 90–108 (2001)
32. Saxena, A.: Topology design of large displacement compliant mechanisms with multiple materials and multiple output ports. Struct. Multidisc. Optim. **30**, 477–490 (2005)
33. Sigmund, O.: On the design of compliant mechanisms using topology optimization. Mech. Struct. Mach. **25**, 495–526 (1997)
34. Sigmund, O., Hougaard, K.G.: Geometric properties of optimal photonic crystals. Phys. Rev. Lett. **100**, 153904 (2008)
35. Svanberg, K.: The method of moving asymptotes-a new method for structural optimization. Int. J. Numer. Meth. Engrg. **24**, 359–373 (1987)
36. Villanueva, C.H., Maute, K.: CutFEM topology optimization of 3D laminar incompressible flow problems. Comput. Methods Appl. Mech. Eng. **320**, 444–473 (2017)
37. Wang, D., Li, H., Wei, X., Wang, X.-P.: An efficient iterative thresholding method for image segmentation. J. Comput. Phys. **350**, 657–667 (2017)
38. Wang, D., Wang, X.-P., Xu, X.: An improved threshold dynamics method for wetting dynamics. J. Comput. Phys. **392**, 291–310 (2019)
39. Wang, D., Wang, X.-P.: The iterative convolution-thresholding method (ICTM) for image segmentation, arXiv preprint arXiv:1904.10917(2019)
40. Wang, D., Osting, B., Wang, X.-P.: Interface dynamics for an Allen-Cahn-type equation governing a matrix-valued field. SIAM J. Multiscale Model. Sim. **17**(4), 1252–1273 (2019)
41. Wang, D., Osting, B.: A diffusion generated method for computing Dirichlet partitions. J. Comput. Appl. Math. **351**, 302–316 (2019)
42. Wang, D., Cherkaev, A., Osting, B.: Dynamics and stationary configurations of heterogeneous foams. PLOS ONE **14**(4) (2019)
43. Wiker, N., Klarbring, A., Borrvall, T.: Topology optimization of regions of Darcy and Stokes flow. Int. J. Numer. Meth. Eng. **69**, 1374–1404 (2007)
44. Xu, X., Wang, D., Wang, X.P.: An efficient threshold dynamics method for wetting on rough surfaces. J. Comput. Phys. **330**, 510–528 (2017)
45. Xu, X., Ying, W.: An adaptive threshold dynamics method for three-dimensional wetting on rough surfaces. Preprint (2019)

46. Yaji, K., Yamada, T., Yoshino, M., Matsumoto, T., Izui, K., Nishiwaki, S.: Topology optimization using the lattice Boltzmann method incorporating level set boundary expressions. J. Comput. Phys. **274**, 158–181 (2014)
47. Zhou, S., Li, Q.: A variational level set method for the topology optimization of steady-state Navier-Stokes flow. J. Comput. Phys. **227**, 10178–10195 (2008)

Dynamics of Complex Singularities of Nonlinear PDEs

Analysis and Computation

J. A. C. Weideman

Abstract Solutions to nonlinear evolution equations exhibit a wide range of interesting phenomena such as shocks, solitons, recurrence, and blow-up. As an aid to understanding some of these features, the solutions can be viewed as analytic functions of a complex space variable. The dynamics of poles and branch point singularities in the complex plane can often be associated with the aforementioned features of the solution. Some of the computational and analytical results in this area are surveyed here. This includes a first attempt at computing the poles in the famous Zabusky–Kruskal experiment that lead to the discovery of the soliton.

1 Introduction

Ever since Kruskal [22] remarked that soliton motion may be thought of as a "parade of poles," the study of complex pole dynamics in nonlinear wave equations has been an active research field. This paper is an overview of the field, using some of the well-known model problems, including the Korteweg–De Vries equation that prompted Kruskal's remark. The plan is to take these equations, some of them dissipative and others dispersive, and start them all with the same set of initial and boundary conditions. Using analysis where we can and numerical computation otherwise, we shall then track the evolution of the complex singularities. The singularity dynamics of the various equations will be contrasted, and also connected to the typical nonlinear features associated with these equations such as shock formation, soliton motion, finite time blow-up, and recurrence. Here, a particular interest is the entry of the singularities when the initial condition has no singularities in the finite complex plane.

J. A. C. Weideman (✉)
Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa
e-mail: weideman@sun.ac.za

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7_13

We consider equations of the form

$$u_t + uu_x = L(u), \quad t > 0, \; -\pi \le x < \pi, \tag{1}$$

and assume 2π-periodic solutions in the space variable, x. The linear operator on the right can be any one of

$$L(u) = \nu\, u_{xx} \qquad \text{(Burgers)}, \tag{2}$$

$$L(u) = -\nu\, u_{xxx} \quad \text{(Korteweg–De Vries)}, \tag{3}$$

$$L(u) = \nu\, H\{u_{xx}\} \quad \text{(Benjamin–Ono)}, \tag{4}$$

where ν is a nonnegative constant and H denotes the periodic Hilbert transform, defined below. As initial condition we consider

$$u(x, 0) = -\sin(x), \tag{5}$$

the particular form of which allows us to make connections to several works of historical interest, namely papers by Cole [10], Hopf [21], Platzman [27], and Zabusky and Kruskal [39].

The numerical procedure we follow is similar to the one proposed in [35]. The first step involves a Fourier spectral method in space and a numerical integrator in time to compute the solution on $[-\pi, \pi] \times [0, T]$. The second step is to continue the solution at any time t in $[0, T]$ into the complex x-plane. For the continuation we use a Fourier–Padé method, although other possibilities are considered as well.

In order to identify and display poles and branch points in the complex plane, we shall plot what is called the "analytical landscape" in [34]. With the solution $f(z)$ expressed in polar form $re^{i\theta}$, the software of [34] can be used to generate a 3D plot in which the horizontal axes represent the real and imaginary components of $z = x + iy$, the height represents the modulus r, and colour represents the phase $e^{i\theta}$. The two examples in Fig. 1 should clarify this visualization.

The outline of the paper is as follows: The inviscid Burgers equation and its viscous counterpart are discussed, respectively, in Sects. 2 and 3. Here, analysis provides the exact locations of the branch point singularities in the inviscid case and approximate locations of the poles in the case of small viscosity. For the other PDEs considered here, namely Benjamin-Ono (BO) in Sect. 4 and Korteweg-de Vries (KdV) in Sect. 5, analytical results are harder to come by and we resort to the numerical procedure mentioned above. The nonlinear Schrödinger equation (NLS) also makes an appearance in our discussion of recurrence in Sect. 6. In the final section we discuss the details of the numerical methods employed in the earlier sections.

Novel results presented here include the pole dynamics of the BO, KdV, and NLS equations. Related studies of KdV were undertaken in [7, 17], but these authors did not consider the Zabusky–Kruskal experiment which is our focus here. Pole behaviour in KdV and NLS was also discussed in the papers [11, 22] and [9, 23], respectively, but those analyses were based on cases where explicit solutions are

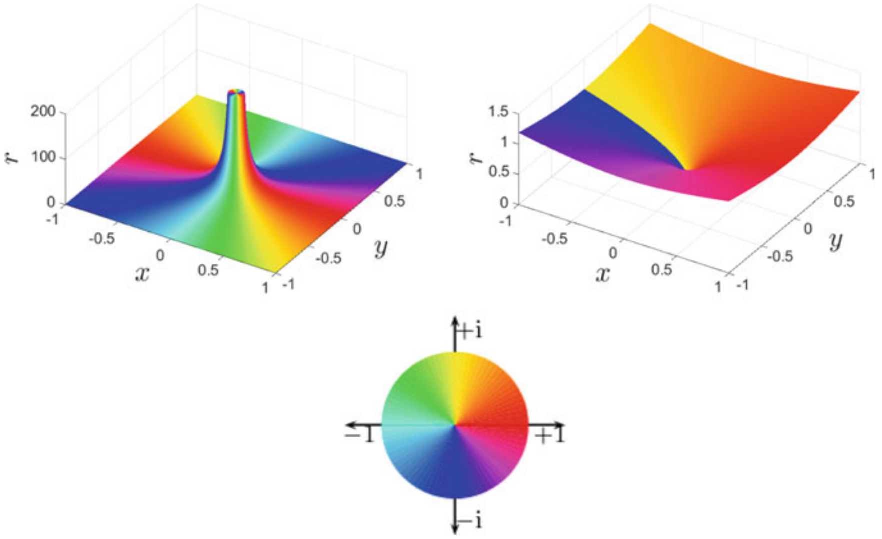


Fig. 1 Analytical landscapes of the functions $f(z) = 1/z^2$ (top left), and $f(z) = z^{1/2}$ (top right). The height represents the modulus and the colour represents the phase, as defined by the NIST standard colour wheel (bottom); see [13]. For details about the software used to produce these figures, see [34]

available. Moreover, in those papers the poles were already present in the initial condition. Here, our interest is in the situation where the singularities are "born" at infinity.

Although this paper focuses only on simple model equations such as (1)–(4), pole dynamics have been studied in more complex models, particularly in the water wave context. Among the many references are [3, 7, 14].

2 The Inviscid Burgers Equation

The inviscid Burgers equation, $u_t + uu_x = 0$, subject to the initial condition (5), develops a shock at $(x, t) = (0, 1)$, as can be verified by the method of characteristics. It also admits an explicit Fourier series solution [27]

$$u(x, t) = -2 \sum_{k=1}^{\infty} c_k(t) \sin(kx), \qquad c_k(t) := \frac{J_k(kt)}{kt}, \tag{6}$$

valid for $0 < t < 1$. The J_k are the Bessel functions of the first kind. This series is of limited use for numerical purposes, however, particularly for continuation into the complex plane. When truncated, it becomes an entire function and will not reveal

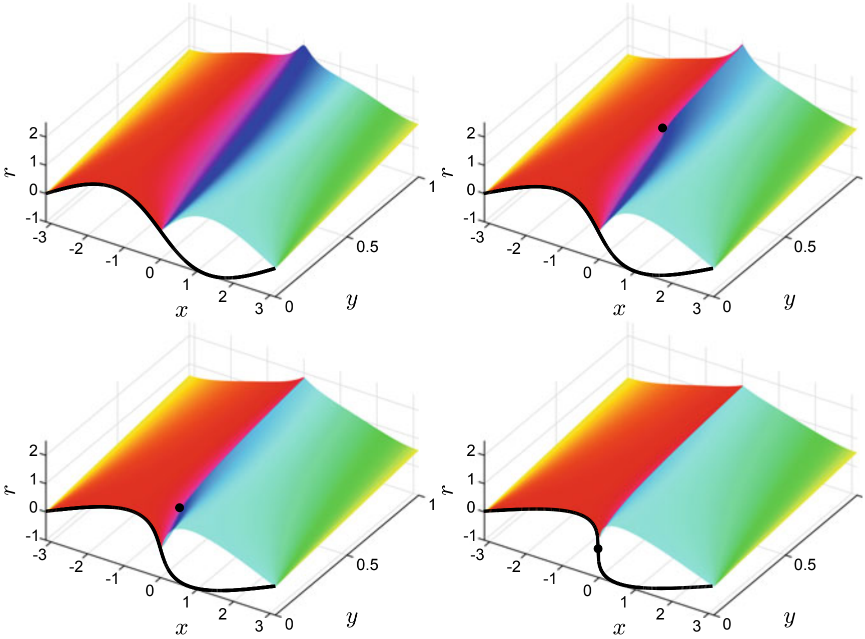


Fig. 2 Solution to the inviscid Burgers equation as computed by applying Newton iteration to the implicit solution formula (7). The four frames correspond to $t = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}$, and 1 (in the usual order). The thicker black curve is the real-valued solution on the real axis, displaying the typical steepening of the curve until the shock forms in the last frame. The solution in the upper half-plane is displayed in the format of Fig. 1. The solution in the lower half-plane is not shown because of symmetry. The black dot represents a branch point singularity that travels along the imaginary axis according to (9). By referring to the colour wheel of Fig. 1, one can see that on the imaginary axis, there is no jump in phase between the origin and the branch point (in some printed versions the abrupt change in phase may appear to be discontinuous but it is not.) From the branch point to $+i\infty$, however, there is a phase jump consistent with a singularity of quadratic type

much singularity information other than perhaps the location and type of the singularity nearest to the real axis [26, 32].

Instead, for numerical purposes we shall use the implicit solution formula

$$u = f(x - ut), \qquad f(x) = -\sin(x). \tag{7}$$

This transcendental equation can be solved by Newton iteration for values of x in the complex plane. One can start at a small time increment, say $t = \Delta t$, use $u = f(x)$ as initial guess, and iterate until convergence. Then t is incremented to $2\Delta t$, the initial guess is updated to the current solution, and the process is repeated. Figure 2 shows the corresponding solutions in the visualization format described in the introduction.

The figure shows one member of a conjugate pair of branch point singularities, born at $+i\infty$, which travels down the positive imaginary axis and meet its conjugate partner (not shown) at $(x, t) = (0, 1)$ when the shock occurs. This behaviour was first reported in [5, 6], where a cubic polynomial was used as initial condition (similar to the first two terms in the Taylor expansion of (5)). In the cubic case, eq. (7) can be solved explicitly by Cardano's formula, which enabled a complete description of the singularity dynamics as summarized in [5, 6, 28, 29]. In our case, the initial condition is trigonometric and therefore Cardano's formula is not applicable. It is nevertheless possible to find the singularity locations and their type explicitly.

The singularity location, say $z = z_s$, and the corresponding solution value, say $u = u_s$, are defined by the simultaneous equations

$$u_s = f(z_s - u_s t), \qquad 1 = -tf'(z_s - u_s t), \tag{8}$$

the latter equation representing the vanishing Jacobian of the mapping; see for example [26]. With $f(x)$ defined by (5), the solution is, for $0 < t < 1$,

$$z_s = \pm i\left(\sqrt{1-t^2} - \tanh^{-1}\sqrt{1-t^2}\right), \quad u_s = \pm i\, t^{-1}\sqrt{1-t^2}. \tag{9}$$

These formulas are consistent with the solution shown in Fig. 2. A graph of the singularity location as a function of time is shown as the dashed curve in Fig. 3 of the next section.

Further analysis shows that the singularity is of quadratic type, consistent with the phase colours in Fig. 2 and in agreement with the analysis of [5, 6, 28, 29] for the cubic initial condition. When $t = 1$, i.e., at the time the shock occurs, the singularity type changes from quadratic to cubic. The Riemann surface structure associated with this is discussed in [5, 6], in connection with the cubic initial condition.

3 The Viscous Burgers Equation

When viscosity is added, i.e., $\nu > 0$ in the Burgers equation (1)–(2), shock formation does not occur. In the complex plane interpretation this means the singularities do not reach the real axis. Moreover, they become strings of poles rather than the branch points observed in the previous section. The poles travel in conjugate pairs from $\pm i\,\infty$, with rapid approach towards the real axis, before turning around. They retrace their steps along the imaginary axes at a more leisurely pace, and eventually recede back to infinity, which ultimately leads to the zero steady state solution.[1]

Analogously to (6), the Burgers equation subject to the initial condition (5) has an explicit series solution, this time not a Fourier series but a ratio of two such series:

$$u(x,t) = -2\nu\frac{\theta_x}{\theta}, \quad \theta(x,t) := I_0\big(\frac{1}{2\nu}\big) + 2\sum_{k=1}^{\infty}(-1)^k I_k\big(\frac{1}{2\nu}\big)e^{-\nu k^2 t}\cos(kx). \tag{10}$$

[1] A movie of the pole dynamics of this solution and some of the other solutions in this paper can be found on the author's web page [36].

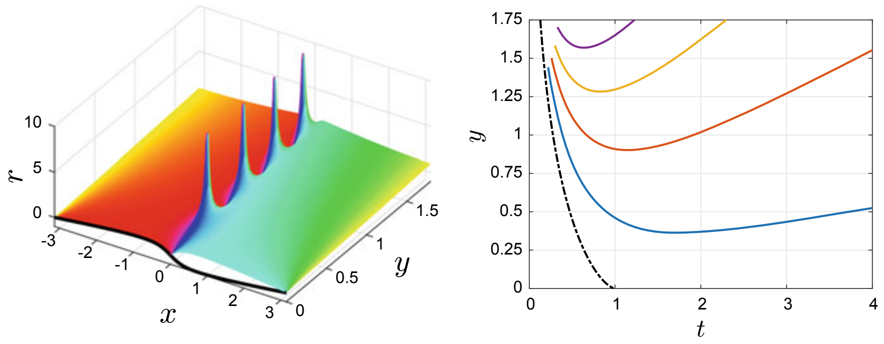


Fig. 3 Left: Solution of the viscous Burgers equation (2), with $\nu = 0.1$, $t = 1$, as computed from the series solution formula (10). Right: The locations on the positive imaginary axis of the first four poles as a function of time. The dash-dot curve is the location of the branch-point singularity when $\nu = 0$, as given by formula (9) (the pole curves approach the dash-dot curve asymptotically as $t \to 0^+$ but could not be computed reliably for small values of t because of ill-conditioning, hence the gaps)

The I_k are the modified Bessel functions of the first kind. This solution is derived from the famous Hopf–Cole transformation; in fact, the above series is a special case of one of the examples presented in the original paper of Cole [10]. Presumably the solutions (6) and (10) can be connected in the limit $\nu \to 0^+$, but we found no such reference in the literature.

The pole locations in Fig. 3 can be computed from the series solution (10). For asymptotic estimates, however, a better representation is the integral form [10, 21]:

$$u(x,t) = \frac{\int_{-\infty}^{\infty} \frac{x-s}{t} \exp\left(\frac{1}{2\nu} F(x,s,t)\right) ds}{\int_{-\infty}^{\infty} \exp\left(\frac{1}{2\nu} F(x,s,t)\right) ds}. \tag{11}$$

In the case of the initial condition (5) the function F is defined by

$$F(x,s,t) = 1 - \cos(s) - \frac{(x-s)^2}{2t}. \tag{12}$$

To estimate the pole locations in the inviscid Burgers equation one can analyze the denominator of the formula in (11). Looking for poles on the positive imaginary axis, we define, for $y > 0$,

$$D(y,t) = \int\limits_{-\infty}^{\infty} \exp\left(\frac{1}{2\nu} F(yi,s,t)\right) ds. \tag{13}$$

A saddle point method can be used to estimate this integral when $0 < \nu \ll 1$. We present an informal analysis here, focussed on an explanation of the situation shown in Fig. 3. A more comprehensive analysis (for the cubic initial condition) can be found in [28].

Figure 4 shows level curves of the real and imaginary parts of $F(iy, s, t)$ in the complex s-plane, with $y = 1$ and $t = 1$. The figure reveals three saddle points, two in the upper half-plane and one in the lower half-plane. The contour of integration in (13) is accordingly deformed into the upper half-plane, in order to pass through the two saddle points.

To estimate the saddle point contributions, we differentiate (13) with respect to s (and suppress the dependence on y and t),

$$F'(s) = \sin(s) - \frac{(s - yi)}{t}, \qquad F''(s) = \cos(s) - \frac{1}{t}. \tag{14}$$

The saddle points are defined by $F'(s) = 0$, i.e.,

$$s - yi - t\sin(s) = 0. \tag{15}$$

No explicit solution of this equation seems to exist, but it can be checked that for $t = 1$ and all $y > 0$ there is precisely one root on the negative imaginary axis, and two roots in the upper half-plane, symmetrically located with respect to the imaginary axis. The configuration shown in Fig. 4 can therefore be taken as representative of all $y > 0$, except that the saddle points coalesce at the origin as $y \to 0^+$.

We label the roots in the first and second quadrants as s_1 and s_2, respectively, with $s_2 = -\overline{s}_1$. The corresponding saddle point contributions are D_1 and D_2, where

$$D_j = 2\sqrt{\frac{\pi\nu}{|F''(s_j)|}}\exp\left(\frac{1}{2\nu}F(s_j) - \frac{1}{2}i(\theta_j \pm \pi)\right), \tag{16}$$

where the upper (resp., lower) sign choice refer to $j = 1$ (resp., $j = 2$). The quantities θ_j are defined by $F''(s_j) = |F''(s_j)|e^{i\theta_j}$.

The approximation to the denominator integral (13) is now given by $D \sim D_1 + D_2$ as $\nu \to 0^+$. After using the symmetry relationships between s_1 and s_2 noted above, as well as the fact that $|F''(s_1)| = |F''(s_2)|$, this becomes

$$D \sim 4\sqrt{\frac{\pi\nu}{|F''(s_1)|}}e^{\frac{1}{2\nu}\lambda_1}\sin\left(\frac{\mu_1}{2\nu} - \frac{1}{2}\theta_1\right), \quad F(s_1) := \lambda_1 + \mu_1 i. \tag{17}$$

In the second frame of Fig. 4 the graph of this function is shown as a function of y. In comparison with a high-accuracy quadrature approximation of the integral (13), the approximation (17) is seen to be quite accurate. The exception is for small values of y, because of the coalescence of the saddle points mentioned above.

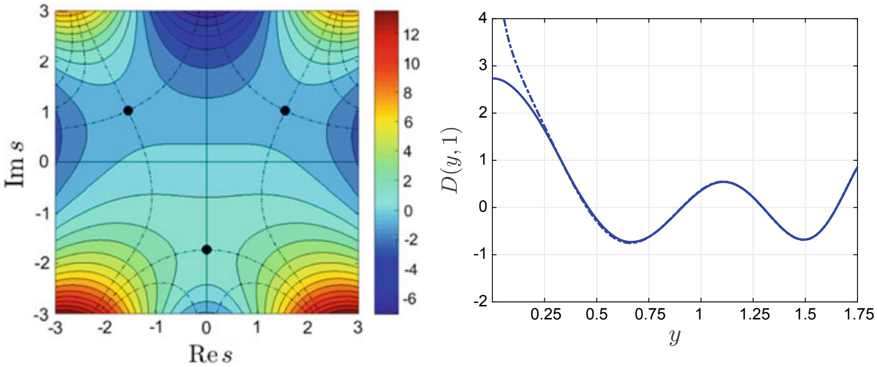


Fig. 4 Saddle point analysis for the viscous Burgers equation shown in Fig. 3. Left: The dots are saddle points of $F(yi, s, t)$, with $y = 1$, $t = 1$. The colour represents level curves of the real part of $F(yi, s, t)$, and the dash-dot curves are level curves of the imaginary part. For the saddle point analysis the path of integration in (13), i.e., the real line, is deformed into the dash-dot curve in the upper half-plane that defines the steepest descent direction. The main contributions to the integral come from the regions in the neighbourhood of the saddle points. Right: The function $D(y, 1)$, computed by numerical integration of (13) (solid curve), in comparison with the saddle point approximation (17) (dash-dot curve). The zeros of this function define the locations of the poles seen in Fig. 3

Table 1 Left: Pole locations on the positive imaginary axis for the solution shown in Fig. 3, i.e., $t = 1$ and $\nu = 0.1$. The 'exact' values were computed by numerical quadrature of (13) and root finding, both processes executed to high precision. The estimated values were computed by a numerical solution of the two equations (15) and (18). Right: Turning points of the poles, i.e., the coordinates of the local minima in the right frame of Fig. 3. This was computed by a numerical solution of the two equations (15) and (18) in combination with a minimization procedure with objective function y

k	Exact	Estimated
1	0.4589	0.4527
2	0.9090	0.9068
3	1.2964	1.2952
4	1.6505	1.6498

k	t	y
1	1.7221	0.3469
2	1.1612	0.8991
3	0.8302	1.2822
4	0.6373	1.5684

Approximate pole locations can be computed as the zeros of (17), i.e.,

$$\mu_1 - \nu\theta_1 = 2\nu k\pi, \quad k = 1, 2, \dots, \tag{18}$$

which is solved simultaneously with the saddle point equation (15). In Table 1 we compare this estimate with the actual pole locations.

The equations (15)–(18) can be used as basis for further analysis, both theoretical and numerical, of the pole locations. For example, by solving these equations numerically and simultaneously minimizing over y, the closest distance any particular pole gets to the real axis can be computed. These results are also summarized in Table 1.

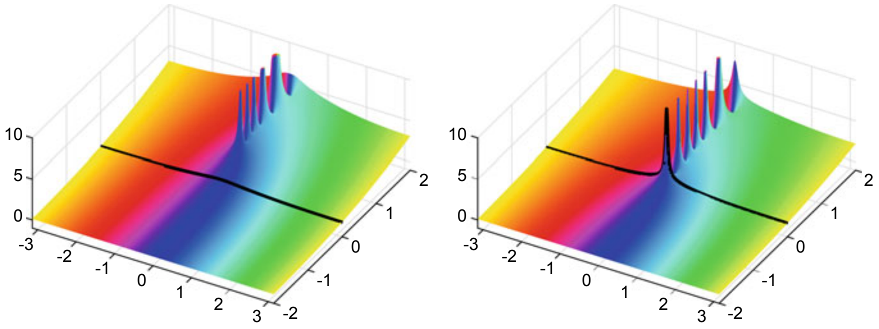


Fig. 5 Finite time blow-up in the Burgers equation (2) with $\nu = 0.1$, subject to the complex initial condition (19). The poles approach the origin from the positive imaginary direction, as can be seen in the left frame, which corresponds to $t = 0.7$. In the right frame the leading pole has reached the real axis, roughly at $t = 1$, which results in a blow-up (note that there is no upper/lower half-plane symmetry as was the case in Fig. 2, so we show both half-planes in this figure)

In conclusion of this section on the Burgers equation we mention a lesser known fact, namely, that nonlinear blow-up is possible with complex initial data. For example, Fig. 5 shows the blow-up in the solution corresponding to the complex Fourier mode initial condition

$$u(x, 0) = -\sin(x) - i\cos(x). \tag{19}$$

Features such as the blow-up time or the minimum value of ν that allows blow-up can be analyzed by the saddle point method outlined above, but we shall not pursue this here.

When dispersion replaces diffusion in (1), the poles drift away from the imaginary axis. The pole behaviour is more complicated than in the Burgers case and the bigger the dispersive effects, the more intricate the behaviour. For this reason we tackle the less famous BO equation first, before getting to the more celebrated KdV equation.

4 The Benjamin-Ono Equation

The periodic Hilbert transform H in (4) can be defined as a convolution integral involving a cotangent kernel [19, Ch. 14], or, equivalently, in terms of Fourier series

$$u(x,t) = \sum_{k=-\infty}^{\infty} c_k(t)e^{ikx} \quad \Rightarrow \quad H\{u_{xx}\} = \sum_{k=-\infty}^{\infty} (-i)\text{sgn}(k)k^2 c_k(t)e^{ikx}. \tag{20}$$

When the nonlinear term in (1) is absent, both the BO and KdV equations are linear dispersive wave equations. They admit travelling wave solutions $u(x,t) =$

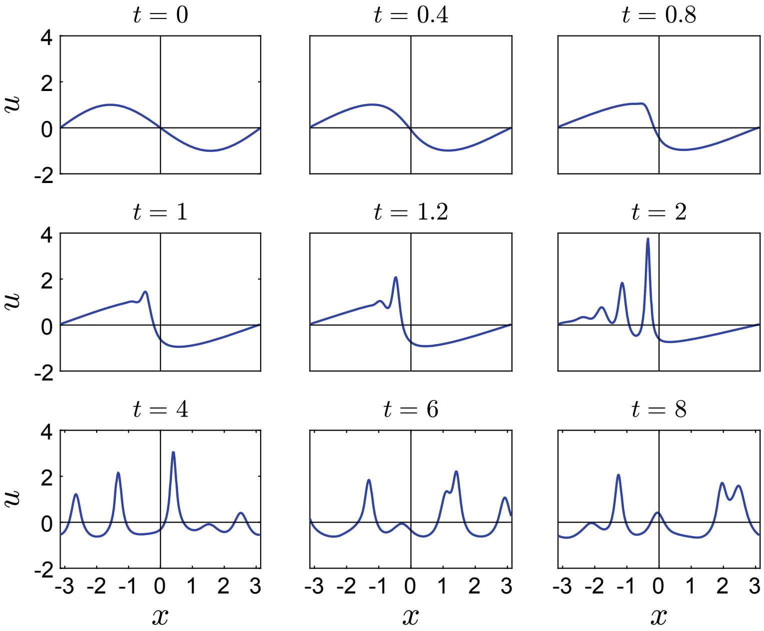


Fig. 6 Solutions to the Benjamin-Ono equation (1) and (4), corresponding to the initial condition (5), with $\nu = 0.1$. The pole dynamics of this solution can be seen in Fig. 7

$e^{i(kx-\omega(k)t)}$ with dispersion relations $\omega = -\nu\,\mathrm{sgn}(k)k^2$ and $\omega = -\nu\,k^3$, respectively. The quadratic vs cubic dependence on the wave number k makes dispersive effects in the BO equation less pronounced than in the KdV equation.

With the nonlinear term in (1) present, both the BO and KdV equations are completely integrable and solvable, in principle, by the inverse scattering transform [1]. For arbitrary initial conditions and particularly with periodic boundary conditions, however, it is unlikely that all steps of the procedure can be completed successfully to obtain explicit solutions. Numerical methods will therefore be used to study singularity dynamics. As mentioned in the introduction, this consists of a standard method of lines procedure to obtain the solution on the real axis, followed by numerical analytical continuation into the complex plane by means of a Fourier-Padé method. Details are postponed to Sect. 7. Our choice of a Padé based method stems from the fact that singularities in both BO and KdV (next section) are expected to be poles. This is related to the complete integrability of these equations and the Painlevé property as discussed in [1, Sect. 2].

Figure 6 shows the solution on the real axis for the BO equation. Like diffusion, dispersion prevents shocks, but the mechanism is different: oscillations appear and separate into travelling wave solutions. In the case of KdV, this behaviour gave rise to the numerical discovery of the soliton, as discussed in Sect. 5. In the present example, about eight such solitons can be seen, perhaps most clearly identifiable in the pole parade shown in Fig. 7.

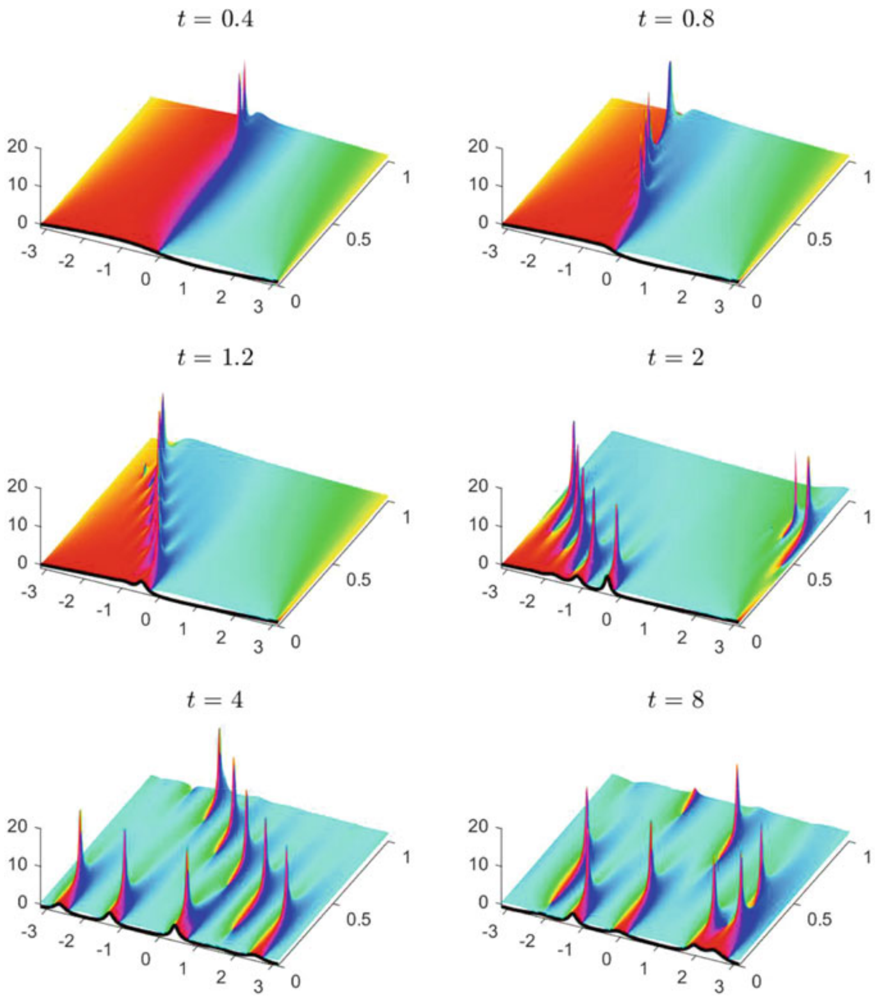


Fig. 7 Pole locations of a subset of the solutions of the BO equation shown in Fig. 6. Each soliton in that figure can be associated with a pair of conjugate simple poles in the complex plane. The poles that exit on the left re-enter on the right because of the periodic boundary conditions

The initial pole behaviour is very similar to that observed in the Burgers equation, namely, the poles are born at infinity and start to travel in conjugate pairs towards the imaginary axes. Unlike the Burgers case, however, the poles do not remain on the imaginary axes but veer off into the left half-plane. Eight pairs can eventually be associated with the solitons shown in Fig. 6.

In the absence of readily computable error estimates for our procedure we have used the following strategy to validate the results. Poles of the BO equation are simple, each with residue $\pm 2i\nu$; see for example [8]. The order and residue of each pole can be checked by contour integration on a small circle surrounding its location [35].[2] Using this technique, spurious poles and other numerical artifacts can be identified (one example of which is the slight irregularity near $-3 + 0.8i$ in the third frame of Fig. 7.)

5 The Korteweg-De Vries Equation

In the case of KdV, the qualitative behaviour of the solutions is similar to that of the BO equation. The dispersion prevents shock formation in the solution by breaking it up into a number of solitons, which is the famous discovery of Zabusky and Kruskal [39]. The iconic figure from that paper is reprinted in Fig. 8. In the left frame of Fig. 9 we reproduce that solution, but rescaled to the domain $[-\pi, \pi]$ in order to facilitate comparisons with the other solutions shown in this paper.

The initial behaviour is the same as for the other equations we have seen thus far, namely, there are poles that enter from infinity and travel towards the real axis in conjugate pairs, roughly similar to the first two frames in Fig. 7. As was the case for the BO equation, dispersion causes the poles to drift into the left half-plane and eventually re-enter in the right half-plane because of periodicity. The eight solitons marked in the Zabusky–Kruskal figure are clearly identifiable in the pole plot of Fig. 9, with the poles closer to the real axis corresponding to the taller solitons.

We have used the same strategy mentioned at the end of Sect. 4 for validation of Fig. 9. In the case of KdV the poles are locally of the form $-12\nu/(z - z_0)^2$. The phase information of Fig. 9, when viewed in colour, makes it clear that the computed poles are indeed of order two, and contour integration confirmed the strength coefficient of -12ν.

It should be noted, however, that numerical analytical continuation is inherently ill-conditioned as one goes further into the complex plane, and that puts some limitations on our investigations. Two examples are as follows:

First, for $t \ll 1$ we found that the Fourier-Padé based method was not able to produce the theoretical pole information accurately, presumably because of the distance between the real axis and the nearest singularity. Therefore no figures of this initial phase of the evolution are presented here. Second, in the literature the existence of 'hidden solitons' in the Zabusky–Kruskal experiment is mentioned; see [12] (and the references therein). In order to investigate these hidden solitons, the solution of Fig. 9 has to be continued much farther into the complex plane. Because of spurious poles and the ill-conditioning alluded to above, our efforts at tracking these hidden solitons were inconclusive. Both of these investigations are offered as a challenge to computational mathematicians.

Here are two suggestions for such investigations. First, for the KdV method it is recommended that the equation be transformed into the potential KdV equation, by

[2] The order of a pole can also be confirmed visually by examining the phase information in the pole plots.

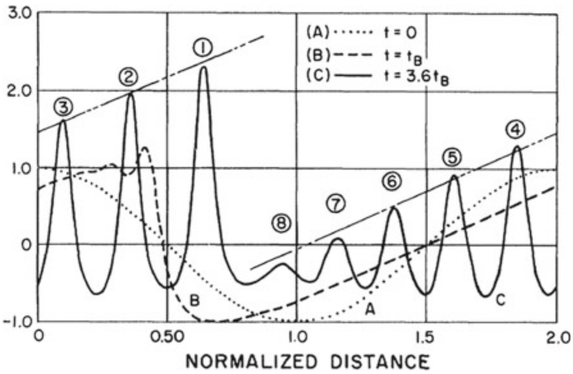


Fig. 8 The iconic figure of soliton formation in the KdV equation. The initial condition is $u(x, 0) = \cos(\pi x)$ on $[0, 2]$, with $\nu = 0.022^2$. Reprinted, with permission, from [39]. Copyright (1965) by the American Physical Society

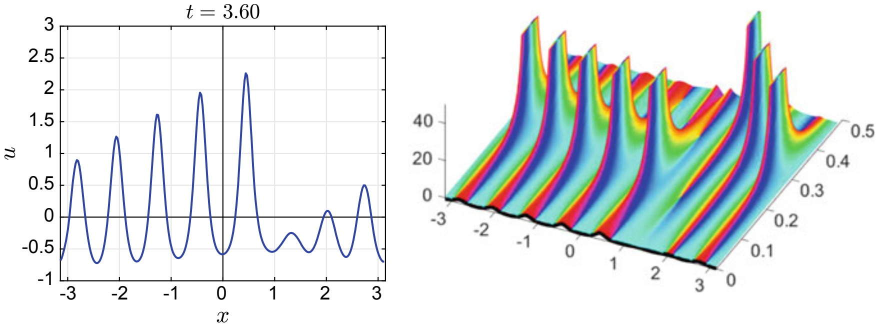


Fig. 9 Left: the Zabusky–Kruskal solution shown in Fig. 8, after rescaling to $[-\pi, \pi]$. Right: the corresponding poles in the complex plane

the substitution $u = v_x$; see [22]. This equation has simple poles, which makes it better suited for approximation by Padé methods. Second, the use of multi-precision arithmetic is advisable. Here, everything was done in IEEE double precision, mainly because of the speed if offers to create animations of the pole parades [36].

6 Recurrence

Historically, the discovery of the soliton in [39] overshadowed the fact that the objective of that paper was something else entirely, namely, the verification of the recurrence phenomenon previously discovered by Fermi, Pasta, Ulam, and Tsingou

(FPUT) in yet another celebrated numerical experiment [16].[3] In short, this means that if a nonlinear system is started in a low mode configuration such as the initial condition (5), then higher modes are created by the nonlinear interaction, causing an energy cascade from low modes to high. The upshot of the FPUT experiment was that this process is not continued indefinitely, but eventually reverses with most of the energy flowing back to the low modes. The effect of this is that the initial condition is reconstructed briefly—approximately so and with a shift in phase—after a certain period of time.

Numerical experiments with KdV such as those reported in Sect. 5 do not reveal the recurrence behaviour in the pole dynamics. Had true recurrence occurred, the poles would have retraced their steps back along the imaginary axes out to infinity or would have cancelled somehow. The most we could observe at the purported recurrence time was a slight widening of the strip of analyticity around the real axis. This lack of a clear recurrence can be attributed to the fact that the phenomenon is rather weak in KdV, as discussed in detail in [20].

For a more convincing demonstration of recurrence one has to look outside the family (1)–(4). Perhaps the best PDE for this purpose is the NLS equation

$$i\,u_t + u_{xx} + \nu |u|^2 u = 0, \tag{21}$$

where the solution, $u(x, t)$, is complex-valued. We shall consider $\nu > 0$ (known as the focussing case) and continue to work with 2π-periodic boundary conditions. It will be necessary, however, to modify our initial condition to have nonzero mean, so we consider

$$u(x, 0) = 1 + \epsilon \cos x. \tag{22}$$

The corresponding solution is an ϵ-perturbation of the x-independent solution $u = e^{i\nu t}$. Linearisation about this solution shows that the side-bands $e^{\pm inx}$ grow exponentially for all integers n satisfying [37, 38]

$$0 < n^2 < 2\nu. \tag{23}$$

That is, for $\nu < \frac{1}{2}$ there is no instability, for $\frac{1}{2} < \nu < 2$ a single pair of side-bands is unstable, a double pair for $2 < \nu < \frac{9}{2}$, and so on. The instability is named after Benjamin and Feir, who derived it not via the NLS but directly from the water wave setting [4]. The growth does not continue unboundedly but subsides, and recurrences occur at periodic time intervals. The connection between Benjamin-Feir instability and FPUT recurrence was pointed out in [38].

The growth and recurrence pattern for a special case with two unstable modes can be seen in Fig. 10. In frames 2, 3 and 7, 8 the unstable mode $e^{\pm ix}$ dominates, while $e^{\pm 2ix}$ dominates in frames 4, 5, and 6. An almost perfect recurrence occurs in frame 9, after which time the process continues periodically.

[3] Since the mid-2000s it has been recognized that Mary Tsingou deserves credit for her computations, and so the FPU experiment was renamed FPUT.

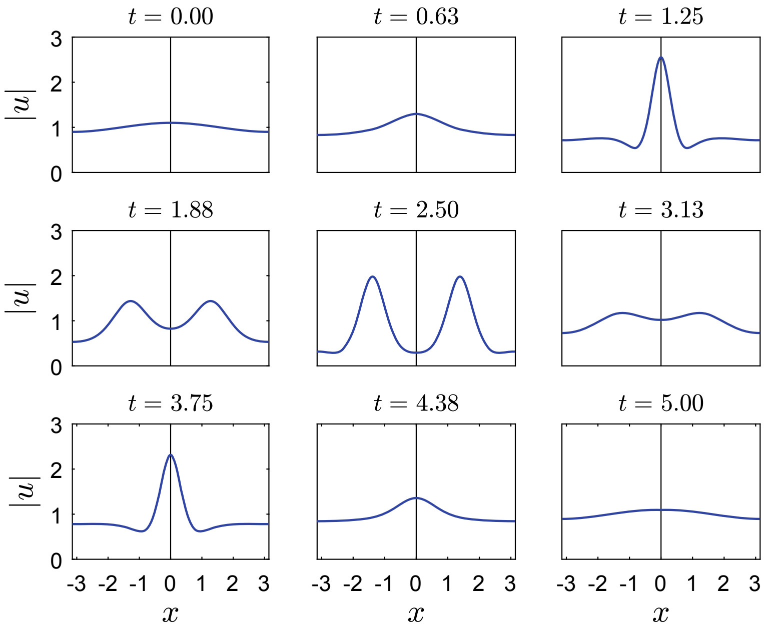
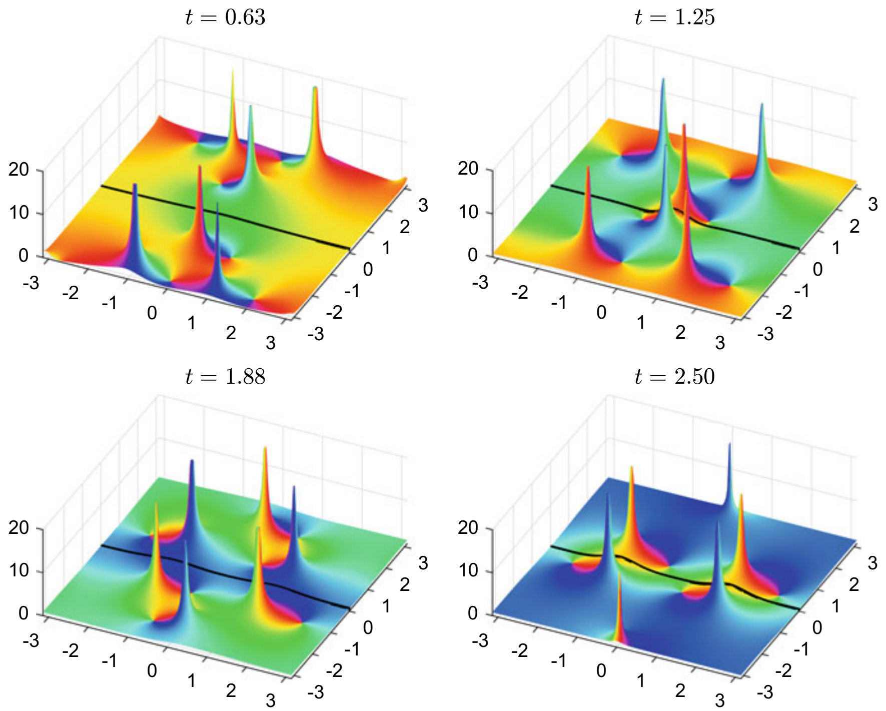


Fig. 10 Solutions to the nonlinear Schrödinger equation (21) corresponding to the initial condition (22), with $\nu = 3$, $\epsilon = 0.1$. The unstable modes $e^{\pm ix}$ and $e^{\pm 2ix}$ take turns in dominating the solution, with a near perfect recurrence at $t = 5$. The pole dynamics of the first phase of this solution can be seen in Fig. 11

Pole locations of some of the solutions in Fig. 10 can be seen in Fig. 11. The first unstable mode is controlled by a conjugate pair of simple poles on the imaginary axis. The second is controlled by two pairs of conjugate poles, each pair symmetrically located with respect to the imaginary axis. The first frame shows the initial onset, with the poles on the imaginary axis leading the procession. The second frame is roughly where the first mode reaches its maximum growth, which corresponds to the point at which the poles reach their minimum distance to the real axis. In the third frame, these poles are receding back along the imaginary axes and are overtaken by the approaching secondary sets of poles. The last frame shows a situation where the second mode has become dominant. At the recurrence time, all of these poles will have receded back to infinity.

7 Numerical Tools

In this final section we review some of the numerical techniques that can be used in this field. Our discussion, which focuses primarily on Padé approximation and its variants, is by no means exhaustive. For other approaches, including tracking the



$t = 0.63$ $t = 1.25$ $t = 1.88$ $t = 2.50$

Fig. 11 Pole locations of a subset of the solutions of the NLS equation shown in Fig. 10. In the first two frames the unstable mode $e^{\pm ix}$ dominates, while $e^{\pm 2ix}$ dominates in the last two frames. This is determined by which pairs of poles are closest to the real axis

poles through the numerical solution of certain dynamical systems, we refer to [7, 26, 32, 33].

We limit the discussion to 2π-periodic solutions that admit a Fourier series expansion of the form

$$u(x,t) = \sum_{k=-\infty}^{\infty} c_k(t) e^{ikx}, \quad -\pi \le x < \pi. \tag{24}$$

In some rare cases the coefficients $c_k(t)$ are known explicitly; cf. (6). Otherwise, the $c_k(t)$ can be computed numerically by a Fourier spectral method and the method of lines [35]. In order to do this step as accurately as possible, it is necessary to truncate the Fourier series to a large number of terms (here we used $|k| \le 256$ or 512), and also use small error tolerances in the time-integration (here on the order of 10^{-12} in the stiff integrator `ode15s` in MATLAB).

When truncated, the series (24) becomes an entire function and will not reveal much singularity information other than perhaps the width of the strip of analyticity around the real axis [32]. A more suitable representation is obtained by converting the truncated series to Fourier-Padé form. For a fixed value of t (suppressed for now in the notation) we convert the series to Taylor-plus-Laurent form by the substitution $z = e^{ix}$:

$$u(x) = \sum_{k=-\infty}^{\infty} c_k e^{ikx} = \sum_{k=0}^{\infty} c_k z^k + \sum_{k=0}^{\infty} c_{-k}(1/z)^k. \tag{25}$$

(It is necessary to redefine $c_0 \to c_0/2$.) Each term on the right can be converted to a type (N, N) rational form as follows. Consider the first term and define

$$f(z) = \sum_{k=0}^{\infty} c_k z^k, \quad p(z) = \sum_{k=0}^{N} a_k z^k, \quad q(z) = \sum_{k=0}^{N} b_k z^k. \tag{26}$$

One then requires that

$$f(z) \approx \frac{p(z)}{q(z)} \quad \Rightarrow \quad p(z) - q(z)f(z) = \mathcal{O}(z^{2N+1}). \tag{27}$$

The latter equation can be set up as a linear system to solve for the coefficients a_k and b_k (after fixing one coefficient, typically $b_0 = 1$). The second term on the right in (25) can be converted to rational form in the same way, which then gives the approximation to $u(x)$ as the ratio of two Fourier-series. The pole plots in Sects. 4, 5 and 6, were all computed using this Fourier-Padé approach.

A promising alternative to the Padé approach to rational approximation is the so-called AAA method, recently proposed in [24], with subsequent extensions to the periodic case [25]. It is not implemented in coefficient space like (24)–(26), but rather uses function values, easily obtained from (26) by an inverse discrete Fourier transform. The representation is the barycentric formula for trigonometric functions [18]

$$u(x) = \frac{\sum_{k=1}^{M} (-1)^k \csc\left(\frac{1}{2}(x - x_k)\right) u_k}{\sum_{k=1}^{M} (-1)^k \csc\left(\frac{1}{2}(x - x_k)\right)}, \tag{28}$$

applicable when M is odd (a similar formula holds for even M). When $x_k = -\pi + (k-1)2\pi/M$ (i.e., evenly spaced nodes in $[-\pi, \pi)$) and $u_k = u(x_k)$, then $u(x)$ is identical to the series (26) when truncated to $|k| \le N$, where $2N + 1 = M$.

In the AAA algorithm the so-called support points x_k are not chosen to be equidistant, which changes the formula (28) from a truncated Fourier series to a rational form. The choice of the x_k proceeds adaptively so as to avoid exponential instabilities.

In preliminary numerical tests the trigonometric AAA algorithm was competitive with the Fourier-Padé method described above. But further experimentation is needed to decide the winner in this particular application field.

Neither of these two methods, however, can give much information on branch point singularities. One way of introducing branches into the approximant is quadratic Padé approximation [30], which is a special case of Hermite-Padé approximation [2]. Define a polynomial $r(x)$ similar to $p(x)$ and $q(x)$ in (26), and in analogy with the rightmost expression in (27) define

$$p(z) + q(z)f(z) + r(z)\big(f(z)\big)^2 = \mathcal{O}(z^{3N+2}). \tag{29}$$

Dropping the order term on the right yields

$$f(z) \approx \frac{-q(z) \pm \sqrt{q(z)^2 - 4p(z)r(z)}}{2\,r(z)}, \tag{30}$$

and when this is used to approximate the two terms on the right of (25) a two-valued approximant to $u(x)$ is obtained. Cubic and higher order approximants can be defined analogously, but will not be considered here.

Recall that Fig. 2 showed a solution of the inviscid Burgers equation with a branch point singularity. To test how accurately this singularity can be approximated by these methods, we solved the equation numerically as described below eq. (24). (We refrained from using the explicit series (6), which is too special.) The numerical solution (24) was then continued into the complex plane using the Fourier-Padé and quadratic Fourier-Padé approximations. Although we have a large number of Fourier coefficients available, we found that best results are obtained if only a fraction of those are used in the Padé approximations. For the results shown here, we used only $N = 35$ terms in the series for $f(z)$ in (26), which translates into a type (17, 17) linear Fourier-Padé approximant, and type (11, 11, 11) in the quadratic case.

The results are shown in Fig. 12. The middle figure is the reference solution, computed to high accuracy by the Newton iteration described in Sect. 2. On the left is the approximation obtained by the linear Fourier-Padé approximant. Away from the imaginary axis the approximation is good, but it is poor on the axis itself. In the absence of branches in the approximant, a series of poles and zeros (the latter not clearly visible) appears as a proxy for the jump in phase. The fact that alternating poles and zeros 'fall in the shadow' of the branch point is a well-known phenomenon in standard Padé approximation [31], and is evidently also present in the trigonometric case.[4] On the other hand, the quadratic Fourier-Padé approximant shown on the right is virtually indistinguishable from the reference solution.

The relative errors in these two approximations are shown in Fig. 13. The linear approximant has low accuracy near the imaginary axis because of the spurious poles mentioned above. By contrast, the quadratic approximant maintains high accuracy, even on the imaginary axis. If one takes the solution generated by the Newton method as exact, the quadratic approximant yields more than five decimal digits of accuracy in almost the whole domain shown in Fig. 13.

Further discussion of numerical aspects of quadratic Padé approximation, including their computation and conditioning, can be found in [15].

[4] Comparing the left frames of Figs. 3 and 12 is interesting. Both solutions can be viewed as a perturbation of the multivalued solution shown in Fig. 2. In Fig. 3 the perturbation is caused by a small amount of diffusion, while in Fig. 12 it is caused by numerical approximation. In both cases the proximity of the multivalued solution is revealed by a sequence of zeros and poles along the phase discontinuity.

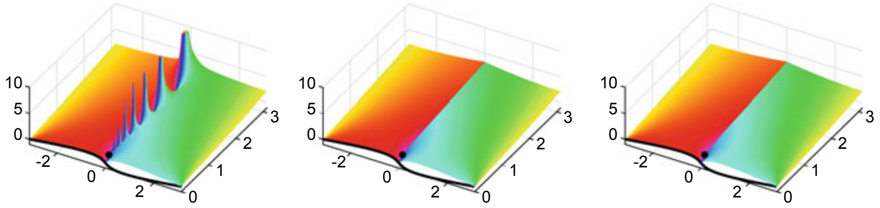


Fig. 12 Approximation of a branch point singularity in the inviscid Burgers equation, at $t = 0.75$. Left: a type (17, 17) linear Padé approximation. Middle: reference solution computed by Newton iteration from (7). Right: a type (11, 11, 11) quadratic Padé approximation

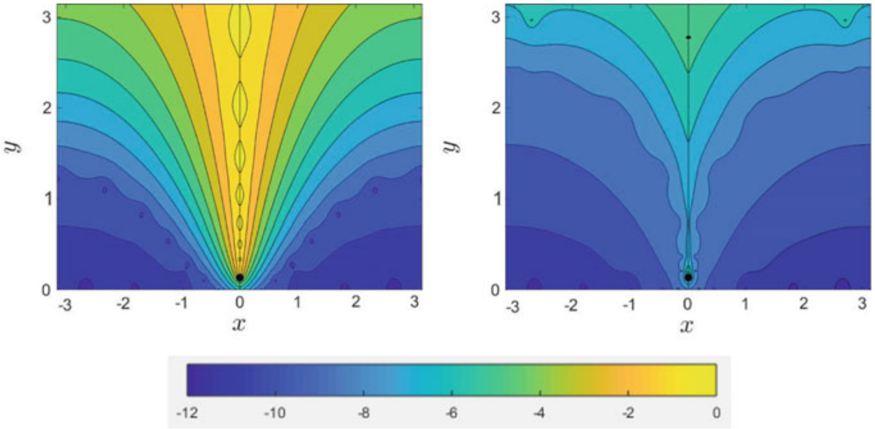


Fig. 13 Relative errors in the approximation of the branch point singularity of Fig. 12. Left: the linear Padé approximation. Right: the quadratic Padé approximation. Bottom: the colour bar in a $\log_{10}$ scale, so each change in shade represents roughly one decimal digit of accuracy

Acknowledgements This paper is an extended version of an invited talk given at ICIAM2019. The author is grateful to the ICIAM secretariat for the invitation. During the preparation the practical assistance and support of Marco Fasondini and Nick Hale were invaluable, as was Nick Trefethen in the role of sounding board. Numerous other people responded to emails and provided input. The software of Elias Wegert [34] made experimentation a pleasure. The author would like to thank the Isaac Newton Institute for Mathematical Sciences for support and hospitality during the programme *Complex analysis: techniques, applications and computations* while this paper was written. This work was supported by: EPSRC grant number EP/R014604/1. A grant from the H.B. Thom foundation of Stellenbosch University is also gratefully acknowledged.

References

1. Ablowitz, M.J., Clarkson, P.A.: Solitons, Nonlinear Evolution Equations and Inverse Scattering, London Mathematical Society Lecture Note Series, vol. 149. Cambridge University Press, Cambridge (1991)
2. Baker, G.A., Jr., Graves-Morris, P.: Padé Approximants, 2nd edn. Cambridge University Press, Cambridge (1996)

3. Baker, G.R., Xie, C.: Singularities in the complex physical plane for deep water waves. J. Fluid Mech. **685**, 83–116 (2011)
4. Benjamin, T.B., Feir, J.E.: The disintegration of wave trains on deep water. J. Fluid Mech. **27**, 417–430 (1967)
5. Bessis, D., Fournier, J.D.: Pole condensation and the Riemann surface associated with a shock in Burgers equation. J. de Phys. Lettres **45**(17), 833–841 (1984)
6. Bessis, D., Fournier, J.D.: Complex singularities and the Riemann surface for the Burgers equation. In: Nonlinear Physics (Shanghai, 1989). Res. Rep. Phys., pp. 252–257. Springer, Berlin (1990)
7. Caflisch, R.E., Gargano, F., Sammartino, M., Sciacca, V.: Complex singularities and PDEs. Riv. Math. Univ. Parma (N.S.) **6**(1), 69–133 (2015)
8. Case, K.M.: The N-soliton solution of the Benjamin-Ono equation. Proc. Nat. Acad. Sci. U.S.A. **75**(8), 3562–3563 (1978)
9. Chiu, T.L., Liu, T.Y., Chan, H.N., Chow, K.W.: The dynamics and evolution of poles and rogue waves for nonlinear Schrödinger equations. Commun. Theor. Phys. (Beijing) **68**(3), 290–294 (2017)
10. Cole, J.D.: On a quasi-linear parabolic equation occurring in aerodynamics. Quart. Appl. Math. **9**, 225–236 (1951)
11. Deconinck, B., Segur, H.: Pole dynamics for elliptic solutions of the Korteweg-de Vries equation. Math. Phys. Anal. Geom. **3**(1), 49–74 (2000)
12. Deng, G., Biondini, G., Trillo, S.: Small dispersion limit of the Korteweg-de Vries equation with periodic initial conditions and analytical description of the Zabusky-Kruskal experiment. Phys. D **333**, 137–147 (2016)
13. Olver, F.W.J., Olde Daalhuis, A.B., Lozier, D.W., Schneider, B.I., Boisvert, R.F., Clark, C.W., Miller, B.R., Saunders, B.V., Cohl, H.S., McClain, M.A. (eds.) NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.0.24 of 2019-09-15. https://dlmf.nist.gov/help/vrml/aboutcolor
14. Dyachenko, A.I., Dyachenko, S.A., Lushnikov, P.M., Zakharov, V.E.: Dynamics of poles in two-dimensional hydrodynamics with free surface: new constants of motion. J. Fluid Mech. **874**, 891–925 (2019)
15. Fasondini, M., Hale, N., Spoerer, R., Weideman, J.A.C.: Quadratic Padé approximation: Numerical aspects and applications. Comp. Res. Model. **11** (2019)
16. Fermi, E., Pasta, J., Ulam, S.: Studies of nonlinear problems. Tech. rep., Los Alamos LA-1940 (1955)
17. Gargano, F., Ponetti, G., Sammartino, M., Sciacca, V.: Complex singularities in KdV solutions. Richerche mat. **65**, 479–490 (2016)
18. Henrici, P.: Barycentric formulas for interpolating trigonometric polynomials and their conjugates. Numer. Math. **33**(2), 225–234 (1979)
19. Henrici, P.: Applied and Computational Complex Analysis. Vol. 3. Pure and Applied Mathematics (New York). Wiley, New York (1986)
20. Herbst, B., Nieddu, G., Trubatch, A.D.: Recurrence in the Korteweg–de Vries equation? In: Nonlinear Wave Equations: Analytic and Computational Techniques. Contemp. Math. vol. 635, pp. 1–12. Amer. Math. Soc., Providence, RI (2015)
21. Hopf, E.: The partial differential equation $u_t + uu_x = \mu u_{xx}$. Comm. Pure Appl. Math. **3**, 201–230 (1950)
22. Kruskal, M.D.: The Korteweg-de Vries equation and related evolution equations. In: Nonlinear Wave Motion. Proceedings of AMS-SIAM Summer Seminar, Clarkson College of Technology, Potsdam, New York. Lectures in Applied Mathematics, Vol. 15, pp. 61–83 (1974)
23. Liu, T.Y., Chiu, T.L., Clarkson, P.A., Chow, K.W.: A connection between the maximum displacements of rogue waves and the dynamics of poles in the complex plane. Chaos **27**(9), 091103, 7 (2017)
24. Nakatsukasa, Y., Sète, O., Trefethen, L.N.: The AAA algorithm for rational approximation. SIAM J. Sci. Comput. **40**(3), A1494–A1522 (2018)
25. Nakatsukasa, Y., Wilbur, H.: Private Communication (2019)

26. Pauls, W., Frisch, U.: A Borel transform method for locating singularities of Taylor and Fourier series. J. Stat. Phys. **127**(6), 1095–1119 (2007)
27. Platzman, G.W.: An exact integral of complete spectral equations for unsteady one-dimensional flow. Tellus **16**, 422–431 (1964)
28. Senouf, D.: Dynamics and condensation of complex singularities for Burgers' equation. I. SIAM J. Math. Anal. **28**(6), 1457–1489 (1997)
29. Senouf, D.: Dynamics and condensation of complex singularities for Burgers' equation. II. SIAM J. Math. Anal. **28**(6), 1490–1513 (1997)
30. Shafer, R.E.: On quadratic approximation. SIAM J. Numer. Anal. **11**, 447–460 (1974)
31. Stahl, H.: The convergence of Padé approximants to functions with branch points. J. Approx. Theory **91**(2), 139–204 (1997)
32. Sulem, C., Sulem, P.L., Frisch, H.: Tracing complex singularities with spectral methods. J. Comput. Phys. **50**(1), 138–161 (1983)
33. Tourigny, Y., Grinfeld, M.: Deciphering singularities by discrete methods. Math. Comp. **62**(205), 155–169 (1994)
34. Wegert, E.: Visual Complex Functions. Birkhäuser/Springer Basel AG, Basel (2012)
35. Weideman, J.A.C.: Computing the dynamics of complex singularities of nonlinear PDEs. SIAM J. Appl. Dyn. Syst. **2**(2), 171–186 (2003)
36. Weideman, J.A.C.: Animations of pole dynamics (2019). http://appliedmaths.sun.ac.za/~weideman/
37. Weideman, J.A.C., Herbst, B.M.: Split-step methods for the solution of the nonlinear Schrödinger equation. SIAM J. Numer. Anal. **23**(3), 485–507 (1986)
38. Yuen, H.C., Ferguson, W.E.: Relationship between Benjamin-Feir instability and recurrence in the nonlinear Schrödinger equation. Phys. Fluids **21**, 1275–1278 (1978)
39. Zabusky, N., Kruskal, M.: Interaction of "solitons" in a collisionless plasma and the recurrence of initial states. Phys. Rev. Lett. **15**, 240–243 (1965)

Author Index

T. Chacón Rebollo et al. (eds.), *Recent Advances in Industrial and Applied Mathematics*, ICIAM 2019 SEMA SIMAI Springer Series 1,
https://doi.org/10.1007/978-3-030-86236-7