

Advances in Mechanics and Mathematics 41

Vinai K. Singh
David Gao
Andreas Fischer *Editors*



Advances in Mathematical Methods and High Performance Computing

 Springer

Advances in Mechanics and Mathematics

Volume 41

Series Editors

David Gao, Federation University Australia
Tudor Ratiu, Shanghai Jiao Tong University

Advisory Board

Antony Bloch, University of Michigan
John Gough, Aberystwyth University
Darryl D. Holm, Imperial College London
Peter Olver, University of Minnesota
Juan-Pablo Ortega, University of St. Gallen
Genevieve Raugel, CNRS and University Paris-Sud
Jan Philip Solovej, University of Copenhagen
Michael Zgurovsky, Igor Sikorsky Kyiv Polytechnic Institute
Jun Zhang, University of Michigan
Enrique Zuazua, Universidad Autónoma de Madrid and DeustoTech

More information about this series at <http://www.springer.com/series/5613>

Vinai K. Singh • David Gao • Andreas Fischer
Editors

Advances in Mathematical Methods and High Performance Computing

 Springer

Editors

Vinai K. Singh
Department of Applied Science and
Humanities
Inderprastha Engineering College
Ghaziabad, Uttar Pradesh, India

David Gao
School of Science and Technology
Federation University Australia
Mt. Helen, VIC, Australia

Andreas Fischer
Institute of Numerical Mathematics
Technische Universität Dresden
Dresden, Germany

ISSN 1571-8689

ISSN 1876-9896 (electronic)

Advances in Mechanics and Mathematics

ISBN 978-3-030-02486-4

ISBN 978-3-030-02487-1 (eBook)

<https://doi.org/10.1007/978-3-030-02487-1>

Library of Congress Control Number: 2018967453

© Springer Nature Switzerland AG 2019

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

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

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

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

Preface

Mathematics and Computer Science stimulate and inspire each other since the formation of the latter in the last century. Mathematical thinking can be traced back to ancient times where nothing like a computer existed. Nevertheless, people were interested in counting, measuring, and comparing objects. Therefore, not only different types of numbers evolved but also the desire of calculating hidden properties of things and processes. This led to algorithms, one of the sources and research objects of nowadays Computer Science.

The increasing power offered by High-Performance Computing (HPC) within the last decades had and has a big impact on the applicability and the design of mathematical algorithms. Existing algorithms could be successfully applied to more and more complex problems from natural and engineering sciences. This has brought new insight into the nature of these problems. Moreover, new ideas for understanding and improving the mathematical methods have been born. Last but not the least, due to these advances, the mathematical modeling of real-world problems has made significant progress. The latter is one of the most important means to combine the chances of HPC with the developments in mathematics and mathematical methods in order to satisfy the increasing needs coming from applications in a variety of fields.

Therefore, we think that it is really worthwhile to demonstrate recent advances in applied mathematics and HPC. The articles in this book show that today's and tomorrow's computational power can be used efficiently if models, methods, and mathematical foundation fit together. It is even foreseen that, due to the expected end of Moore's law, software and mathematical methods will gain more importance.

Many of the articles in the book are devoted to applied problems from biology, chemistry, computational mechanics, environmental sciences, mechanical engineering, operations research, physiology, and several other fields. Frequently, modeling is based on partial differential equations. Further models are related to approximation and optimization. Several articles provide new insights into challenges and developments of high performance and scientific computing. Of course, most of the articles cannot be assigned to one particular type of research. Rather, several aspects of computing, modeling, algorithms, and theory are successfully combined.

Therefore, the assignment of the 33 articles to the 3 parts of the book should be understood as a first orientation. We hope that the reader will find interesting and inspiring articles and that the book will well serve practitioners and researches as well as beginners and experts.

Our sincere thanks go to all authors who submitted a manuscript. Moreover, the work of the reviewers is greatly acknowledged. Last but not the least, the editors would like to express their thanks to Marc Strauss, the Editorial Director, and his team at Springer for their professional support. David Gao's work has been supported by the US Air Force Office of Scientific Research under the grants FA2386-16-1-4082 and FA9550-17-1-0151.

Ghaziabad, India
Ballarat, VIC, Australia
Dresden, Germany
April 2018

Vinai K. Singh
David Gao
Andreas Fischer

Contents

Part I Mathematical Modeling, Applications, and Theoretical Foundations	
Canonical Duality-Triality Theory: Unified Understanding for Modeling, Problems, and NP-Hardness in Global Optimization of Multi-Scale Systems	3
David Gao	
Numerical Investigation of Stochastic Neural Field Equations	51
Pedro M. Lima	
Nonstationary Signal Decomposition for Dummies	69
Antonio Cicone	
Modeling the Socio-Economic Waste Generation Factors Using Artificial Neural Network: A Case Study of Gurugram (Haryana State, India)	83
Ajay Satija, Dipti Singh, and Vinai K. Singh	
Regularization of Highly Ill-Conditioned RBF Asymmetric Collocation Systems in Fractional Models	105
K. S. Prashanthi and G. Chandhini	
The Effect of Toxin and Human Impact on Marine Ecosystem	117
S. Chakraborty and S. Pal	
A Computational Study of Reduction Techniques for the Minimum Connectivity Inference Problem	135
Muhammad Abid Dar, Andreas Fischer, John Martinovic, and Guntram Scheithauer	
Approximate Controllability of Nonlocal Impulsive Stochastic Differential Equations with Delay	149
Surendra Kumar	

Convergence of an Operator Splitting Scheme for Abstract Stochastic Evolution Equations	163
Joshua L. Padgett and Qin Sheng	
Modified Post-Widder Operators Preserving Exponential Functions	181
Vijay Gupta and Vinai K. Singh	
The Properties of Certain Linear and Nonlinear Differential Equations	193
Galina Filipuk and Alexander Chichurin	
Fixed Points for (ϕ, ψ)-Contractions in Menger Probabilistic Metric Spaces	201
Vandana Tiwari and Tanmoy Som	
A Novel Canonical Duality Theory for Solving 3-D Topology Optimization Problems	209
David Gao and Elaf Jaafar Ali	
Part II High Performance and Scientific Computing	
High Performance Computing: Challenges and Risks for the Future	249
Michael M. Resch, Thomas Boenisch, Michael Gienger, and Bastian Koller	
Modern Parallel Architectures to Speed Up Numerical Simulation	259
Mikhail Lavrentiev, Konstantin Lysakov, Alexey Romanenko, and Mikhail Shadrin	
Parallel Algorithms for Low Rank Tensor Arithmetic	271
Lars Grasedyck and Christian Löbbert	
The Resiliency of Multilevel Methods on Next-Generation Computing Platforms: Probabilistic Model and Its Analysis	283
Mark Ainsworth and Christian Glusa	
Visualization of Data: Methods, Software, and Applications	295
Gintautas Dzemyda, Olga Kurasova, Viktor Medvedev, and Giedrė Dzemydaitė	
HPC Technologies from Scientific Computing to Big Data Applications	309
L. M. Patnaik and Srinidhi Hiriyannaiah	
Part III Models, Methods, and Applications Based on Partial Differential Equations	
Analysis and Simulation of Time-Domain Elliptical Cloaks by the Discontinuous Galerkin Method	323
Yunqing Huang, Chen Meng, and Jichun Li	
Dynamic Pore-Network Models Development	337
X. Yin, E. T. de Vries, A. Raoof, and S. M. Hassanizadeh	

Mean Field Magnetohydrodynamic Dynamo in Partially Ionized Plasma: Nonlinear, Numerical Results..... 357
 K. A. P. Singh

Outcome of Wall Features on the Creeping Sinusoidal Flow of MHD Couple Stress Fluid in an Inclined Channel with Chemical Reaction 371
 Mallinath Dhang and Gurunath Sankad

A Fractional Inverse Initial Value Problem 387
 Amin Boumenir and Vu Kim Tuan

Three-Dimensional Biomagnetic Flow and Heat Transfer over a Stretching Surface with Variable Fluid Properties..... 403
 M. G. Murtaza, E. E. Tzirtzilakis and M. Ferdows

Effects of Slip on the Peristaltic Motion of a Jeffrey Fluid in Porous Medium with Wall Effects..... 415
 Gurunath Sankad and Pratima S. Nagathan

Linear and Nonlinear Double Diffusive Convection in a Couple Stress Fluid Saturated Anisotropic Porous Layer with Soret Effect and Internal Heat Source..... 429
 Kanchan Shakya

Modeling of Wave-Induced Oscillation in Pohang New Harbor by Using Hybrid Finite Element Model 449
 Prashant Kumar, Rupali and Rajni

Similarity Solution of Hydromagnetic Flow Near Stagnation Point Over a Stretching Surface Subjected to Newtonian Heating and Convective Condition 457
 KM Kanika, Santosh Chaudhary, and Mohan Kumar Choudhary

Modelling Corrosion Phenomenon of Magnesium Alloy AZ91 in Simulated Body Fluids..... 471
 Ramalingam Vaira Vignesh and Ramasamy Padmanaban

Approximate and Analytic Solution of Some Nonlinear Diffusive Equations 487
 Amitha Manmohan Rao and Arundhati Suresh Warke

Index..... 501

Part I
Mathematical Modeling, Applications,
and Theoretical Foundations

Canonical Duality-Triality Theory: Unified Understanding for Modeling, Problems, and NP-Hardness in Global Optimization of Multi-Scale Systems



David Gao

1 Introduction and Motivation

General problems in mathematical optimization are usually formulated in the following form:

$$(\mathcal{P}_{ap}) : \min f(\mathbf{x}), \quad \text{s.t. } \mathbf{h}(\mathbf{x}) = 0, \quad \mathbf{g}(\mathbf{x}) \leq 0, \quad (1)$$

where the unknown $\mathbf{x} \in \mathbb{R}^n$ is a vector, $f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ is the so-called “objective” function,¹ and $\mathbf{h}(\mathbf{x}) = \{h_i(\mathbf{x})\} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $\mathbf{g}(\mathbf{x}) = \{g_j(\mathbf{x})\} : \mathbb{R}^n \rightarrow \mathbb{R}^p$ are two vector-valued constraint functions. It must be emphasized that, different from the basic concept of *objectivity* in continuum physics and nonlinear analysis, the objective function used extensively in optimization literature is allowed to be any arbitrarily given function, even the linear function. Therefore, the (\mathcal{P}_{ap}) is an *abstractly (or arbitrarily) proposed problem* (APP). Although it enables one to “model” a very wide range of problems, it comes at a price: many global optimization problems are considered to be NP-hard. Without detailed information on these arbitrarily given functions, it is impossible to have a powerful theory for solving the artificial nonconvex problem (1).

Canonical duality-triality is a newly developed and continuously improved methodological theory. This theory comprises mainly: 1) a canonical transformation,

¹This terminology is used mainly in the English literature. The function $f(\mathbf{x})$ is correctly called the target function in all Chinese and Japanese literature.

D. Gao (✉)
School of Science and Technology, Federation University Australia, Mt Helen, VIC 3353,
Australia
e-mail: d.gao@federation.edu.au

which is a versatile methodology that can be used to model complex systems within a unified framework, 2) a complementary-dual principle, which can be used to formulate a perfect dual problem with a unified analytic solution, and 3) a triality theory, which can identify both global and local extrema and to develop effective canonical dual algorithms for solving real-world problems in both continuous and discrete systems. This theory was developed from Gao and Strang's original work on nonconvex variational/boundary-value problems in large deformation mechanics [43]. It was shown in Gao's book [18] and in recent articles [40, 53] that the (external) penalty and Lagrange multiplier methods are special applications of the canonical duality theory in convex optimization. It is now understood that this theory reveals an intrinsic multi-scale duality pattern in complex systems, many popular theories and methods in nonconvex analysis, global optimization, and computational science can be unified within the framework of the canonical duality-triality theory. Indeed, it is easy to show that the KKT theory in mathematical programming, the *semi-definite programming* (SDP) method in global optimization, and the *half-quadratic regularization* in information technology are naturally covered by the canonical duality theory [39, 56, 86].

Mathematics and mechanics have been complementary partners since the Newton times. Many fundamental ideas, concepts, and mathematical methods extensively used in calculus of variations and optimization are originally developed from mechanics. It is known that the classical Lagrangian duality theory and the associated Lagrange multiplier method were developed by Lagrange in analytical mechanics [51]. The modern concepts of super-potential and sub-differential in convex analysis were proposed by J.J. Moreau from frictional mechanics [63]. The canonical duality theory is also developed from the fundamental concepts of objectivity and work-conjugate principle in continuum physics. The Gao-Strang gap function discovered in finite deformation theory provides a global optimality condition for general nonconvex/nonsmooth variational analysis and global optimization. Application of this theory to nonlinear elasticity leads to a pure complementary energy principle which was a 50-year-old open problem [58]. Generalization to global optimization was made in 2000 [20]. Since then, this theory has been used successfully for solving a large class of challenging problems in multi-disciplinary fields of applied mathematics, computational science, engineering mechanics, operations research, and industrial and systems engineering [11–17, 22–25, 36–38, 40, 42, 44, 45, 70, 73].

However, as V.I. Arnold indicated [2]: “In the middle of the twentieth century it was attempted to divide physics and mathematics. The consequences turned out to be catastrophic.” Indeed, due to the ever-increasing gap between physics and other fields, some well-defined concepts in continuum physics, such as objectivity, Lagrangian, tensor, and fully nonlinearity, etc., have been seriously misused in optimization, which leads to not only ridiculous arguments but also wrong mathematical models and many artificially proposed problems. Also, the canonical dual transformation theory and methodology have been rediscovered in different forms by researchers from different fields. The main goal of this paper is to bridge this gap by presenting the canonical duality theory in a systematical way

from a unified modeling, basic assumptions to the theory, method, and general applications. The methodology, examples, and conjectures presented in this paper are important not only for better understanding this unconventional theory but also for solving many challenging problems in complex systems. This paper will bring some fundamentally new insights into multi-scale complex systems, global optimization, and computational science.

2 Multi-Scale Modeling and Properly Posed Problems

In linguistics, a complete and grammatically correct sentence should be composed by at least three words: subject, object, and a predicate.² As a language of science, the mathematics should follow this rule. Based on the canonical duality principle [18], a unified mathematical problem for multi-scale complex systems was proposed in [26, 28]:

$$(\mathcal{P}) : \min\{\Pi(\boldsymbol{\chi}) = G(\mathbf{D}\boldsymbol{\chi}) - F(\boldsymbol{\chi}) \mid \boldsymbol{\chi} \in \mathcal{X}_c\}, \quad (2)$$

where $F : \mathcal{X}_a \subset \mathcal{X} \rightarrow \mathbb{R}$ is a subjective function such that the external duality relation $\boldsymbol{\chi}^* = \nabla F(\boldsymbol{\chi}) = \bar{\boldsymbol{\chi}}^*$ is a given input (or source), its domain \mathcal{X}_a contains only geometrical constraints (such as boundary or initial conditions), which depends on each given problem; $\mathbf{D} : \mathcal{X}_a \rightarrow \mathcal{G}_a$ is a linear operator which links the configuration variable $\boldsymbol{\chi} \in \mathcal{X}_a$ with an internal variable $\mathbf{g} = \mathbf{D}\boldsymbol{\chi} \in \mathcal{G}_a$ at different physical scales; $G : \mathcal{G}_a \subset \mathcal{G} \rightarrow \mathbb{R}$ is an objective function such that the internal duality relation $\mathbf{g}^* = \nabla G(\mathbf{g})$ is governed by the constitutive law, its domain \mathcal{G}_a contains only physical constraints, which depends on mathematical modeling; The feasible set is defined by:

$$\mathcal{X}_c = \{\boldsymbol{\chi} \in \mathcal{X}_a \mid \mathbf{D}\boldsymbol{\chi} \in \mathcal{G}_a\}. \quad (3)$$

The predicate in (\mathcal{P}) is the operator “−” and the difference $\Pi(\boldsymbol{\chi})$ is called the target function in general problems. The object and subject are in balance only at the optimal states.

2.1 Objectivity, Isotropy, and Symmetry in Modeling

Objectivity is a central concept in our daily life, related to reality and truth. According to Wikipedia, the objectivity in philosophy means the state or quality of

²By the facts that (object, subject) is a duality pair in a noun (or pronoun) space, which is dual to a verb space, the multi-level duality pattern $\{(object, subject); predicate\}$ is called triality, which is essential for languages and sciences.

being true even outside a subject's individual biases, interpretations, feelings, and imaginings.³ In science, the objectivity is often attributed to the property of scientific measurement, as the accuracy of a measurement can be tested independent from the individual scientist who first reports it.⁴ In continuum mechanics, the objectivity is also called the principle of frame-indifference [65, 80], which is a basic concept in mathematical modeling [8, 18, 60] but is still subjected to seriously study in continuum physics [59, 64]. Let \mathcal{R} be a special orthogonal group $\text{SO}(n)$, i.e., $\mathbf{R} \in \mathcal{R}$ if and only if $\mathbf{R}^T = \mathbf{R}^{-1}$ and $\det \mathbf{R} = 1$. The following mathematical definition was given in Gao's book (Definition 6.1.2 [18]).

Definition 1 (Objectivity and Isotropy) A set \mathcal{G}_a is said to be objective if $\mathbf{R}\mathbf{g} \in \mathcal{G}_a$ $\forall \mathbf{g} \in \mathcal{G}_a, \forall \mathbf{R} \in \mathcal{R}$. A real-valued function $G : \mathcal{G}_a \rightarrow \mathbb{R}$ is said to be objective if

$$G(\mathbf{R}\mathbf{g}) = G(\mathbf{g}) \quad \forall \mathbf{g} \in \mathcal{G}_a, \forall \mathbf{R} \in \mathcal{R}. \quad (4)$$

A set \mathcal{G}_a is said to be isotropic if $\mathbf{g}\mathbf{R} \in \mathcal{G}_a$ $\forall \mathbf{g} \in \mathcal{G}_a, \forall \mathbf{R} \in \mathcal{R}$. A real-valued function $G : \mathcal{G}_a \rightarrow \mathbb{R}$ is said to be isotropic if

$$G(\mathbf{g}\mathbf{R}) = G(\mathbf{g}) \quad \forall \mathbf{g} \in \mathcal{G}_a, \forall \mathbf{R} \in \mathcal{R}. \quad (5)$$

Lemma 1 A real-valued function $G(\mathbf{g})$ is objective if and only if there exists a real-valued function $\Phi(\mathbf{C})$ such that $G(\mathbf{g}) = \Phi(\mathbf{g}^T \mathbf{g})$.

Geometrically speaking, an objective function is rotational symmetry, which should be an $\text{SO}(n)$ -invariant in n -dimensional Euclidean space. Physically, an objective function doesn't depend on observers. Because of Noether's theorem,⁵ rotational symmetry of a physical system is equivalent to the *angular momentum conservation law* (see Section 6.1.2 [18]). Therefore, the objectivity is essential for any real-world mathematical models. In Euclidean space $\mathcal{G}_a \subset \mathbb{R}^n$, the simplest objective function is the ℓ_2 -norm $\|\mathbf{g}\|$ in \mathbb{R}^n as we have $\|\mathbf{R}\mathbf{g}\|^2 = \mathbf{g}^T \mathbf{R}^T \mathbf{R} \mathbf{g} = \|\mathbf{g}\|^2$ $\forall \mathbf{R} \in \mathcal{R}$. In continuum physics, the objectivity implies that the equilibrium condition of angular momentum (symmetry of the Cauchy stress tensor $\boldsymbol{\sigma} = \partial G(\mathbf{g})$, Section 6.1 [18]) holds. It was emphasized by P.G. Ciarlet that *the objectivity is not an assumption, but an axiom* [8]. In Gao and Strang's work, the internal energy $W(\mathbf{g})$ must be an objective function such that its variation (Gâteaux derivative) $\boldsymbol{\sigma} = \partial W(\mathbf{g})$ is the so-called *constitutive duality law*, which depends only on the intrinsic property of the system.

³[https://en.wikipedia.org/wiki/Objectivity_\(philosophy\)](https://en.wikipedia.org/wiki/Objectivity_(philosophy)).

⁴[https://en.wikipedia.org/wiki/Objectivity_\(science\)](https://en.wikipedia.org/wiki/Objectivity_(science)).

⁵That is, every differentiable symmetry of the action of a physical system has a corresponding conservation law.

2.2 Subjectivity, Symmetry Breaking, and Well-Posed Problem

Dual to the objective function that depends on modeling, the subjective function $F(\chi)$ depends on each problem such that its variation is governed by the *action-reaction duality law*: $\bar{\chi}^* = \partial F(\chi) \in \mathcal{X}^*$. From the point view of systems theory, the action $\bar{\chi}^* \in \mathcal{X}^*$ can be considered as the input or source of the system, and the reaction $\chi \in \mathcal{X}$ should be the output (or the configuration, the state) of the system. A system is conservative if the action is independent of the reaction. Therefore, the subjective function must be linear on its domain \mathcal{X}_a and, by Riesz representation theorem, we should have $F(\chi) = \langle \chi, \bar{\chi}^* \rangle$, where the bilinear form $\langle \chi, \chi^* \rangle : \mathcal{X} \times \mathcal{X}^* \rightarrow \mathbb{R}$ puts \mathcal{X} and \mathcal{X}^* in duality. The target function $\Pi(\chi) = G(\mathbf{D}\chi) - F(\chi)$ can have different physical meanings in real-world applications. For example, in continuum mechanics the subjective function $F(\chi)$ is the external energy, the objective function $G(\mathbf{g})$ is the stored energy, then $\Pi(\chi)$ is the total potential energy. In this case, the minimum total potential energy principle leads to the general variational problem (2). The criticality condition $\partial \Pi(\chi) = 0$ leads to the equilibrium (Euler-Lagrange) equation:

$$A(\chi) = \mathbf{D}^* \partial G(\mathbf{D}\chi) = \bar{\chi}^* \quad (6)$$

where $\mathbf{D}^* : \mathcal{G}_a^* \rightarrow \mathcal{X}^*$ is an adjoint operator of \mathbf{D} and $A : \mathcal{X}_c \rightarrow \mathcal{X}^*$ is called *equilibrium operator*. The triality structure $\mathbb{S}^e = \{(\mathcal{X}, \mathcal{X}^*); A\}$ forms an elementary system in Gao's book (Section 4.3, [18]). This abstract form covers the most well-known equilibrium problems in real-world applications ranging from mathematical physics in continuous analysis to mathematical programming in discrete systems [18, 34]. Particularly, if $G(\mathbf{g})$ is quadratic such that $\partial^2 G(\mathbf{g}) = \mathbf{H}$, then the operator $A : \mathcal{X}_c \rightarrow \mathcal{X}^*$ is linear and can be written in the triality form: $A = \mathbf{D}^* \mathbf{H} \mathbf{D}$, which appears extensively in mathematical physics, optimization, and linear systems (see the celebrated text by Strang [77]). Clearly, any convex quadratic function $G(\mathbf{D}\chi)$ is objective due to the Cholesky decomposition $A = \Lambda^* \Lambda \geq 0$.

According to the action-reaction duality in physics, if there is no action or demand (i.e., $\bar{\chi}^* = 0$), the system has no reaction (i.e., $\chi = 0$). Dually, a real-world problem should have at least one nontrivial solution for any given nontrivial input.

Definition 2 (Properly and Well-Posed Problems) A problem is called *properly posed* if for any given nontrivial input it has at least one nontrivial solution. It is called *well-posed* if the solution is unique.

Clearly, this definition is more general than Hadamard's well-posed problems in dynamical systems since the continuity condition is not required. Physically speaking, any real-world problems should be well-posed since all natural phenomena exist uniquely. But practically, it is difficult to model a real-world problem precisely. Therefore, properly posed problems are allowed for the canonical duality theory. This definition is important for understanding the triality theory and NP-hard problems.

2.3 Management Optimization

In management science, the decision variable χ is simply a vector $\mathbf{x} \in \mathbb{R}^n$, which could represent the products of a manufacture company. The input $\bar{\chi}^*$ can be considered as market price (or demanding), denoted by $\mathbf{f} \in \mathbb{R}^n$. Therefore, the subjective function $\langle \mathbf{x}, \mathbf{f} \rangle = \mathbf{x}^T \mathbf{f}$ in this example is the total income of the company. The products are produced by workers $\mathbf{g} \in \mathbb{R}^m$. Due to the cooperation, we have $\mathbf{g} = \mathbf{D}\mathbf{x}$ and $\mathbf{D} \in \mathbb{R}^{m \times n}$ is a matrix. Workers are paid by salary $\mathbf{g}^* = \partial G(\mathbf{g})$, and therefore, the objective function $G(\mathbf{g})$ is the cost (in this example G is not necessarily to be objective since the company is a man-made system). Then, the target $\Pi(\mathbf{x}) = G(\mathbf{D}\mathbf{x}) - \mathbf{x}^T \mathbf{f}$ is the total loss and the minimization problem $\min \Pi(\mathbf{x})$ leads to the equilibrium equation:

$$\mathbf{D}^T \partial_{\mathbf{g}} G(\mathbf{D}\mathbf{x}) = \mathbf{f}.$$

The cost function $G(\mathbf{g})$ could be convex for a small company, but usually nonconvex for big companies to allow some people having the same salaries.

If the company has to make a profit $\frac{1}{2}\alpha\|\mathbf{x}\|^2$, where $\alpha > 0$ is a parameter, then the target function is $\Pi(\mathbf{x}) = G(\mathbf{D}\mathbf{x}) + \frac{1}{2}\alpha\|\mathbf{x}\|^2 - \mathbf{x}^T \mathbf{f}$ and the minimization problem $\min \Pi(\mathbf{x})$ leads to:

$$\alpha\mathbf{x} = \mathbf{f} - \mathbf{D}^T \partial_{\mathbf{g}} G(\mathbf{D}\mathbf{x}). \quad (7)$$

This is a fixed point problem. In this case, if we let $\bar{\mathbf{g}} = \bar{\mathbf{D}}\mathbf{x} = (\mathbf{D}\mathbf{x}, \mathbf{x})$ and $\bar{G} = G(\mathbf{g}) + \frac{1}{2}\alpha\|\mathbf{x}\|^2$, then the fixed point problem (7) can be written in the unified form of:

$$\bar{\mathbf{D}}^T \partial_{\bar{\mathbf{g}}} \bar{G}(\bar{\mathbf{D}}\mathbf{x}) = \mathbf{f}.$$

This shows that the fixed point problem is a special case of the general equilibrium equation (6), a necessary condition of the general minimization problem (\mathcal{P}_g).

2.4 Nonconvex Analysis and Boundary-Value Problems

For static systems, the unknown of a mixed boundary-value problem is a vector-valued function:

$$\begin{aligned} \chi(\mathbf{x}) \in \mathcal{X}_a &= \{\chi \in \mathcal{C}[\Omega, \mathbb{R}^m] \mid \chi(\mathbf{x}) = 0 \quad \forall \mathbf{x} \in \Gamma_\chi\}, \\ \Omega &\subset \mathbb{R}^d, \quad d \leq 3, \quad m \geq 1, \quad \partial\Omega = \Gamma_\chi \cup \Gamma_t, \end{aligned}$$

and the input is $\bar{\chi}^* = \{\mathbf{f}(\mathbf{x}) \ \forall \mathbf{x} \in \Omega, \ \mathbf{t}(\mathbf{x}) \ \forall \mathbf{x} \in \Gamma_t\}$ [43]. In this case, the external energy is $F(\chi) = \langle \chi, \bar{\chi}^* \rangle = \int_{\Omega} \chi \cdot \mathbf{f} \, d\Omega + \int_{\Gamma_t} \chi \cdot \mathbf{t} \, d\Gamma$. In nonlinear analysis, \mathbf{D} is a gradient-like partial differential operator and $\mathbf{g} = \mathbf{D}\chi \in \mathcal{G}_a \subset \mathcal{L}^p[\Omega; \mathbb{R}^{m \times d}]$ is a *two-point tensor field* [18] over Ω . The internal energy $G(\mathbf{g})$ is defined by:

$$G(\mathbf{g}) = \int_{\Omega} U(\mathbf{x}, \mathbf{g}) \, d\Omega, \quad (8)$$

where $U(\mathbf{x}, \mathbf{g}) : \Omega \times \mathcal{G}_a \rightarrow \mathbb{R}$ is the *stored energy density*. The system is (space) *homogeneous* if $U = U(\mathbf{g})$. Thus, $G(\mathbf{g})$ is objective if and only if $U(\mathbf{x}, \mathbf{g})$ is objective on an objective set \mathcal{G}_a . By the facts that $\mathbf{g} = \mathbf{D}\mathbf{u}$ is a two-point tensor, which is not considered as a strain measure, but the (right) Cauchy-Green tensor $\mathbf{C} = \mathbf{g}^T \mathbf{g}$ is an objective strain tensor, there must exist a function $\Phi(\mathbf{C})$ such that $G(\mathbf{g}) = \Phi(\mathbf{C})$. In nonlinear elasticity, the function $\Phi(\mathbf{C})$ is usually convex and the duality $\mathbf{C}^* = \partial\Phi(\mathbf{C})$ is invertible (i.e., Hill's work-conjugate principle [18]). These basic truths in continuum physics laid a foundation for the canonical duality theory.

By finite element method, the domain Ω is divided into m -elements $\{\Omega^e\}$ such that the unknown function is piecewisely discretized by $\chi(\mathbf{x}) \simeq \mathbf{N}_e(\mathbf{x})\chi_e \ \forall \mathbf{x} \in \Omega^e$. Thus, the nonconvex variational problem (2) can be numerically reformulated in a global optimization problem:

$$\min\{\Pi(\chi) = G(\mathbf{D}\chi) - \langle \chi, \mathbf{f} \rangle \mid \chi \in \mathcal{X}_c\}, \quad (9)$$

where $\chi = \{\chi_e\}$ is the discretized unknown $\chi(\mathbf{x})$, \mathbf{D} is a generalized matrix depending on the interpolation $\mathbf{N}_e(\mathbf{x})$, and \mathcal{X}_c is a convex constraint set including the boundary conditions. The canonical dual finite element method was first proposed in 1996 [11]. Applications have been given recently in engineering and sciences [30, 45, 73].

2.5 Lagrangian Mechanics and Initial-Value Problems

In Lagrange mechanics [51, 52], the unknown $\chi(t) \in \mathcal{X}_a \subset \mathcal{C}^1[I; \mathbb{R}^n]$ is a vector field over a time domain $I \subset \mathbb{R}$. Its components $\{\chi_i(t)\}$ ($i = 1, \dots, n$) are known as the *Lagrangian coordinates*. Its dual variable $\bar{\chi}^*$ is the action vector function in \mathbb{R}^n , denoted by $\mathbf{f}(t)$. The external energy $F(\chi) = \langle \chi, \bar{\chi}^* \rangle = \int_I \chi(t) \cdot \mathbf{f}(t) \, dt$. While the internal energy $G(\mathbf{D}\chi)$ is the so-called action:

$$G(\mathbf{D}\chi) = \int_I L(t, \chi, \dot{\chi}) \, dt, \quad L = T(\dot{\chi}) - U(t, \chi), \quad (10)$$

where $\mathbf{D}\chi = \{1, \partial_t\}\chi = \{\chi, \dot{\chi}\}$ is a vector-valued mapping, T is the kinetic energy density, U is the potential density, and $L = T - U$ is the *Lagrangian density*. Together, $\Pi(\chi) = G(\mathbf{D}\chi) - F(\chi)$ is called the *total action*. This standard form

holds from the classical Newton mechanics to quantum field theory.⁶ Its stationary condition leads to the well-known *Euler-Lagrange equation*:

$$A(\boldsymbol{\chi}) = \mathbf{D}^* \partial G(\mathbf{D}\boldsymbol{\chi}) = \{-\partial_t, 1\} \cdot \partial L(\boldsymbol{\chi}, \dot{\boldsymbol{\chi}}) = -\partial_t \partial_{\dot{\boldsymbol{\chi}}} T(\dot{\boldsymbol{\chi}}) - \partial_{\boldsymbol{\chi}} U(t, \boldsymbol{\chi}) = \mathbf{f}. \quad (11)$$

The system is called (time) *homogeneous* if $L = L(\boldsymbol{\chi}, \dot{\boldsymbol{\chi}})$. In general, the kinetic energy T must be an objective function of the velocity⁷ $\mathbf{v}_k = \dot{\mathbf{x}}_k(\boldsymbol{\chi})$ of each particle $\mathbf{x}_k = \mathbf{x}_k(\boldsymbol{\chi}) \in \mathbb{R}^3 \quad \forall k \in \{1, \dots, K\}$, while the potential density U depends on each problem. For Newtonian mechanics, we have $\boldsymbol{\chi}(t) = \mathbf{x}(t)$ and $T(\mathbf{v}) = \frac{1}{2}m\|\mathbf{v}\|^2$ is quadratic. If $U = 0$, the equilibrium equation $A(\boldsymbol{\chi}) = -m\ddot{\mathbf{x}}(t) = \mathbf{f}$ includes the Newton second law: $\mathbf{F} = m\ddot{\mathbf{x}}$ and the third law: $-\mathbf{F} = \mathbf{f}$. The first law $\mathbf{v} = \dot{\mathbf{x}} = \mathbf{v}_0$ holds only if $\mathbf{f} = 0$. In this case, the system has either a trivial solution $\mathbf{x} = 0$ or infinitely many solutions $\mathbf{x}(t) = \mathbf{v}_0 t + \mathbf{x}_0$, depending on the initial conditions in \mathcal{X}_a . This simple fact in elementary physics plays a key role in understanding the canonical duality theory and NP-hard problems in global optimization.

By using the methods of finite difference and least squares [39, 54], the general nonlinear dynamical system (11) can also be formulated as the same global optimization problem (9), where $\boldsymbol{\chi} = \{\chi_i(t_k)\}$ is the Lagrangian coordinates $\chi_i (i = 1, \dots, n)$ at each discretized time $t_k (k = 1, \dots, m)$, \mathbf{D} is a finite difference matrix, and \mathcal{X}_c is a convex constraint set including the initial condition [54]. By the canonical duality theory, an intrinsic relation between chaos in nonlinear dynamics and NP-hardness in global optimization was revealed recently in [54].

2.6 Mono-Duality and Duality Gap

Lagrangian duality was developed from Lagrange mechanics since 1788 [51], where the kinetic energy $T(\mathbf{v}) = \sum_k \frac{1}{2}m_k\|\mathbf{v}_k\|^2$ is a quadratic (objective) function. For convex static systems (or dynamical systems but $U(\boldsymbol{\chi}) = 0$), the stored energy $G : \mathcal{G}_a \rightarrow \mathbb{R}$ is convex and its Legendre conjugate $G^*(\boldsymbol{\sigma}) = \{\langle \mathbf{g}; \boldsymbol{\sigma} \rangle - G(\mathbf{g}) \mid \boldsymbol{\sigma} = \partial G(\mathbf{g})\}$ is uniquely defined on \mathcal{G}_a^* . Thus, by $G(\mathbf{D}\boldsymbol{\chi}) = \langle \mathbf{D}\boldsymbol{\chi}; \boldsymbol{\sigma} \rangle - G^*(\boldsymbol{\sigma})$ the total action or potential $\Pi(\boldsymbol{\chi})$ can be written in the Lagrangian form⁸ $L : \mathcal{X}_a \times \mathcal{G}_a^* \rightarrow \mathbb{R}$:

$$L(\boldsymbol{\chi}, \boldsymbol{\sigma}) = \langle \mathbf{D}\boldsymbol{\chi}; \boldsymbol{\sigma} \rangle - G^*(\boldsymbol{\sigma}) - \langle \boldsymbol{\chi}, \mathbf{f} \rangle = \langle \boldsymbol{\chi}, \mathbf{D}^* \boldsymbol{\sigma} - \mathbf{f} \rangle - G^*(\boldsymbol{\sigma}), \quad (12)$$

where $\boldsymbol{\chi} \in \mathcal{X}_a$ can be viewed as a Lagrange multiplier for the equilibrium equation $\mathbf{D}^* \boldsymbol{\sigma} = \mathbf{f} \in \mathcal{X}_a^*$. In linear elasticity, $L(\boldsymbol{\chi}, \boldsymbol{\sigma})$ is the well-known *Hellinger-Reissner*

⁶See Wikipedia: https://en.wikipedia.org/wiki/Lagrangian_mechanics.

⁷The objectivity of $T(\mathbf{v})$ is also called the isotropy in Lagrange mechanics since \mathbf{v} is a vector (see [52]).

⁸In the Physics literature, the same notation L is used for both action $L(\boldsymbol{\chi}, \dot{\boldsymbol{\chi}})$ and the Lagrangian $L(\boldsymbol{\chi}, \mathbf{p})$ since both represent the same physical quantity.

complementary energy [18]. Let $\mathcal{S}_c = \{\boldsymbol{\sigma} \in \mathcal{G}_a^* \mid \mathbf{D}^*\boldsymbol{\sigma} = \mathbf{f}\}$ be the so-called *statically admissible space*. Then, the Lagrangian dual of the general problem (\mathcal{P}) is given by:

$$(\mathcal{P}^*) : \quad \max\{\Pi^*(\boldsymbol{\sigma}) = -G^*(\boldsymbol{\sigma}) \mid \boldsymbol{\sigma} \in \mathcal{S}_c\}, \quad (13)$$

and the saddle Lagrangian leads to a well-known min-max duality in convex (static) systems:

$$\min_{\boldsymbol{\chi} \in \mathcal{X}_c} \Pi(\boldsymbol{\chi}) = \min_{\boldsymbol{\chi} \in \mathcal{X}_a} \max_{\boldsymbol{\sigma} \in \mathcal{G}_a^*} L(\boldsymbol{\chi}, \boldsymbol{\sigma}) = \max_{\boldsymbol{\sigma} \in \mathcal{G}_a^*} \min_{\boldsymbol{\chi} \in \mathcal{X}_a} L(\boldsymbol{\chi}, \boldsymbol{\sigma}) = \max_{\boldsymbol{\sigma} \in \mathcal{S}_c} \Pi^*(\boldsymbol{\sigma}). \quad (14)$$

This one-to-one duality is the so-called *mono-duality* in Chapter 1 [18], or the *complementary-dual variational principle* in continuum physics. In finite elasticity, the Lagrangian dual is also known as the *Levison-Zubov principle*. However, this principle holds only for convex problems. In real-world problems, the stored energy $G(\mathbf{g})$ is usually nonconvex in order to model complex phenomena. Its complementary energy can't be determined uniquely by the Legendre transformation. Although its Fenchel conjugate $G^\sharp : \mathcal{G}_a^* \rightarrow \mathbb{R} \cup \{+\infty\}$ can be uniquely defined, the Fenchel-Moreau dual problem:

$$(\mathcal{P}^\sharp) : \quad \max\{\Pi^\sharp(\boldsymbol{\sigma}) = -G^\sharp(\boldsymbol{\sigma}) \mid \boldsymbol{\sigma} \in \mathcal{S}_c\} \quad (15)$$

is not considered as a complementary-dual problem due to Fenchel-Young inequality:

$$g_a = \min\{\Pi(\boldsymbol{\chi}) \mid \boldsymbol{\chi} \in \mathcal{X}_c\} \geq \max\{\Pi^\sharp(\boldsymbol{\sigma}) \mid \boldsymbol{\sigma} \in \mathcal{S}_c\} = g_p, \quad (16)$$

and $g_{ap} = g_a - g_p \neq 0$ is the well-known *duality gap*. This duality gap is intrinsic to all Lagrange-Fenchel-Moreau types of duality problems since the linear operator \mathbf{D} can't change the nonconvexity of $G(\mathbf{D}\boldsymbol{\chi})$. It turns out that the existence of a pure stress $\boldsymbol{\sigma}$ based complementary-dual principle was a well-known debate in nonlinear elasticity for more than fifty years [58].

Remark 1 (Equilibrium Constraints and Lagrange Multiplier Law) Strictly speaking, the Lagrange multiplier method can be used mainly for equilibrium constraint in \mathcal{S}_c and the Lagrange multiplier must be the solution to the primal problem (see Section 1.5.2 [18]). The equilibrium equation $\mathbf{D}^*\boldsymbol{\sigma} = \mathbf{f}$ must be an invariant under certain coordinates transformation, say the law of angular momentum conservation, which is guaranteed by the objectivity of the stored energy $G(\mathbf{D}\boldsymbol{\chi})$ in continuum mechanics (see Definition 6.1.2, [18]), or by the isotropy of the kinetic energy $T(\dot{\boldsymbol{\chi}})$ in Lagrangian mechanics [52]. Specifically, the equilibrium equation for Newtonian mechanics is an invariant under the *Galilean transformation*; while for Einstein's special relativity theory, the equilibrium equation $\mathbf{D}^*\boldsymbol{\sigma} = \mathbf{f}$ is an invariant under the *Lorentz transformation*. For linear equilibrium equation, the quadratic $G(\mathbf{g})$ is naturally an objective function for convex systems. Unfortunately, since the concept of the objectivity is misused in optimization and the notation of the Euclidian

coordinate $\mathbf{x} = \{x_i\}$ is used as the unknown, the Lagrange multiplier method and the associated augmented methods have been mistakenly used for solving general nonconvex optimization problems, which produces many artificial duality gaps [53]. \clubsuit

2.7 Bi-Duality and Conceptual Mistakes

For convex Hamiltonian systems, the action $G(\mathbf{D}\boldsymbol{\chi})$ in (10) is a d.c. (difference of convex) functional and the Lagrangian has its standard form in Lagrangian mechanics (see Section 2.5.2 [18] with $\mathbf{q}(t) = \boldsymbol{\chi}$ and $\mathbf{p} = \boldsymbol{\sigma}$):

$$L(\mathbf{q}, \mathbf{p}) = \langle \dot{\mathbf{q}}; \mathbf{p} \rangle - \int_I [T^*(\mathbf{p}) + U(\mathbf{q})] dt - \langle \mathbf{q}, \mathbf{f} \rangle, \quad (17)$$

where $\mathbf{q} \in \mathcal{X}_a \subset \mathcal{C}^1[I, \mathbb{R}^n]$ is the Lagrange coordinate and $\mathbf{p} \in \mathcal{S}_a \subset \mathcal{C}[I, \mathbb{R}^n]$ is the momentum. In this case, the Lagrangian is a bi-concave functional on $\mathcal{X}_a \times \mathcal{S}_a$, but the Hamiltonian:

$$H(\mathbf{q}, \mathbf{p}) = \langle \mathbf{D}\mathbf{q}; \mathbf{p} \rangle - L(\mathbf{q}, \mathbf{p}) = \int_I [T^*(\mathbf{p}) + U(\mathbf{q})] dt \quad (18)$$

is convex.⁹ The total action and its canonical dual are [18]

$$\Pi(\mathbf{q}) = \max\{L(\mathbf{q}, \mathbf{p}) \mid \mathbf{p} \in \mathcal{V}_a^*\} = \int_I [T(\dot{\mathbf{q}}) - U(\mathbf{q})] dt - \langle \mathbf{q}, \mathbf{f} \rangle \quad \forall \mathbf{q} \in \mathcal{X}_c \quad (19)$$

$$\Pi^d(\mathbf{p}) = \max\{L(\mathbf{q}, \mathbf{p}) \mid \mathbf{q} \in \mathcal{X}_a\} = \int_I [U^*(\dot{\mathbf{p}}) - T^*(\mathbf{p})] dt \quad \forall \mathbf{p} \in \mathcal{S}_c \quad (20)$$

Clearly, both Π and Π^d are d.c. functionals. In this case, the so-called *bi-duality* was first presented in author's book Chapter 2 [18]:

Theorem 1 (Bi-Duality Theory) *For a given convex Hamiltonian system, if $(\bar{\mathbf{q}}, \bar{\mathbf{p}})$ is a critical point of $L(\mathbf{q}, \mathbf{p})$ over the time interval $I \subset \mathbb{R}$, then we have*

$$\Pi(\bar{\mathbf{q}}) = \min \Pi(\mathbf{q}) \Leftrightarrow \min \Pi^d(\mathbf{p}) = \Pi^d(\bar{\mathbf{p}}) \quad (21)$$

$$\Pi(\bar{\mathbf{q}}) = \max \Pi(\mathbf{q}) \Leftrightarrow \max \Pi^d(\mathbf{p}) = \Pi^d(\bar{\mathbf{p}}). \quad (22)$$

The mathematical proof of this theory was given in Section 2.6 [18] for convex Hamiltonian systems and in Corollary 5.3.6 [18] for d.c. programming problems.

⁹This is the reason that instead of the Lagrangian, the Hamiltonian is extensively used in dynamics.

This bi-duality revealed not only an interesting dynamical extremum principle in periodic motion but also an important truth in convex Hamiltonian systems.

Remark 2 (Least Action Principle and Conceptual Mistakes) The least action principle plays a central role in physics from the classical Newtonian mechanics, general relativity (Einstein-Hilbert action), to the modern string theory. Credit for the formulation of this principle is commonly given to Pierre Louis Maupertuis, who felt that “Nature is thrifty in all its actions.” It was historically called “least” because its solution requires finding the path that has the least value [9], say Fermat’s principle in optics. However, in Hamiltonian systems it should be accurately called the principle of stationary action since its solution does not minimize the total action. Actually, the validity of the least action principle has remained obscure in physics for several centuries. As a footnote in their celebrated book (Section 1.2, [52]), Landau and Lifshitz pointed out that the least action principle holds only for a sufficient small time interval, not for the whole trajectory of the system. Unfortunately, this is not true in general since the total action could be a concave functional within a sufficient small time interval.

Theorem 1 shows that a convex Hamiltonian system is controlled by the bi-duality, which revealed the following truths (see page 77 [18]):

*The least action principle is not valid for any periodic motion.
It holds for the whole trajectory of the system if the potential $U(\mathbf{q}) = 0$.*

The bi-duality theory has been challenged by M.D. Voisei, C. Zălinescu, and his former student R. Strugariu in a paper published in a dynamical systems journal [78]. Instead of finding any possible mistakes in author’s work, they created an artificial “Lagrangian”:

$$L(x, y) := -\frac{1}{2}\alpha\|x\|^2 - \frac{1}{2}\beta\|y\|^2 + \langle a, x \rangle \langle b, y \rangle, \quad (\text{Equation (1) in [78]})$$

and the associated “total action” $f(x)$ as well as its “dual action” $g(y)$:

$$f(x) = \max\{L(x, y) \mid y \in Y\} = -\frac{1}{2}\alpha\|x\|^2 + \frac{1}{2}\beta^{-1}\langle a, x \rangle^2\|b\|^2 \quad \forall x \in X$$

$$g(y) = \max\{L(x, y) \mid x \in X\} = -\frac{1}{2}\beta\|y\|^2 + \frac{1}{2}\alpha^{-1}\langle b, y \rangle^2\|a\|^2 \quad \forall y \in Y$$

By using these elementary functions in linear algebra, they produced a series of strange counterexamples to against the bi-duality theory in convex Hamiltonian systems presented by the author in Chapter 2 [18]. They claimed: “Because our counter-examples are very simple, using quadratic functions defined on whole Hilbert (even finite-dimensional) spaces, it is difficult to reinforce the hypotheses of the above mentioned results in order to keep the same conclusions and not obtain trivialities.”

Clearly, the quadratic function $L(x, y)$ created by Zălinescu *et al.* is totally irrelevant to the Lagrangian $L(\mathbf{q}, \mathbf{p})$ in Lagrangian mechanics and in Gao’s book

[18]. Without the differential operator $\mathbf{D} = \partial_t$, the quadratic d.c. function $f(x)$ (or $g(y)$) is defined on one-scale space X (or Y) and is unbounded. Therefore, its critical point does not produce any motion. This basic mistake shows that these people don't have necessary knowledge not only in Lagrangian mechanics (the time derivative $\mathbf{D} = \partial_t$ is necessary for any dynamical systems) but also in d.c. programming (unconstrained quadratic d.c. programming does not make any sense). It also shows that these people even don't know what the Lagrangian coordinate is, otherwise, they would never use a time-independent vector $x \in \mathbb{R}^n$ as an unknown in dynamical systems.

Moreover, since there is neither input in $L(x, y)$ nor initial/boundary conditions in X , all the counterexamples produced by Zălinescu *et al.* are simply not problems but only artificial "models." Since they don't follow the basic rules in mathematical modeling, such as the objectivity, symmetry, conservation, and constitutive laws, etc., these artificial "models" are very strange and even ugly (see Examples 3.3, 4.2, 4.4 [78]). This type of mistakes shows that these people don't know the difference between the modeling and problems. ♣

3 Unified Problem and Canonical Duality-Triality Theory

In this section, we simply restrict our discussion in finite-dimensional space \mathcal{X} . Its element $\chi \in \mathcal{X}$ could be a vector, a matrix, or a tensor.¹⁰ In this case, the linear operator \mathbf{D} is a generalized matrix¹¹ $\mathbf{D} : \mathcal{X} \rightarrow \mathcal{G}$ and \mathcal{G} is a generalized matrix space equipped with a natural norm $\|\mathbf{g}\|$. Let $\mathcal{X}_a \subset \mathcal{X}$ be a convex subset (with only linear constraints) and \mathcal{X}_a^* be its dual set such that for any given input $\bar{\chi}^* = \mathbf{f} \in \mathcal{X}_a^*$ the subjective function $\langle \chi, \mathbf{f} \rangle \geq 0 \quad \forall \chi \in \mathcal{X}_a$. Although the objectivity is necessary for real-world modeling, the numerical discretization of $\mathbf{D}\chi$ could lead to a complicated function $G(\mathbf{D}\chi)$ which may not be objective in $\mathbf{g} = \mathbf{D}\chi$. Also in operations research, many challenging problems are artificially proposed. Thus, the objectivity required in Gao and Strang's original work on nonlinear elasticity has been relaxed by the canonical duality since 2000 [20].

¹⁰Tensor is a geometrical object in mathematics and physics, which is defined as a multi-dimensional array satisfying a transformation law, see <https://en.wikipedia.org/wiki/Tensor>. A tensor must be independent of a particular choice of coordinate system (frame-invariance). But, this terminology has been also misused recent years in the optimization literature such that any multi-dimensional array of data is called tensor [4]. This mistake has been recognized recently in the preface of [67].

¹¹A generalized matrix $\mathbf{D} = \{D_{\alpha \dots \gamma}^{i \dots j}\}$ is a multi-dimensional array but not necessary to satisfy a transformation law, so it is not a tensor. In order to avoid confusion, it can be called a generalized matrix, or simply *gentrix*.

3.1 Canonical Transformation and Gap Function

Definition 3 (Canonical Function and Transformation)

A real-valued function $\Phi : \mathcal{E}_a \rightarrow \mathbb{R}$ is called a canonical function if its domain \mathcal{E}_a is convex and the duality relation $\xi^* = \partial\Phi(\xi) : \mathcal{E}_a \rightarrow \mathcal{E}_a^*$ is bijective.

For a given real-valued function $G : \mathcal{G}_a \rightarrow \mathbb{R}$, if there exist a mapping $\xi : \mathcal{G}_a \rightarrow \mathcal{E}_a$ and a canonical function $\Phi : \mathcal{E}_a \rightarrow \mathbb{R}$ such that

$$G(\mathbf{g}) = \Phi(\xi(\mathbf{g})) \quad \forall \mathbf{g} \in \mathcal{G}_a, \quad (23)$$

the transformation (23) is called the canonical transformation.

The canonical function is not necessary to be convex. Actually, in many real-world applications, $\Phi(\xi)$ is usually concave. For example, in differential geometry and finite deformation theory (see [18]) the objective function $G(\mathbf{g})$ is the deformation Jacobian:

$$G(\mathbf{g}) = \sqrt{\det(\mathbf{g}^T \mathbf{g})}, \quad \mathbf{g} = \nabla \chi. \quad (24)$$

If we chose $\xi = \det(\mathbf{g}^T \mathbf{g})$ as the geometrical measure, which is the third invariant of the *Riemannian metric tensor* $\mathbf{g}^T \mathbf{g}$ and usually denoted as I_3 , the canonical function $\Phi(I_3)$ is then a concave function of the scale measure $I_3(\mathbf{g})$.

The canonical duality is a fundamental principle in sciences and oriental philosophy, which underlies all natural phenomena. Therefore, instead of the objectivity in continuum physics, a generalized objective function $G(\mathbf{g})$ is used in the canonical duality theory under the following assumption.

Definition 4 (Generalized Objective Function) For a given real-valued function $G : \mathcal{G}_a \rightarrow \mathbb{R}$, if there exist a measure $\xi : \mathcal{G}_a \rightarrow \mathcal{E}_a$ and a canonical function $\Phi : \mathcal{E}_a \rightarrow \mathbb{R}$ such that the following conditions hold:

(D1) Positivity: $G(\mathbf{g}) \geq 0 \quad \forall \mathbf{g} \in \mathcal{G}_a$;

(D2) Canonicity: $G(\mathbf{g}) = \Phi(\xi(\mathbf{g})) \quad \forall \mathbf{g} \in \mathcal{G}_a$,

then $G : \mathcal{G}_a \rightarrow \mathbb{R}$ is called to be a generalized objective function.

The canonical transformation plays an important role in mathematical modeling and nonlinear analysis. Let $\Lambda = \xi \circ \mathbf{D} : \mathcal{X}_a \rightarrow \mathcal{E}_a$ be the so-called *geometrically admissible operator* and $\langle \xi; \zeta \rangle : \mathcal{E} \times \mathcal{E}^* \rightarrow \mathbb{R}$ be the bilinear form which puts \mathcal{E} and \mathcal{E}^* in duality. By (D2), we have $\mathcal{X}_c = \{\chi \in \mathcal{X}_a \mid \Lambda(\chi) \in \mathcal{E}_a\}$. The problem (\mathcal{P}) can be equivalently written in the following canonical form:

$$(\mathcal{P}_g) : \min \{\Pi(\chi) = \Phi(\Lambda(\chi)) - \langle \chi, \mathbf{f} \rangle \mid \chi \in \mathcal{X}_c\}. \quad (25)$$

By the facts that the canonical duality is a universal principle in nature, the canonical measure $\Lambda(\chi)$ is not necessarily to be objective, and the spaces \mathcal{X} , \mathcal{E} could be at

different physical scale with totally different dimensions, the canonical problem (\mathcal{P}_g) can be used to study general optimization problems in multi-scale complex systems. The criticality condition of (\mathcal{P}_g) is governed by the *fundamental principle of virtual work*:

$$\langle \Lambda_t(\boldsymbol{\chi})\delta\boldsymbol{\chi}; \boldsymbol{\varsigma} \rangle = \langle \delta\boldsymbol{\chi}, \Lambda_t^*(\boldsymbol{\chi})\boldsymbol{\varsigma} \rangle = \langle \delta\boldsymbol{\chi}, \mathbf{f} \rangle \quad \forall \delta\boldsymbol{\chi} \in \mathcal{X}_c, \quad (26)$$

where $\Lambda_t(\boldsymbol{\chi}) = \partial\Lambda(\boldsymbol{\chi})$ represents a generalized Gâteaux (or directional) derivative of $\Lambda(\boldsymbol{\chi})$, its adjoint Λ_t^* is called the balance operator, $\boldsymbol{\varsigma} = \boldsymbol{\xi}^*(\boldsymbol{\xi}) = \partial\Phi(\boldsymbol{\xi})$ and $\boldsymbol{\xi}^* : \mathcal{E}_a \rightarrow \mathcal{E}_a^*$ is a canonical dual (or constitutive) operator. The strong form of this virtual work principle is called *the canonical equilibrium equation*:

$$\mathbf{A}(\boldsymbol{\chi}) = \Lambda_t^*(\boldsymbol{\chi})\boldsymbol{\xi}^*(\Lambda(\boldsymbol{\chi})) = \mathbf{f}. \quad (27)$$

A system governed by this equation is called a *canonical system* and is denoted as (see Chapter 4, [18]):

$$\mathbb{S}_a = \{(\mathcal{X}_a, \mathcal{X}_a^*), (\mathcal{E}_a; \mathcal{E}_a^*); (\Lambda, \boldsymbol{\xi}^*)\}.$$

Definition 5 (Classification of Nonlinearities) The system \mathbb{S}_a is called *geometrically nonlinear* (resp., linear) if the geometrical operator $\Lambda : \mathcal{X}_a \rightarrow \mathcal{E}_a$ is nonlinear (resp., linear); the system is called *physically (or constitutively) nonlinear* (resp., linear) if the canonical dual operator $\boldsymbol{\xi}^* : \mathcal{E}_a \rightarrow \mathcal{E}_a^*$ is nonlinear (resp., linear); the system is called *fully nonlinear* (resp., linear) if it is both geometrically and physically nonlinear (resp., linear).

Both geometrical and physics nonlinearities are basic concepts in nonlinear field theory. The mathematical definition was first given by the author in 2000 under the canonical transformation [20]. A diagrammatic representation of this canonical system is shown in Figure 1.

This diagram shows a symmetry broken in the canonical equilibrium equation, i.e., instead of Λ^* , the balance operator is Λ_t^* . It was discovered by Gao and Strang [43] that by introducing a complementary operator $\Lambda_c(\boldsymbol{\chi}) = \Lambda(\boldsymbol{\chi}) - \Lambda_t(\boldsymbol{\chi})\boldsymbol{\chi}$, this locally broken symmetry is recovered by a so-called complementary gap function:

$$\begin{array}{ccc} \boldsymbol{\chi} \in \mathcal{X}_a & \xrightarrow{\Lambda} & \mathcal{X}_a^* \ni \boldsymbol{\chi}^* \\ & \langle \boldsymbol{\chi}, \boldsymbol{\chi}^* \rangle & \\ \Lambda_t + \Lambda_c = \Lambda & \downarrow & \uparrow \Lambda_t^* = (\Lambda - \Lambda_c)^* \\ \boldsymbol{\xi} \in \mathcal{E}_a & \xrightarrow{\langle \boldsymbol{\xi}; \boldsymbol{\varsigma} \rangle} & \mathcal{E}_a^* \ni \boldsymbol{\varsigma} \\ & \partial\Phi(\boldsymbol{\xi}) & \end{array}$$

Fig. 1 Diagrammatic representation for a canonical system

$$G_{ap}(\boldsymbol{\chi}, \boldsymbol{\varsigma}) = \langle -\Lambda_c(\boldsymbol{\chi}); \boldsymbol{\varsigma} \rangle, \quad (28)$$

which plays a key role in global optimization and the triality theory. Clearly, if $\Lambda = \mathbf{D}$ is linear, then $G_{ap} = 0$. Thus, the following statement is important to understand complexity:

Only the geometrical nonlinearity leads to nonconvexity in optimization, bifurcation in analysis, chaos in dynamics, and NP-hard problems in complex systems.

3.2 Complementary-Dual Principle and Analytical Solution

For a given canonical function $\Phi : \mathcal{E}_a \rightarrow \mathbb{R}$, its conjugate $\Phi^* : \mathcal{E}_a^* \rightarrow \mathbb{R}$ can be uniquely defined by the Legendre transformation:

$$\Phi^*(\boldsymbol{\varsigma}) = \text{sta}\{\langle \boldsymbol{\xi}; \boldsymbol{\varsigma} \rangle - \Phi(\boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \mathcal{E}_a\}, \quad (29)$$

where $\text{sta}\{f(\boldsymbol{\chi}) \mid \boldsymbol{\chi} \in \mathcal{X}\}$ stands for finding the stationary value of $f(\boldsymbol{\chi})$ on \mathcal{X} , and the following canonical duality relations hold on $\mathcal{E}_a \times \mathcal{E}_a^*$:

$$\boldsymbol{\varsigma} = \partial\Phi(\boldsymbol{\varepsilon}) \Leftrightarrow \boldsymbol{\varepsilon} = \partial\Phi^*(\boldsymbol{\varsigma}) \Leftrightarrow \Phi(\boldsymbol{\varepsilon}) + \Phi^*(\boldsymbol{\varsigma}) = \langle \boldsymbol{\varepsilon}; \boldsymbol{\varsigma} \rangle. \quad (30)$$

If the canonical function is convex and lower semicontinuous, the Gâteaux derivative ∂ should be replaced by the sub-differential and Φ^* is replaced by the Fenchel conjugate $\Phi^\sharp(\boldsymbol{\varsigma}) = \sup\{\langle \boldsymbol{\xi}; \boldsymbol{\varsigma} \rangle - \Phi(\boldsymbol{\xi}) \mid \boldsymbol{\xi} \in \mathcal{E}_a\}$. In this case, (30) is replaced by the generalized canonical duality:

$$\boldsymbol{\varsigma} \in \partial\Phi(\boldsymbol{\varepsilon}) \Leftrightarrow \boldsymbol{\varepsilon} \in \partial\Phi^\sharp(\boldsymbol{\varsigma}) \Leftrightarrow \Phi(\boldsymbol{\varepsilon}) + \Phi^\sharp(\boldsymbol{\varsigma}) = \langle \boldsymbol{\varepsilon}; \boldsymbol{\varsigma} \rangle \quad \forall (\boldsymbol{\xi}, \boldsymbol{\varsigma}) \in \mathcal{E}_a \times \mathcal{E}_a^*. \quad (31)$$

If the convex set \mathcal{E}_a contains inequality constraints, then (31) includes all the *internal KKT conditions* [14, 53]. In this sense, a KKT point of the canonical form $\Pi(\boldsymbol{\chi})$ is a generalized critical point of $\Pi(\boldsymbol{\chi})$.

By the complementarity $\Phi(\Lambda(\boldsymbol{\chi})) = \langle \Lambda(\boldsymbol{\chi}); \boldsymbol{\varsigma} \rangle - \Phi^*(\boldsymbol{\varsigma})$, the canonical form of $\Pi(\boldsymbol{\chi})$ can be equivalently written in Gao and Strang's *total complementary function* $\Xi : \mathcal{X}_a \times \mathcal{E}_a^* \rightarrow \mathbb{R}$ [43]:

$$\Xi(\boldsymbol{\chi}, \boldsymbol{\varsigma}) = \langle \Lambda(\boldsymbol{\chi}); \boldsymbol{\varsigma} \rangle - \Phi^*(\boldsymbol{\varsigma}) - \langle \boldsymbol{\chi}, \mathbf{f} \rangle. \quad (32)$$

Then, the canonical dual function $\Pi^g : \mathcal{S}_c \rightarrow \mathbb{R}$ can be obtained by the *canonical dual transformation*:

$$\Pi^g(\boldsymbol{\chi}) = \text{sta}\{\Xi(\boldsymbol{\chi}, \boldsymbol{\varsigma}) \mid \boldsymbol{\chi} \in \mathcal{X}_a\} = G_{ap}^\Lambda(\boldsymbol{\varsigma}) - \Phi^*(\boldsymbol{\varsigma}), \quad (33)$$

where $G_{ap}^\Lambda(\boldsymbol{\varsigma}) = \text{sta}\{\langle \Lambda(\boldsymbol{\chi}); \boldsymbol{\varsigma} \rangle - \langle \boldsymbol{\chi}, \mathbf{f} \rangle \mid \boldsymbol{\chi} \in \mathcal{X}_a\}$, which is defined on the canonical dual feasible space $\mathcal{S}_c = \{\boldsymbol{\varsigma} \in \mathcal{E}_a^* \mid \Lambda_t^*(\boldsymbol{\chi})\boldsymbol{\varsigma} = \mathbf{f} \ \forall \boldsymbol{\chi} \in \mathcal{X}_a\}$. Clearly, $\mathcal{S}_c \neq \emptyset$ if (\mathcal{P}) is properly posed.

Theorem 2 (Complementary-Dual Principle [18]) *The pair $(\bar{\boldsymbol{\chi}}, \bar{\boldsymbol{\varsigma}})$ is a critical point of $\Xi(\boldsymbol{\chi}, \boldsymbol{\varsigma})$ if and only if $\bar{\boldsymbol{\chi}}$ is a critical point of $\Pi(\boldsymbol{\chi})$ and $\bar{\boldsymbol{\varsigma}}$ is a critical point of $\Pi^g(\boldsymbol{\varsigma})$. Moreover:*

$$\Pi(\bar{\boldsymbol{\chi}}) = \Xi(\bar{\boldsymbol{\chi}}, \bar{\boldsymbol{\varsigma}}) = \Pi^g(\bar{\boldsymbol{\varsigma}}). \quad (34)$$

Proof The criticality condition $\partial \Xi(\bar{\boldsymbol{\chi}}, \bar{\boldsymbol{\varsigma}}) = 0$ leads to the following canonical equations:

$$\Lambda(\bar{\boldsymbol{\chi}}) = \partial \Phi^*(\bar{\boldsymbol{\varsigma}}), \quad \Lambda_t^*(\bar{\boldsymbol{\chi}})\bar{\boldsymbol{\varsigma}} = \mathbf{f}. \quad (35)$$

The theorem is proved by the canonical duality (30) and the definition of Π^g . \square

Theorem 2 shows a one-to-one correspondence of the critical points between the primal function and its canonical dual. In large deformation theory, this theorem solved the fifty-year-old open problem on complementary variational principle and is known as the Gao principle in literature [58]. In real-world applications, the geometrical operator Λ is usually *quadratic homogeneous*, i.e., $\Lambda(\alpha \boldsymbol{\chi}) = \alpha^2 \Lambda(\boldsymbol{\chi}) \ \forall \alpha \in \mathbb{R}$. In this case, we have [43] $\Lambda_t(\boldsymbol{\chi})\boldsymbol{\chi} = 2\Lambda(\boldsymbol{\chi})$, $\Lambda_c(\boldsymbol{\chi}) = -\Lambda(\boldsymbol{\chi})$, and

$$\Xi(\boldsymbol{\chi}, \boldsymbol{\varsigma}) = G_{ap}(\boldsymbol{\chi}, \boldsymbol{\varsigma}) - \Phi^*(\boldsymbol{\varsigma}) - \langle \boldsymbol{\chi}, \mathbf{f} \rangle = \frac{1}{2} \langle \boldsymbol{\chi}, \mathbf{G}(\boldsymbol{\varsigma})\boldsymbol{\chi} \rangle - \Phi^*(\boldsymbol{\varsigma}) - \langle \boldsymbol{\chi}, \mathbf{f} \rangle, \quad (36)$$

where $\mathbf{G}(\boldsymbol{\varsigma}) = \partial_{\boldsymbol{\chi}}^2 G_{ap}(\boldsymbol{\chi}, \boldsymbol{\varsigma})$. Then, the canonical dual function $\Pi^g(\boldsymbol{\varsigma})$ can be written explicitly as:

$$\Pi^g(\boldsymbol{\varsigma}) = \{\Xi(\boldsymbol{\chi}, \boldsymbol{\varsigma}) \mid \mathbf{G}(\boldsymbol{\varsigma})\boldsymbol{\chi} = \mathbf{f} \ \forall \boldsymbol{\chi} \in \mathcal{X}_a\} = -\frac{1}{2} \langle [\mathbf{G}(\boldsymbol{\varsigma})]^+ \mathbf{f}, \mathbf{f} \rangle - \Phi^*(\boldsymbol{\varsigma}), \quad (37)$$

where \mathbf{G}^+ represents a generalized inverse of \mathbf{G} .

Theorem 3 (Analytical Solution Form [18]) *If $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_c$ is a critical point of $\Pi^g(\boldsymbol{\varsigma})$, then:*

$$\bar{\boldsymbol{\chi}} = [\mathbf{G}(\bar{\boldsymbol{\varsigma}})]^+ \mathbf{f} \quad (38)$$

is a critical point of $\Pi(\boldsymbol{\chi})$ and $\Pi(\bar{\boldsymbol{\chi}}) = \Xi(\bar{\boldsymbol{\chi}}, \bar{\boldsymbol{\varsigma}}) = \Pi^g(\bar{\boldsymbol{\varsigma}})$. Dually, if $\bar{\boldsymbol{\chi}} \in \mathcal{X}_c$ is a critical point of $\Pi(\boldsymbol{\chi})$, it must be in the form of (38) for a critical point $\bar{\boldsymbol{\varsigma}} \in \mathcal{S}_c$ of $\Pi^g(\boldsymbol{\varsigma})$.

This unified analytical solution form holds not only for general global optimization problems in finite-dimensional systems [25] but also for a large class

of nonlinear boundary-/initial-value problems in nonconvex analysis and dynamic systems [21, 23, 54].

3.3 Triality Theory and NP-Hard Criterion

In order to study extremality property for the general problem (\mathcal{P}_g) , we need additional assumptions for the generalized objective function $G(\mathbf{D}\chi)$.

Assumption 1 (Canonically Convex Function) *Let $G(\mathbf{D}\chi)$ be a generalized objective function, i.e., there exist a measure $\Lambda : \mathcal{X}_a \rightarrow \mathcal{E}_a$ and a canonical function $\Phi : \mathcal{E}_a \rightarrow \mathbb{R}$ such that $G(\mathbf{D}\chi) = \Phi(\Lambda(\chi))$. We assume:*

- (A1) *Nonlinearity: $\Lambda(\chi) : \mathcal{X}_a \rightarrow \mathcal{E}_a$ is a quadratic measure,*
- (A2) *Regularity: $\Phi(\Lambda(\chi))$ is twice continuously differentiable for all $\chi \in \mathcal{X}_a$,*
- (A3) *Convexity: $\Phi : \mathcal{E}_a \rightarrow \mathbb{R}$ is strictly convex.*

One should emphasize that although $\Phi(\xi)$ is required to be strictly convex on \mathcal{E}_a in this section, the composition $\Phi(\Lambda(\chi))$ is usually nonconvex on \mathcal{X}_a due to the geometrical nonlinearity. The canonical duality theory presented in this section can be generalized to the problems governed by high-order nonlinear $\Lambda\chi$ and canonically concave function $\Phi(\xi)$ (see [18, 22, 46, 56]).

Definition 6 (Degenerate and Nondegenerate Critical Points, Morse Function)

Let $\bar{\chi} \in \mathcal{X}_c$ be a critical point of a real-valued function $\Pi : \mathcal{X}_c \rightarrow \mathbb{R}$. $\bar{\chi}$ is called degenerate (resp. nondegenerate) if the Hessian matrix of $\Pi(\chi)$ is singular (resp., nonsingular) at $\bar{\chi}$. The function $\Pi : \mathcal{X}_c \rightarrow \mathbb{R}$ is called a Morse function if it has no degenerate critical points.

Theorem 4 (Triality Theory [20]) *Suppose that $\Phi : \mathcal{E}_a \rightarrow \mathbb{R}$ is convex, $(\bar{\chi}, \bar{\varsigma})$ is a nondegenerate critical point of $\Xi(\chi, \varsigma)$, and $\mathcal{X}_o \times \mathcal{S}_o$ is a neighborhood¹² of $(\bar{\chi}, \bar{\varsigma})$.*

If $\bar{\varsigma} \in \mathcal{S}_c^+ = \{\varsigma \in \mathcal{S}_c \mid \mathbf{G}(\varsigma) \geq 0\}$, then

$$\Pi(\bar{\chi}) = \min_{\chi \in \mathcal{X}_c} \Pi(\chi) = \max_{\varsigma \in \mathcal{S}_c^+} \Pi^g(\varsigma) = \Pi^g(\bar{\varsigma}). \tag{39}$$

If $\bar{\varsigma} \in \mathcal{S}_c^- = \{\varsigma \in \mathcal{S}_c \mid \mathbf{G}(\varsigma) < 0\}$, then we have either

$$\Pi(\bar{\chi}) = \max_{\chi \in \mathcal{X}_o} \Pi(\chi) = \max_{\varsigma \in \mathcal{S}_o} \Pi^g(\varsigma) = \Pi^g(\bar{\varsigma}), \tag{40}$$

or (if $\dim \Pi = \dim \Pi^g$)

$$\Pi(\bar{\chi}) = \min_{\chi \in \mathcal{X}_o} \Pi(\chi) = \min_{\varsigma \in \mathcal{S}_o} \Pi^g(\varsigma) = \Pi^g(\bar{\varsigma}). \tag{41}$$

¹²The neighborhood \mathcal{X}_o of $\bar{\chi}$ means that on which, $\bar{\chi}$ is the only critical point (see page 140 [18]).

The statement (39) is the so-called *canonical min-max duality*, which can be proved easily by Gao and Strang's work in 1989 [43]. Clearly, $\boldsymbol{\zeta} \in \mathcal{S}_c^+$ if and only if $G_{ap}(\boldsymbol{\chi}, \boldsymbol{\zeta}) \geq 0 \quad \forall \boldsymbol{\chi} \in \mathcal{X}$. This duality theory shows that the Gao-Strang gap function provides a global optimum criterion. The statements (40) and (41) are called the *canonical double-max* and *double-min dualities*, respectively, which can be used to find local extremum solutions. The triality theory shows that the nonconvex minimization problem (\mathcal{P}) is canonically dual to the following maximum stationary problem:

$$(\mathcal{P}^d) : \max \text{sta}\{\Pi^g(\boldsymbol{\zeta}) \mid \boldsymbol{\zeta} \in \mathcal{S}_c^+\}. \quad (42)$$

Theorem 5 (Existence and Uniqueness Criteria [25]) *For a properly posed (\mathcal{P}), if the canonical function $\Phi : \mathcal{E}_a \rightarrow \mathbb{R}$ is convex, $\text{int}\mathcal{S}_c^+ \neq \emptyset$, and*

$$\lim_{\alpha \rightarrow 0^+} \Pi^g(\boldsymbol{\zeta}_o + \alpha \boldsymbol{\zeta}) = -\infty \quad \forall \boldsymbol{\zeta}_o \in \partial\mathcal{S}_c^+, \quad \forall \boldsymbol{\zeta} \in \mathcal{S}_c^+, \quad (43)$$

then (\mathcal{P}^d) has at least one solution $\bar{\boldsymbol{\zeta}} \in \mathcal{S}_c^+$ and $\bar{\boldsymbol{\chi}} = [\mathbf{G}(\bar{\boldsymbol{\zeta}})]^+ \mathbf{f}$ is a solution to (\mathcal{P}). The solution is unique if $\mathbf{H} = \mathbf{G}(\bar{\boldsymbol{\zeta}}) > 0$.

Proof Under the required conditions, $-\Pi^g : \mathcal{S}_c^+ \rightarrow \mathbb{R}$ is convex and coercive and $\text{int}\mathcal{S}_c^+ \neq \emptyset$. Therefore, (\mathcal{P}^g) has at least one solution. If $\mathbf{H} > 0$, then $\Pi^g : \mathcal{S}_c^+ \rightarrow \mathbb{R}$ is strictly concave and (\mathcal{P}^g) has a unique solution. \square

This theorem shows that if $\text{int}\mathcal{S}_c^+ \neq \emptyset$ the nonconvex problem (\mathcal{P}_g) is canonically dual to (\mathcal{P}^g) which can be solved easily. Otherwise, the problem (\mathcal{P}_g) is canonically dual to the following minimal stationary problem, i.e., to find a global minimum stationary value of Π^g on \mathcal{S}_c :

$$(\mathcal{P}^s) : \min \text{sta}\{\Pi^g(\boldsymbol{\zeta}) \mid \boldsymbol{\zeta} \in \mathcal{S}_c\}, \quad (44)$$

which could be really NP-hard since $\Pi^g(\boldsymbol{\zeta})$ is nonconvex on the nonconvex set \mathcal{S}_c . Therefore, a conjecture was proposed in [24].

Conjecture 1 (Criterion of NP-Hardness) *A properly posed problem (\mathcal{P}_g) is NP-hard only if $\text{int}\mathcal{S}_c^+ = \emptyset$.*

Remark 3 (History of Triality and Challenges) The triality theory was discovered by the author during his research on post-buckling of a large deformed elastic beam in 1996 [12], where the primal variable $\mathbf{u}(\mathbf{x})$ is a displacement vector in \mathbb{R}^2 and $\boldsymbol{\zeta}(\mathbf{x})$ is a canonical dual stress also in \mathbb{R}^2 . Therefore, the triality theory was correctly proposed in nonconvex analysis, which provides for the first time a complete set of solutions to the post-buckling problem. Physically, the global minimizer $\bar{\mathbf{u}}(\mathbf{x})$ represents a stable buckled beam configuration (happened naturally), the local minimizer is an unstable buckled state (happened occasionally), while the local maximizer is the unbuckled beam state. Mathematical proof of the triality theory was given in [18] for one-D nonconvex variational problems (Theorem 2.6.2) and

for finite-dimensional optimization problems (Theorem 5.3.6 and Corollary 5.3.1). In 2002, the author discovered some counterexamples to the canonical double-min duality when $\dim \Pi \neq \dim \Pi^g$. Therefore, the triality theory was presented in an “either-or” form since the double-max duality is always true but the double-min duality was remarked by certain additional condition (see Remark 1 in [22] and Remark for Theorem 3 in [23]). Recently, the author and his co-workers proved that the canonical double-min duality holds weakly when $\dim \Pi \neq \dim \Pi^g$ [6, 61]. It was also discovered by using the canonical dual finite element method that the local minimum solutions in nonconvex mechanics are very sensitive not only to the input and boundary conditions of a given system but also to such artificial conditions as the numerical discretization and computational precision, etc. The triality theory provides a precise mathematical tool for studying and understanding complicated natural phenomena.

The triality theory has been repeatedly challenged by M.D. Voisei and C. Zălinescu in a set of at least 11 papers (see [29]). These papers fall into three groups. In the first group (say [78, 83]), they oppositely choose piecewise linear functions for G and quadratic functions for F as counterexamples to against the canonical duality theory with six conclusions on Gao and Strang’s original work including [83]: “About the (complementary) gap function one can conclude that it is useless at least in the current context. The hope for reading an optimization theory with diverse applications is ruined . . .” Clearly, they made conceptual mistakes. In the second group, Voisei and Zălinescu chose an artificial problem with certain symmetry such that $S_a^+ = \emptyset$. Such a problem can be solved easily by linear perturbation (see [62]). The counterexamples in the third group are simply those such that $\dim \Pi \neq \dim \Pi^g$. This type of counterexamples were first discovered by Gao in 2002 so it was emphasized in [22, 23] that the canonical double-min duality holds under certain additional constraints (see Remark 1 in [22] and Remark for Theorem 3 in [23]). But neither [23] nor [22] was cited by Zălinescu and his co-authors in their papers. Honest people can easily understand the motivation of these challenges.

The canonical duality-triality theory has been successfully used for solving a wide class problems in both global optimization and nonconvex analysis [39], including certain challenging problems in nonconvex analysis [19], nonlinear PDEs [33], large deformation mechanics [27], and NP-hard integer programming problems [24, 31]. ♣

4 Applications in Complex Systems

Applications to nonconvex constrained global optimization have been discussed in [40, 53]. This section presents applications to two general global optimization problems.

4.1 Unconstrained Nonconvex Optimization Problem

$$(\mathcal{P}_g): \quad \min \left\{ \Pi(\boldsymbol{\chi}) = \sum_{s=1}^m \Phi_s(\Lambda_s(\boldsymbol{\chi})) - \langle \boldsymbol{\chi}, \mathbf{f} \rangle \mid \boldsymbol{\chi} \in \mathcal{X}_c \right\}, \quad (45)$$

where the canonical measures $\boldsymbol{\xi}_s = \Lambda_s(\boldsymbol{\chi})$ could be either a scalar or a generalized matrix, $\Phi_k(\boldsymbol{\xi}_k)$ are any given canonical functions, such as polynomial, exponential, logarithm, and their compositions, etc. For example, if $\boldsymbol{\chi} \in \mathcal{X}_c \subset \mathbb{R}^n$ and

$$\begin{aligned} G(\mathbf{D}\boldsymbol{\chi}) &= \sum_{i \in \mathbb{I}} \frac{1}{2} \alpha_i \boldsymbol{\chi}^T \mathbf{Q}_i \boldsymbol{\chi} + \sum_{j \in \mathbb{J}} \frac{1}{2} \alpha_j \left(\frac{1}{2} \boldsymbol{\chi}^T \mathbf{Q}_j \boldsymbol{\chi} + \beta_j \right)^2 \\ &\quad + \sum_{k \in \mathbb{K}} \alpha_k \exp \left(\frac{1}{2} \boldsymbol{\chi}^T \mathbf{Q}_k \boldsymbol{\chi} \right) + \sum_{\ell \in \mathbb{L}} \frac{1}{2} \alpha_\ell \boldsymbol{\chi}^T \mathbf{Q}_\ell \boldsymbol{\chi} \log \left(\frac{1}{2} \boldsymbol{\chi}^T \mathbf{Q}_\ell \boldsymbol{\chi} \right), \end{aligned} \quad (46)$$

where $\{\mathbf{Q}_s\}$ are positive-definite matrices to allow the Cholesky decomposition $\mathbf{Q}_s = \mathbf{D}_s^T \mathbf{D}_s$ for all $s \in \{\mathbb{I}, \mathbb{J}, \mathbb{K}, \mathbb{L}\}$ and $\{\alpha_s, \beta_s\}$ are physical constants, which could be either positive or negative under Assumption 1. This general function includes naturally the so-called d.c. functions (i.e., difference of convex functions). Let $p = \dim \mathbb{I}$, $q = \dim \mathbb{J} + \dim \mathbb{K} + \dim \mathbb{L}$. By using the canonical measure:

$$\boldsymbol{\xi} = \{\xi_s\} = \left\{ \frac{1}{2} \alpha_i \boldsymbol{\chi}^T \mathbf{Q}_i \boldsymbol{\chi}, \frac{1}{2} \boldsymbol{\chi}^T \mathbf{Q}_r \boldsymbol{\chi} \right\} \in \mathcal{E}_a = \mathbb{R}^p \times \mathbb{R}_+^q,$$

where $\mathbb{R}_+^q = \{\mathbf{x} \in \mathbb{R}^q \mid x_i \geq 0 \forall i = 1, \dots, q\}$, $G(\mathbf{g})$ can be written in the canonical form:

$$\Phi(\boldsymbol{\xi}) = \sum_{i \in \mathbb{I}} \xi_i + \sum_{j \in \mathbb{J}} \frac{1}{2} \alpha_j (\xi_j + \beta_j)^2 + \sum_{k \in \mathbb{K}} \alpha_k \exp \xi_k + \sum_{\ell \in \mathbb{L}} \alpha_\ell \xi_\ell \log \xi_\ell.$$

Thus, $\partial \Phi(\boldsymbol{\xi}) = \{1, \varsigma_r\}$ in which $\boldsymbol{\varsigma} = \{\alpha_j (\xi_j + \beta_j), \alpha_k \exp \xi_k, \alpha_\ell (\log \xi_\ell - 1)\} \in \mathcal{E}_a^*$, and

$$\mathcal{E}_a^* = \{\boldsymbol{\varsigma} \in \mathbb{R}^q \mid \varsigma_j \geq -\alpha_j \beta_j \quad \forall j \in \mathbb{J}, \quad \varsigma_k \geq \alpha_k \quad \forall k \in \mathbb{K}, \quad \varsigma_\ell \in \mathbb{R} \quad \forall \ell \in \mathbb{L}\}.$$

The conjugate of Φ can be easily obtained as:

$$\Phi^*(\boldsymbol{\varsigma}) = \sum_{j \in \mathbb{J}} \left(\frac{1}{2\alpha_j} \varsigma_j^2 + \beta_j \varsigma_j \right) + \sum_{k \in \mathbb{K}} \varsigma_k (\ln(\alpha_k^{-1} \varsigma_k) - 1) + \sum_{\ell \in \mathbb{L}} \alpha_\ell \exp(\alpha_\ell^{-1} \varsigma_\ell - 1). \quad (47)$$

Since $\Lambda(\chi)$ is quadratic homogenous, the gap function and G_{ap}^Λ in this case are

$$G_{ap}(\chi, \varsigma) = \frac{1}{2} \chi^T \mathbf{G}(\varsigma) \chi, \quad G_{ap}^\Lambda(\varsigma) = \frac{1}{2} \mathbf{f}^T [\mathbf{G}(\varsigma)]^+ \mathbf{f},$$

$$\mathbf{G}(\varsigma) = \sum_{i \in \mathbb{I}} \alpha_i \mathbf{Q}_i + \sum_{s \in \{\mathbb{J}, \mathbb{K}, \mathbb{L}\}} \varsigma_s \mathbf{Q}_s.$$

Since $\Pi^g(\varsigma) = -G_{ap}^\Lambda(\varsigma) - \Phi^*(\varsigma)$ is concave and \mathcal{S}_c^+ is a closed convex set, if for the given physical constants and the input \mathbf{f} such that $\mathcal{S}_c^+ \neq \emptyset$, the canonical dual problem (\mathcal{P}^g) has at least one solution $\bar{\varsigma} \in \mathcal{S}_c^+ \subset \mathbb{R}^q$ and $\bar{\chi} = [\mathbf{G}(\bar{\varsigma})]^+ \mathbf{f} \in \mathcal{X}_c \subset \mathbb{R}^n$ is a global minimum solution to (\mathcal{P}). If $n \gg q$, the problem (\mathcal{P}^g) can be much easier than (\mathcal{P}).

4.2 D.C. Programming

It is known that in Euclidean space every continuous global optimization problem on a compact set can be reformulated as a d.c. optimization problem, i.e., a nonconvex problem which can be described in terms of *d.c. functions* (difference of convex functions) and *d.c. sets* (difference of convex sets) [82]. By the fact that any constraint set can be equivalently relaxed by a nonsmooth indicator function, general nonconvex optimization problems can be written in the following standard d.c. programming form:

$$\min \{ f(\mathbf{x}) = g(\mathbf{x}) - h(\mathbf{x}) \mid \forall \mathbf{x} \in \mathcal{X} \}, \quad (48)$$

where $\mathcal{X} = \mathbb{R}^n$, $g(\mathbf{x}), h(\mathbf{x})$ are convex proper lower semicontinuous functions on \mathbb{R}^n . A more general model is that $g(\mathbf{x})$ can be an arbitrary function [82]. Clearly, this d.c. programming problem is too abstract. Although it can be used to “model” a very wide range of mathematical problems [47], it is impossible to have an elegant theory and powerful algorithms for solving this problem without detailed structures on these arbitrarily given functions. As a result of extensive studying during the last thirty years (cf. [48, 79]), even some very simple d.c. programming problems are considered as NP-hard [82].

Based on the canonical duality theory, a generalized d.c. programming problem (\mathcal{P}_{dc}) can be presented in a canonical d.c. minimization problem form:

$$(\mathcal{P}_{dc}) : \min \left\{ \Pi(\chi) = \Phi(\Lambda(\chi)) - \frac{1}{2} \langle \chi, \mathbf{H}\chi \rangle - \langle \chi, \bar{\chi}^* \rangle \mid \chi \in \mathcal{X}_c \right\}, \quad (49)$$

where \mathbf{H} is a given positive-definite generalized matrix.

Since the canonical measure $\xi = \Lambda(\chi) \in \mathcal{E}_a$ is nonlinear and $\Phi(\xi)$ is convex on \mathcal{E}_a , the composition $\Phi(\Lambda(\chi))$ has a higher-order nonlinearity than a quadratic function. Therefore, the coercivity for the target function $\Pi(\chi)$:

$$\lim_{\|\chi\| \rightarrow \infty} \Pi(\chi) = \infty \quad (50)$$

should be naturally satisfied for many real-world problems, which is a sufficient condition for existence of a global minimal solution to (\mathcal{P}_{dc}) (otherwise, the set \mathcal{X}_c should be bounded). Clearly, this generalized d.c. minimization problem can be used to model a reasonably large class of real-world problems in multi-disciplinary fields [34, 49].

4.3 Fixed Point Problems

Fixed point problem is a well-established subject in the area of nonlinear analysis, which is usually formulated in the following form:

$$(\mathcal{P}_{fp}) : \quad \mathbf{x} = F(\mathbf{x}), \quad (51)$$

where $F : \mathcal{X}_a \rightarrow \mathcal{X}_a$ is nonlinear mapping and \mathcal{X}_a is a subset of a normed space \mathcal{X} . Problem (\mathcal{P}_{fp}) appears extensively in engineering and sciences, such as equilibrium problems, mathematical economics, game theory, and numerical methods for nonlinear dynamical systems. It is realized [72] that this well-studied field is actually a special subject of global optimization.

Lemma 2 *If F is a potential operator, i.e., there exists a real-valued function $P : \mathcal{X}_a \rightarrow \mathbb{R}$ such that $F(\mathbf{x}) = \nabla P(\mathbf{x})$, then (\mathcal{P}_{fp}) is equivalent to the following stationary point problem:*

$$\bar{\mathbf{x}} = \arg \text{sta} \left\{ \Pi(\mathbf{x}) = P(\mathbf{x}) - \frac{1}{2} \|\mathbf{x}\|^2 \mid \forall \mathbf{x} \in \mathcal{X}_a \right\}. \quad (52)$$

Otherwise, (\mathcal{P}_{fp}) is equivalent to the following global minimization problem:

$$\bar{\mathbf{x}} = \arg \min \left\{ \Pi(\mathbf{x}) = \frac{1}{2} \|F(\mathbf{x}) - \mathbf{x}\|^2 \mid \forall \mathbf{x} \in \mathcal{X}_a \right\}. \quad (53)$$

Proof First, we assume that $F(\mathbf{x})$ is potential operator, then \mathbf{x} is a stationary point of $\Pi(\mathbf{x})$ if and only if $\nabla \Pi(\mathbf{x}) = \nabla P(\mathbf{x}) - \mathbf{x} = 0$, thus \mathbf{x} is also a solution to (\mathcal{P}_{fp}) since $F(\mathbf{x}) = \nabla P(\mathbf{x})$.

Now, we assume that $F(\mathbf{x})$ is not a potential operator. By the fact that $\Pi(\mathbf{x}) = \frac{1}{2} \|F(\mathbf{x}) - \mathbf{x}\|^2 \geq 0 \quad \forall \mathbf{x} \in \mathcal{X}$, the vector $\bar{\mathbf{x}}$ is a global minimizer of $\Pi(\mathbf{x})$ if and only if $F(\bar{\mathbf{x}}) - \bar{\mathbf{x}} = 0$. Thus, $\bar{\mathbf{x}}$ must be a solution to (\mathcal{P}_{fp}) . \square

By the facts that the global minimizer of an unconstrained optimization problem must be a stationary point, and

$$\frac{1}{2}\|F(\mathbf{x})-\mathbf{x}\|^2 = P(\mathbf{x})-\frac{1}{2}\|\mathbf{x}\|^2, \quad P(\mathbf{x}) = \frac{1}{2}\langle F(\mathbf{x}), F(\mathbf{x}) \rangle - \langle \mathbf{x}, F(\mathbf{x}) \rangle + \|\mathbf{x}\|^2, \quad (54)$$

the global minimization problem (53) is a special case of the stationary point problem (52). Mathematically speaking, if a fixed point problem has a trivial solution, then $F(\mathbf{x})$ must be a homogeneous operator, i.e., $F(0) = 0$. For general problems, $F(\mathbf{x})$ should have a nonhomogeneous term $\mathbf{f} \in \mathbb{R}^n$. Thus, we can let $P(\mathbf{x}) = G(\mathbf{D}\mathbf{x}) - \langle \mathbf{x}, \mathbf{f} \rangle$ such that $\mathbf{D} : \mathcal{X} \rightarrow \mathcal{G} \subset \mathbb{R}^m$ is a linear operator and $G : \mathcal{G} \rightarrow \mathbb{R}$ is a generalized objective function. Thus, the fixed point problem (\mathcal{P}_{fp}) can be reformulated in the following stationary point problem:

$$(\mathcal{P}_{fp}) : \quad \bar{\mathbf{x}} = \arg \text{sta} \left\{ \Pi(\mathbf{x}) = G(\mathbf{D}\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2 - \langle \mathbf{x}, \mathbf{f} \rangle \mid \forall \mathbf{x} \in \mathcal{X}_c \right\}. \quad (55)$$

Clearly, the fixed point problem is actually equivalent to a d.c. programming problem. Canonical duality theory for solving this fixed point problem is given recently in [72].

4.4 Mixed Integer Nonlinear Programming (MINLP)

The decision variable for (MINLP) is $\chi = \{\mathbf{y}, \mathbf{z}\} \in \mathcal{Y}_a \times \mathcal{Z}_a$, where \mathcal{Y}_a is a continuous variable set and \mathcal{Z}_a is a set of integers. It was shown in [69] that for any given integer set \mathcal{Z}_a , there exists a linear transformation $\mathbf{D}_z : \mathcal{Z}_a \rightarrow \mathbb{Z} = \{\pm 1\}^n$. Thus, based on the unified problem (\mathcal{P}_g), a general MINLP problem can be proposed as:

$$(\mathcal{P}_{mi}) : \quad \min\{\Pi(\mathbf{y}, \mathbf{z}) = G(\mathbf{D}_y\mathbf{y}, \mathbf{D}_z\mathbf{z}) - \langle \mathbf{y}, \mathbf{s} \rangle - \langle \mathbf{z}, \mathbf{t} \rangle \mid (\mathbf{y}, \mathbf{z}) \in \mathcal{Y}_c \times \mathcal{Z}_c\}, \quad (56)$$

where $\mathbf{f} = (\mathbf{s}, \mathbf{t})$ is a given input, $\mathbf{D}\chi = (\mathbf{D}_y\mathbf{y}, \mathbf{D}_z\mathbf{z}) \in \mathcal{G}_y \times \mathbb{Z}$ is a multi-scale operator, and

$$\mathcal{Y}_c = \{\mathbf{y} \in \mathcal{Y}_a \mid \mathbf{D}_y\mathbf{y} \in \mathcal{G}_y\}, \quad \mathcal{Z}_c = \{\mathbf{z} \in \mathcal{Z}_a \mid \mathbf{D}_z\mathbf{z} \in \mathbb{Z}\}.$$

In \mathcal{Y}_a , certain linear constraints are given. Since the set \mathbb{Z}_a is bounded, by Assumption 1 either $G : \mathcal{G}_y \rightarrow \mathbb{R}$ is coercive or \mathcal{G}_y is bounded. This general problem (\mathcal{P}_{mi}) covers many real-world applications, including the so-called fixed cost problem [41]. Let

$$\mathbf{g} = \Lambda_z(\mathbf{z}) = (\mathbf{D}_z\mathbf{z}) \circ (\mathbf{D}_z\mathbf{z}) \in \mathcal{E}_z = \mathbb{R}_+^n, \quad (57)$$

where $\mathbf{x} \circ \mathbf{y} = \{x_i y_i\}^n$ is the Hadamard product in \mathbb{R}^n , the integer constraint in \mathbb{Z} can be relaxed by the canonical function $\Psi(\mathbf{g}) = \{0 \text{ if } \mathbf{g} \leq \mathbf{e}, \infty \text{ otherwise}\}$, where $\mathbf{e} = \{1\}^n$. Therefore, the canonical form of (\mathcal{P}_{mi}) is

$$\min\{\Pi(\mathbf{y}, \mathbf{z}) = \Phi(\mathbf{\Lambda}(\mathbf{y}, \mathbf{z})) + \Psi(\mathbf{\Lambda}_z(\mathbf{z})) - \langle \mathbf{y}, \mathbf{s} \rangle - \langle \mathbf{z}, \mathbf{t} \rangle \mid \mathbf{y} \in \mathcal{Y}_c\}. \quad (58)$$

Since the canonical function $\Psi(\mathbf{g})$ is convex, semicontinuous, its Fenchel conjugate is

$$\Psi^\sharp(\boldsymbol{\sigma}) = \sup\{\langle \mathbf{g}, \boldsymbol{\sigma} \rangle - \Psi(\mathbf{g}) \mid \mathbf{g} \in \mathbb{R}^n\} = \{\langle \mathbf{e}, \boldsymbol{\sigma} \rangle \text{ if } \boldsymbol{\sigma} \geq 0, \infty \text{ otherwise}\}.$$

The generalized canonical duality relations (31) are $\boldsymbol{\sigma} \geq 0 \Leftrightarrow \mathbf{g} \leq \mathbf{e} \Leftrightarrow \langle \mathbf{g} - \mathbf{e}, \boldsymbol{\sigma} \rangle = 0$. The complementarity shows that the canonical integer constraint $\mathbf{g} = \mathbf{e}$ can be relaxed by the $\boldsymbol{\sigma} > \mathbf{0}$ in continuous space. Thus, if $\boldsymbol{\xi} = \mathbf{\Lambda}(\boldsymbol{\chi})$ is a quadratic homogenous operator and the canonical function $\Phi(\boldsymbol{\xi})$ is convex on \mathcal{E}_a , the canonical dual to (\mathcal{P}_{mi}) is

$$(\mathcal{P}_{mi}^g) : \max \left\{ \Pi^g(\boldsymbol{\zeta}, \boldsymbol{\sigma}) = -\frac{1}{2} \langle [\mathbf{G}(\boldsymbol{\zeta}, \boldsymbol{\sigma})]^+ \mathbf{f}, \mathbf{f} \rangle - \Phi^\sharp(\boldsymbol{\zeta}) - \langle \mathbf{e}, \boldsymbol{\sigma} \rangle \mid (\boldsymbol{\zeta}, \boldsymbol{\sigma}) \in \mathcal{S}_c^+ \right\}, \quad (59)$$

where $\mathbf{G}(\boldsymbol{\zeta}, \boldsymbol{\sigma})$ depends on the quadratic operators $\mathbf{\Lambda}(\boldsymbol{\chi})$ and $\mathbf{\Lambda}_z(\mathbf{z})$, \mathcal{S}_c^+ is a convex open set:

$$\mathcal{S}_c^+ = \{(\boldsymbol{\zeta}, \boldsymbol{\sigma}) \in \mathcal{E}_a^* \times \mathbb{R}_+^n \mid \mathbf{G}(\boldsymbol{\zeta}, \boldsymbol{\sigma}) \geq 0, \boldsymbol{\sigma} > 0\}. \quad (60)$$

The canonical duality-triality theory has been used successfully for solving mixed integer programming problems [35, 41]. Particularly, for the quadratic integer programming problem:

$$(\mathcal{P}_{qi}) : \min \left\{ \Pi(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} - \mathbf{x}^T \mathbf{f} \mid \mathbf{x} \in \{-1, 1\}^n \right\}, \quad (61)$$

we have $\mathcal{S}_c^+ = \{\boldsymbol{\sigma} \in \mathbb{R}_+^n \mid \mathbf{G}(\boldsymbol{\sigma}) = \mathbf{Q} + 2\text{Diag}(\boldsymbol{\sigma}) \geq 0, \boldsymbol{\sigma} > 0\}$ and

$$(\mathcal{P}_{qi}^g) : \max \left\{ \Pi^g(\boldsymbol{\sigma}) = -\frac{1}{2} \mathbf{f}^T [\mathbf{G}(\boldsymbol{\sigma})]^+ \mathbf{f} - \mathbf{e}^T \boldsymbol{\sigma} \mid \boldsymbol{\sigma} \in \mathcal{S}_c^+ \right\} \quad (62)$$

which can be solved easily if $\text{int}\mathcal{S}_c^+ \neq \emptyset$. Otherwise, (\mathcal{P}_{qi}) could be NP-hard since \mathcal{S}_c^+ is an open set, which is a conjecture proposed in [24]. In this case, (\mathcal{P}_{qi}) is canonically dual to an unconstrained nonsmooth/nonconvex minimization problem [25].

4.5 General Knapsack Problem and Analytical Solution

Knapsack problems appear in real-world decision-making processes in a wide variety of fields, such as finding the least wasteful way to cut raw materials, resource allocation where there are financial constraints, selection of investments and portfolios, selection of assets for asset-backed securitization, and generating keys for the Merkle-Hellman and other knapsack cryptosystems. Mathematically, a general quadratic knapsack problem can be formulated as an integer programming problem:

$$(\mathcal{P}_{qk}) : \min \left\{ \Pi_{qk}(\mathbf{z}) = \frac{1}{2} \mathbf{z}^T \mathbf{Q} \mathbf{z} - \mathbf{c}^T \mathbf{z} \mid \mathbf{z} \in \{0, 1\}^n, \mathbf{v}^T \mathbf{z} \leq V_c \right\}, \quad (63)$$

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is a given symmetrical, usually indefinite, matrix, $\mathbf{c}, \mathbf{v} \in \mathbb{R}^n$ are two given vectors, and $V_c > 0$ is a design parameter.

The knapsack problem has been studied for more than a century, with early works dating as far back as 1897. The main difficulty in this problem is the integer constraint $\mathbf{z} \in \{0, 1\}^n$, so that even the most simple linear knapsack problem:

$$(\mathcal{P}_{lk}) : \max \left\{ \Pi_{lk}(\mathbf{z}) = -\mathbf{c}^T \mathbf{z} \mid \mathbf{z} \in \{0, 1\}^n, \mathbf{v}^T \mathbf{z} \leq V_c \right\}, \quad (64)$$

is listed as one of Karp's 21 NP-complete problems [50].

By the fact that $\alpha \circ \mathbf{z}^2 = \alpha \circ \mathbf{z} \quad \forall \mathbf{z} \in \{0, 1\}^n, \forall \alpha \in \mathbb{R}^n$, for any given symmetrical $\mathbf{Q} \in \mathbb{R}^{n \times n}$ we can choose an α such that $\mathbf{Q}_\alpha = \mathbf{Q} + 2\text{Diag}(\alpha) \geq 0$. Thus, by $\mathbf{c}_\alpha = \mathbf{c} + \alpha$, the problem (\mathcal{P}_q) can be equivalently written in the so-called α -perturbation form [25]:

$$(\mathcal{P}_\alpha) : \min \left\{ \Pi_\alpha(\mathbf{z}) = \frac{1}{2} \mathbf{z}^T \mathbf{Q}_\alpha \mathbf{z} - \mathbf{c}_\alpha^T \mathbf{z} \mid \mathbf{v}^T \mathbf{z} \leq V_c, \mathbf{z} \in \{0, 1\}^n \right\}. \quad (65)$$

Let $\text{rank } \mathbf{Q}_\alpha = r \leq n$, there must exist (see [77]) an $\mathbf{L} \in \mathbb{R}^{r \times n}$ and $\mathbf{H} \in \mathbb{R}^{r \times r}$ with $\text{rank } \mathbf{L} = \text{rank } \mathbf{H} = r$ and $\mathbf{H} \succ 0$ such that $\mathbf{Q}_\alpha = 4\mathbf{L}^T \mathbf{H} \mathbf{L}$. Similar to the α -perturbed canonical dual problem (\mathcal{P}_{ip}^g) given in [25], the canonical dual problem (\mathcal{P}_q^g) can be reformulated as:

$$(\mathcal{P}_\alpha^g) : \max_{\boldsymbol{\zeta} \in \mathcal{S}_c^+} \left\{ \Pi_\alpha^g(\boldsymbol{\sigma}, \tau) = -\frac{1}{2} \text{Abs}[\boldsymbol{\phi}(\boldsymbol{\sigma}, \tau)] - \frac{1}{2} \boldsymbol{\sigma}^T \mathbf{H}^{-1} \boldsymbol{\sigma} - \tau V_b + d \right\}, \quad (66)$$

where $V_b = V_c - \frac{1}{2} \sum_{i=1}^n v_i$, $d = \frac{1}{8} \sum_{i=1}^n (2\alpha_i + \sum_{j=1}^n Q_{ij}) - \frac{1}{2} \sum_{i=1}^n (c_i + \alpha_i)$,

$$\mathcal{S}_c^+ = \{ \boldsymbol{\zeta} = (\boldsymbol{\sigma}, \tau) \in \mathbb{R}^{m+1} \mid \tau \geq 0 \}. \quad (67)$$

$$\boldsymbol{\phi}(\boldsymbol{\sigma}, \tau) = \mathbf{c} - \tau \mathbf{v} - 2\mathbf{L}^T \boldsymbol{\sigma} - \frac{1}{2} \mathbf{Q} \mathbf{e}, \quad (68)$$

The notation $\text{Abs}[\boldsymbol{\phi}(\boldsymbol{\sigma}, \tau)]$ denotes $\text{Abs}[\boldsymbol{\phi}(\boldsymbol{\sigma}, \tau)] = \sum_{i=1}^n |\phi_i(\boldsymbol{\sigma}, \tau)|$.

Theorem 6 (Analytical Solution to Quadratic Knapsack Problem) *For any given $V_c > 0$, $\mathbf{v}, \mathbf{c} \in \mathbb{R}_+^n$, $\boldsymbol{\alpha} \in \mathbb{R}_+^n$ such that $\mathbf{Q}_\alpha = \mathbf{Q} + 2\text{Diag}(\boldsymbol{\alpha}) = 4\mathbf{L}^T \mathbf{H} \mathbf{L}$ and $\mathbf{H} \succ 0$, if $\bar{\boldsymbol{\xi}} = \{\bar{\boldsymbol{\sigma}}, \bar{\tau}\}$ is a solution to (\mathcal{P}_α^g) , then*

$$\bar{\mathbf{z}} = \frac{1}{2} \left\{ \frac{\phi_i(\bar{\boldsymbol{\sigma}}, \bar{\tau})}{|\phi_i(\bar{\boldsymbol{\sigma}}, \bar{\tau})|} + 1 \right\}^n \quad (69)$$

is a global optimal solution to (\mathcal{P}_α) and

$$\Pi_\alpha(\bar{\mathbf{z}}) = \min_{\mathbf{z} \in \mathcal{Z}_\alpha} \Pi_\alpha(\mathbf{z}) = \max_{\boldsymbol{\xi} \in \mathcal{S}_\alpha^+} \Pi_\alpha^g(\boldsymbol{\xi}) = \Pi_\alpha^g(\bar{\boldsymbol{\xi}}). \quad (70)$$

Theorem 7 (Existence and Uniqueness Theorem to Quadratic Knapsack Problem) *For any given $V_c > 0$, $\mathbf{v}, \mathbf{c} \in \mathbb{R}_+^n$, $\boldsymbol{\alpha} \in \mathbb{R}_+^n$ such that $\mathbf{Q}_\alpha = \mathbf{Q} + 2\text{Diag}(\boldsymbol{\alpha}) = 4\mathbf{L}^T \mathbf{H} \mathbf{L}$, $\mathbf{H} \succ 0$, and $\bar{\boldsymbol{\xi}} = \{\bar{\boldsymbol{\sigma}}, \bar{\tau}\}$ is a solution to (\mathcal{P}_α^g) , if*

$$\phi_i(\bar{\boldsymbol{\sigma}}, \bar{\tau}) \neq 0 \quad \forall i = 1, \dots, n \quad (71)$$

then the canonical dual feasible set $\mathcal{S}_\alpha^+ \neq \emptyset$ and the knapsack problem (\mathcal{P}_α) has a unique solution. Otherwise, if $\phi_i(\bar{\boldsymbol{\sigma}}, \bar{\tau}) = 0$ for at least one $i \in \{1, \dots, n\}$, then $\mathcal{S}_\alpha^+ = \emptyset$ and (\mathcal{P}_α) has at least two solutions.

The canonical dual for the linear knapsack problem has a very simple form:

$$(\mathcal{P}_{lk}^g) : \max_{\tau \geq 0} \left\{ \Pi_{lk}^g(\tau) = -\frac{1}{2} \sum_{i=1}^n (|c_i - \tau v_i| - \tau v_i) - \tau V_c \right\}. \quad (72)$$

Corollary 1 (Analytical Solution to Linear Knapsack Problem) *For any given $V_c > 0$, $\mathbf{v}, \mathbf{c} \in \mathbb{R}_+^n$, if $\bar{\tau} > 0$ is a solution to (\mathcal{P}_{lk}^g) , then*

$$\bar{\mathbf{z}} = \frac{1}{2} \left\{ \frac{c_i - \bar{\tau} v_i}{|c_i - \bar{\tau} v_i|} + 1 \right\}^n \quad (73)$$

is a global optimal solution to (\mathcal{P}_l) and

$$\Pi_{lk}(\bar{\mathbf{z}}) = \Pi_{lk}^g(\bar{\tau}) \quad (74)$$

Corollary 2 (Existence and Uniqueness Theorem to Linear Knapsack Problem) *For any given $\mathbf{v}, \mathbf{c} \in \mathbb{R}_+^n$, if there exists a constant $\tau_c > 0$ such that*

$$\psi_i(\tau_c) = \tau_c v_i - c_i \neq 0 \quad \forall i = 1, \dots, n \quad (75)$$

then the knapsack problem (\mathcal{P}_{lk}) has a unique solution. Otherwise, if $\psi_i(\tau_c) = 0$ for at least one $i \in \{1, \dots, n\}$, then (\mathcal{P}_{lk}) has at least two solutions.

Detailed proof of these results is given by Gao in [31].

The so-called multi-dimensional knapsack problem (MKP) is a generalization of the linear knapsack problem, that is:

$$(\mathcal{P}_{mk}) : \max \mathbf{c}^T \mathbf{z}, \quad s.t. \quad \mathbf{Wz} \leq \boldsymbol{\omega}, \quad \mathbf{z} \in \{0, 1\}^n, \quad (76)$$

where $\mathbf{c} \in \mathbb{R}_+^n$ and $\boldsymbol{\omega} \in \mathbb{R}_+^m$ ($m < n$) are two given nonnegative vectors,

$$\mathbf{W} \in \mathbb{R}_+^{m \times n} = \{\mathbf{W} = \{w_{ij}\} \in \mathbb{R}^{m \times n} \mid w_{ij} \geq 0 \quad \forall i = 1, \dots, m, \quad j = 1, \dots, n\}$$

is a given nonnegative matrix such that $w_{ij} \leq \omega_j$, $\sum_{j=1}^n w_{ij} \geq \omega_i$. Clearly, this problem has multi-knapsacks $\{\omega_i\}^m$. Therefore, instead of the multi-dimensional, the correct name for (\mathcal{P}_{mk}) should be the *multi-knapsacks problem*. This problem has applications in many fields including capital budgeting problems and resource allocation [66]. The canonical dual problem for (\mathcal{P}_{mk}) is

$$(\mathcal{P}_{mk}^g) : \max_{\boldsymbol{\tau} \in \mathbb{R}_+^m} \left\{ \Pi_{mk}^g(\boldsymbol{\tau}) = -\frac{1}{2} \sum_{i=1}^n (|c_i - \sum_{j=1}^m w_{ji} \tau_j| - \sum_{j=1}^m w_{ji} \tau_j) - \boldsymbol{\omega}^T \boldsymbol{\tau} \right\}. \quad (77)$$

Thus, if $\bar{\boldsymbol{\tau}} = \{\bar{\tau}_i\}$ is a global maximizer of (\mathcal{P}_{mk}^g) , the analytic solution to (\mathcal{P}_{mk}) is

$$\mathbf{z} = \frac{1}{2} \left(\frac{c_i - \sum_{j=1}^m w_{ji} \bar{\tau}_j}{|c_i - \sum_{j=1}^m w_{ji} \bar{\tau}_j|} + 1 \right). \quad (78)$$

4.6 Bilevel Optimization and Optimal Control

Bilevel optimization appears extensively in optimal design and control of complex systems. A general formulation of the bilevel optimization problem can be written as follows:

$$(\mathcal{P}_{bo}) : \min \{T(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in \mathcal{X}_a, \quad \mathbf{y} \in \mathcal{Y}_a\}, \quad (79)$$

$$s.t. \quad \mathbf{y} \in \arg \min \{\Pi(\mathbf{v}, \mathbf{x}) \mid \mathbf{v} \in \mathcal{Y}_a\}, \quad (80)$$

where T represents the top-level target (or leader) function, Π is the lower-level target (or follower) function. Similarly, $\mathbf{x} \in \mathcal{X}_a$ represents upper-level decision

vector and $\mathbf{y} \in \mathcal{Y}_a$ represents the lower-level variable. Clearly, this is a coupled nonlinear optimization problem, which is fundamentally difficult even for convex systems.

To solve this coupling problem numerically, an alternative iteration method can be used [30]:

- (1) For a given \mathbf{x}_{k-1} , solve the lower-level problem first to obtain

$$\mathbf{y}_k \in \arg \min \{ \Pi(\mathbf{y}, \mathbf{x}_{k-1}) \mid \mathbf{y} \in \mathcal{Y}_a \}. \quad (81)$$

- (2) Then for the fixed \mathbf{y}_k , solve the upper-level problem for

$$\mathbf{x}_k = \arg \min \{ T(\mathbf{x}, \mathbf{y}_k) \mid \mathbf{x} \in \mathcal{X}_a \}. \quad (82)$$

These two single-level optimization problems can be solved by the canonical duality theory, and the sequence $\{\mathbf{x}_k, \mathbf{y}_k\}$ can converge to an optimal solution of (\mathcal{P}_{bo}) under certain conditions.

As an example, let us consider the following optimal control problem:

$$(\mathcal{P}_{oc}) : \quad \min \{ \Phi(\mathbf{v}, \boldsymbol{\chi}) \mid \mathbf{v} \in \mathcal{U}, \boldsymbol{\chi} \in \mathcal{X} \}, \quad (83)$$

$$\text{s.t. } \dot{\boldsymbol{\chi}} = \mathbf{a}(\boldsymbol{\chi}, \mathbf{v}, t) \quad \forall t \in I = [t_o, t_b], \quad (84)$$

where $\boldsymbol{\chi}(t)$ is the state, $\mathbf{v}(t)$ is the control; \mathcal{X} is a feasible set including the boundary conditions: $\boldsymbol{\chi}(t_o) = \boldsymbol{\chi}_o$ and $\psi(\boldsymbol{\chi}(t_b), t_b) = 0$. The upper-level target Φ is usually a quadratic continuous-time cost functional:

$$\Phi(\mathbf{v}, \boldsymbol{\chi}) = \frac{1}{2} \int_I \left[\boldsymbol{\chi}^T(t) \mathbf{Q}(t) \boldsymbol{\chi}(t) + \mathbf{v}^T(t) \mathbf{R}(t) \mathbf{v}(t) - 2 \boldsymbol{\chi}^T(t) \mathbf{P}(t) \mathbf{v}(t) \right] dt + \Phi_b(\boldsymbol{\chi}(t_b)) \quad (85)$$

where $\mathbf{Q}(t) \in \mathbb{R}^{d \times d}$, $\mathbf{R}(t) \in \mathbb{R}^{p \times p}$ are positive semi-definite and positive definite, respectively, on the time domain $I = [t_o, t_b]$, $\Phi_b(\boldsymbol{\chi}(t_b)) = \frac{1}{2} \boldsymbol{\chi}^T(t_b) \mathbf{Q}_b \boldsymbol{\chi}(t_b)$. New to this cost function is the coupling term $\boldsymbol{\chi}^T(t) \mathbf{P}(t) \mathbf{v}(t)$, where $\mathbf{P}(t) \in \mathbb{R}^{d \times p}$ is a given matrix function of t , which plays an important role in alternative iteration methods for solving the general nonlinear optimal control problem (\mathcal{P}_{oc}) .

For conservative systems, the nonlinear operator $\mathbf{a}(\boldsymbol{\chi}, \mathbf{v}, t)$ is a potential operator, i.e., there exists an action (or Lagrangian) $\Pi(\boldsymbol{\chi}, \dot{\boldsymbol{\chi}}, \mathbf{v})$ such that for any given control $\mathbf{v}(t) \in \mathcal{U}$ the differential equation (84) can be written in the following least action form:

$$\boldsymbol{\chi} \in \arg \min \{ \Pi(\boldsymbol{\chi}, \dot{\boldsymbol{\chi}}, \mathbf{v}) \mid \boldsymbol{\chi} \in \mathcal{X} \} \quad (86)$$

Although such a Lagrangian does not exist for dissipative systems, the least squares method can always be used so that (84) can also be written in this minimization form.

In order to reformulate the challenging control problem (\mathcal{P}_{oc}) in function space, the finite element method can be used such that the time domain I is divided into n elements $\{I_e = [t_k, t_{k+1}]\}$ and in each I_e , the unknown fields can be numerically discretized as:

$$\mathbf{v}(t) = \mathbf{N}_u^e(t)\mathbf{u}_e, \quad \boldsymbol{\chi}(t) = \mathbf{N}_x^e(t)\mathbf{x}_e \quad \forall t \in I_e, \quad e = 1, \dots, n, \quad (87)$$

where $\mathbf{N}_u^e(t)$ is an interpolation matrix for $\mathbf{v}(t)$, $\mathbf{u}_e = (\mathbf{v}(t_k), \mathbf{v}(t_{k+1}))$ is a nodal control vector; similarly, $\mathbf{N}_x^e(t)$ is an interpolation matrix for $\boldsymbol{\chi}(t)$ and $\mathbf{x}_e = (\boldsymbol{\chi}(t_k), \boldsymbol{\chi}(t_{k+1}))$ is a nodal state vector. Let $\mathcal{U}_a \subset \mathbb{R}^{p \times n}$ be an admissible nodal control space, $\mathcal{X}_a \subset \mathbb{R}^{d \times n}$ be an admissible state space, $\mathbf{u} = \{\mathbf{v}_k\} \in \mathcal{U}_a$, and $\mathbf{x} = \{\boldsymbol{\chi}_k\} \in \mathcal{X}_a$, then both the cost functional Φ and the action Π can be numerically written as:

$$\Phi(\mathbf{v}, \boldsymbol{\chi}) \approx \Phi_h(\mathbf{u}, \mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{Q}_h \mathbf{x} + \frac{1}{2}\mathbf{u}^T \mathbf{R}_h \mathbf{u} - \mathbf{x}^T \mathbf{P}_h \mathbf{u}, \quad (88)$$

$$\Pi(\boldsymbol{\chi}, \dot{\boldsymbol{\chi}}, \mathbf{v}) \approx \Pi_h(\mathbf{x}, \mathbf{u}) = G(\mathbf{D}\mathbf{x}, \mathbf{u}) - F(\mathbf{x}, \mathbf{u}), \quad (89)$$

where $G(\mathbf{D}\mathbf{x}, \mathbf{u})$ and $F(\mathbf{x}, \mathbf{u})$ depend on the action $\Pi(\boldsymbol{\chi}, \dot{\boldsymbol{\chi}}, \mathbf{u})$,

$$\mathbf{Q}_h = \sum_{e=1}^n \int_{I_e} \mathbf{N}_x^T(t) \mathbf{Q}(t) \mathbf{N}_x(t) dt + \frac{1}{2} \mathbf{N}_x^T(t_b) \mathbf{Q}_b \mathbf{N}_x(t_b),$$

$$\mathbf{R}_h = \sum_{e=1}^n \int_{I_e} \mathbf{N}_u^T(t) \mathbf{R}(t) \mathbf{N}_u(t) dt,$$

$$\mathbf{P}_h = \sum_{e=1}^n \int_{I_e} \mathbf{N}_x^T(t) \mathbf{P}(t) \mathbf{N}_u(t) dt.$$

Therefore, the optimal control problem (\mathcal{P}_{oc}) can be written in a bilevel optimization problem:

$$(\mathcal{P}_{oc}^h) : \quad \min \{ \Phi_h(\mathbf{u}, \mathbf{x}) \mid \mathbf{u} \in \mathcal{U}_a, \mathbf{x} \in \mathcal{X}_a \}, \quad (90)$$

$$\text{s.t. } \mathbf{x} \in \arg \min \{ \Pi_h(\mathbf{y}, \mathbf{u}) \mid \mathbf{y} \in \mathcal{X}_a \} \quad (91)$$

The canonical duality theory has been successfully applied for solving nonlinear dynamical systems [54, 71] and the relation between chaos and NP-hardness was first discovered by Latorre and Gao [54]. Combined with an alternative iteration method, the canonical duality theory can be used to efficiently solve the general bilevel optimization problems.

4.7 Multi-Level Multi-Targets MINLP and Topology Optimization

Multi-target optimization is concerned with mathematical optimization problems involving more than one target function to be optimized simultaneously. Since the target is a vector-valued function, it is also known as vector optimization, multi-criteria optimization, or Pareto optimization. By the fact that the objectivity has been misused in optimization literature, this important research area has been misguidedly called multi-objective optimization. Multi-target optimization problems appear extensively in multi-scale complex systems where optimal decisions need to be taken in the presence of trade-offs between two or more conflicting targets. Therefore, the multi-level optimization and MINLP problems are naturally involved with the multi-target optimization. In real-world applications, the multi-level multi-target mixed integer nonlinear programming (MMM) could have many different formulations. Based on the canonical duality theory, a simple form of MMM problems can be proposed as the following:

$$(\mathcal{P}_{3m}) : \quad \min \{T(\mathbf{z}, \bar{\mathbf{x}}, \bar{\mathbf{y}}) \mid \bar{\mathbf{x}} \in \mathcal{X}_a, \bar{\mathbf{y}} \in \mathcal{Y}_a, \mathbf{z} \in \mathcal{Z}_a\}, \quad (92)$$

$$\text{s.t. } \bar{\mathbf{x}} \in \arg \min \{\Pi_x(\mathbf{x}, \mathbf{y}, \mathbf{z}) \mid \mathbf{x} \in \mathcal{X}_a\}, \quad (93)$$

$$\bar{\mathbf{y}} \in \arg \min \{\Pi_y(\mathbf{x}, \mathbf{y}, \mathbf{z}) \mid \mathbf{y} \in \mathcal{Y}_a\}. \quad (94)$$

Without loss of generality, we assume that the leader variable $\mathbf{z} \in \mathcal{Z}_a$ is a discrete vector, the follower variables $\mathbf{x} \in \mathcal{X}_a$ and $\mathbf{y} \in \mathcal{Y}_a$ are continuous vectors; the top-level (leader) target $T : \mathcal{Z}_a \times \mathcal{X}_a \times \mathcal{Y}_a \rightarrow \mathbb{R}^m$ is a vector-valued function, which is not necessary to be objective, while the lower-level (follower) targets Π_x and Π_y are real-valued functions such that the follower problems can be written respectively in the canonical form (\mathcal{P}) , where the objectivity and subjectivity are required. If we let

$$\mathbf{u} = \{\mathbf{x}, \mathbf{y}, \dots\} \in \mathcal{U}_a = \mathcal{X}_a \times \mathcal{Y}_a \times \dots,$$

$$\Pi(\mathbf{u}, \mathbf{z}) = \{\Pi_x(\mathbf{u}, \mathbf{z}), \Pi_y(\mathbf{u}, \mathbf{z}), \dots\} : \mathcal{U}_a \times \mathcal{Z}_a \rightarrow \mathbb{R}^d, \quad d \geq 2,$$

then the MMM problem can be written in a general form:

$$(\mathcal{P}_{3m}) : \quad \min \{T(\mathbf{z}, \mathbf{u}) \mid \mathbf{u} \in \mathcal{U}_a, \mathbf{z} \in \mathcal{Z}_a\}, \quad (95)$$

$$\text{s.t. } \mathbf{u} \in \arg \min \{\Pi(\mathbf{v}, \mathbf{z}) \mid \mathbf{v} \in \mathcal{U}_a\}. \quad (96)$$

Clearly, the (\mathcal{P}_{3m}) should be one of the most challenging problems proposed so far in global optimization even if both T and Π are linear vector-valued functions.

Topology optimization is a mathematical tool that optimizes the best mass density distribution $\rho(\mathbf{x})$ within a design domain $\Omega \subset \mathbb{R}^d$ in order to obtain the best structural performance governed by the minimum total potential principle:

$$\min \left\{ \Pi(\mathbf{u}, \rho) = \int_{\Omega} U(\nabla \mathbf{u}) \rho \, d\Omega - \int_{\Gamma_t} \mathbf{u}^T \mathbf{t} \, d\Gamma \mid \mathbf{u} \in \mathcal{U} \right\} \quad (97)$$

where $\mathbf{u} : \Omega \rightarrow \mathbb{R}^d$ is a displacement vector field, the design variable $\rho(\mathbf{x}) \in \{0, 1\}$ is a discrete scalar field, which takes $\rho(\mathbf{x}) = 1$ at a solid material point $\mathbf{x} \in \Omega$ and $\rho(\mathbf{x}) = 0$ at a void point $\mathbf{x} \in \Omega$. By using finite element method, the design domain Ω is meshed with n disjointed finite elements $\{\Omega_e\}$ and let

$$\mathbf{u}(\mathbf{x}) = \mathbf{N}(\mathbf{x})\mathbf{u}_e \quad \rho(\mathbf{x}) = z_e \in \{0, 1\} \quad \forall \mathbf{x} \in \Omega_e,$$

the total potential energy can be numerically written as:

$$\Pi(\mathbf{u}, \rho) \approx \Pi_h(\mathbf{u}, \mathbf{z}) = \mathbf{z}^T \mathbf{c}(\mathbf{u}) - \mathbf{u}^T \mathbf{f}$$

where $\mathbf{u} = \{\mathbf{u}_e\} \in \mathcal{U}_a \subset \mathbb{R}^m$ is a nodal displacement vector, $\mathbf{z} \in \mathcal{Z}_a \subset \{0, 1\}^n$ is a discretized design vector, and

$$\mathbf{c}(\mathbf{u}) = \left\{ \int_{\Omega_e} U(\nabla \mathbf{N}(\mathbf{x})\mathbf{u}_e) \, d\Omega \right\} \in \mathbb{R}_+^n, \quad \mathbf{f} = \left\{ \int_{\Gamma_{te}} \mathbf{N}(\mathbf{x})^T \mathbf{t}(\mathbf{x}) \, d\Gamma \right\} \in \mathbb{R}^m.$$

Let

$$\mathcal{Z}_a = \{\mathbf{z} \in \{0, 1\}^n \mid \mathbf{v}^T \mathbf{z} \leq \omega\},$$

where $\mathbf{v} = \{v_e\} \in \mathbb{R}_+^n$ and $v_e \geq 0$ is the volume of the e -th element, and $\omega > 0$ is the desired volume of structure. The correct mathematical problem for general topology optimization has been proposed recently by Gao [30, 31]:

Problem 1 (Topology Optimization for General Materials) *For a given external load \mathbf{f} and the desired volume $\omega > 0$, to solve the bilevel MINLP problem:*

$$(\mathcal{P}_{to}) : \quad \min \{T(\mathbf{z}, \mathbf{u}) \mid \mathbf{u} \in \mathcal{U}_a, \mathbf{z} \in \mathcal{Z}_a\}, \quad (98)$$

$$s.t. \quad \mathbf{u} \in \arg \min \{\Pi_h(\mathbf{v}, \mathbf{z}) \mid \mathbf{v} \in \mathcal{U}_a\}. \quad (99)$$

The top-level target function T depends on each design problem, which could be

$$T_1(\mathbf{z}, \mathbf{u}) = \mathbf{u}^T \mathbf{f} - \mathbf{z}^T \mathbf{c}(\mathbf{u}), \quad (100)$$

$$T_2(\mathbf{z}, \mathbf{u}) = \frac{1}{2} \mathbf{z}^T \mathbf{Q}(\mathbf{u}) \mathbf{z} - \mathbf{z}^T \mathbf{c}(\mathbf{u}). \quad (101)$$

where $\mathbf{Q}(\mathbf{u})$ is a symmetrical matrix whose diagonal elements $\{Q_{ii}\}_{i=1}^n = 0$ and $Q_{ij}(\mathbf{u})$ is the negative effect to the structure if the i -th and j -th elements are elected. Clearly, the top-level is a linear knapsack problem for $T = T_1$, or a quadratic

knapsack problem if $T = T_2$. If $\mathcal{Z}_a = \{\mathbf{z} \in \{0, 1\}^n \mid \mathbf{W}\mathbf{z} \leq \boldsymbol{\omega}\}$, then the top level is a multi-knapsack problem.

In topology limit design, the top-level target could be a vector-valued function depending on certain design parameter α , say:

$$T_3(\alpha, \mathbf{z}, \mathbf{u}) = \{\alpha, T_i(\mathbf{z}, \mathbf{u})\}, \quad i = 1, 2. \quad (102)$$

If α is the volume ω , then (\mathcal{P}_{to}) is a topology optimization for lightweight design problem:

Problem 2 (Topology Lightweight Design) *For the given external load and $\omega^b > \omega^a > 0$, to solve*

$$(\mathcal{P}_{lw}) : \min \{T_3(\omega, \mathbf{z}, \mathbf{u}) \mid \mathbf{v}^T \mathbf{z} \leq \omega, \omega \in [\omega^a, \omega^b], \mathbf{u} \in \mathcal{U}_a, \mathbf{z} \in \{0, 1\}^n\}, \quad (103)$$

$$s.t. \mathbf{u} \in \arg \min \{\Pi_h(\mathbf{v}, \mathbf{z}) \mid \mathbf{v} \in \mathcal{U}_a\}. \quad (104)$$

This is a bilevel multi-target knapsack problem.

If $\alpha = -\eta$ and $\eta > 0$ is the external loading factor, then by simply choosing $T_3(\alpha, \mathbf{z}, \mathbf{u}) = -\mathbf{z}^T \mathbf{c}(\mathbf{u})$ we have the following problem.

Problem 3 (Topology Limit Design) *For the given external load distribution \mathbf{f} and the plastic yield condition in \mathcal{U}_a , to solve*

$$(\mathcal{P}_{ld}) : \max \{\mathbf{z}^T \mathbf{c}(\mathbf{u}) \mid \eta > 0, \mathbf{u} \in \mathcal{U}_a, \mathbf{z} \in \mathcal{Z}_a\}, \quad (105)$$

$$s.t. \mathbf{u} \in \arg \min \{\mathbf{z}^T \mathbf{c}(\mathbf{v}) - \eta \mid \mathbf{v}^T \mathbf{f} = 1, \mathbf{v} \in \mathcal{U}_a\}. \quad (106)$$

If $\alpha = \{\omega, -\eta\}$, the combination of (\mathcal{P}_{lw}) and (\mathcal{P}_{ld}) forms a new problem:

Problem 4 (Topology Lightweight Limit Design) *For the given $\omega^b > \omega^a > 0$, the external load distribution \mathbf{f} and the plastic yield condition in \mathcal{U}_a , to solve*

$$(\mathcal{P}_{ll}) : \min \{\omega, -\mathbf{z}^T \mathbf{c}(\mathbf{u})\} \quad \forall \omega \in [\omega^a, \omega^b], \eta > 0, \mathbf{u} \in \mathcal{U}_a, \mathbf{z} \in \mathcal{Z}_a, \quad (107)$$

$$s.t. \mathbf{u} \in \arg \min \{\mathbf{z}^T \mathbf{c}(\mathbf{v}) - \eta \mid \mathbf{v}^T \mathbf{f} = 1, \mathbf{v} \in \mathcal{U}_a\}. \quad (108)$$

Due to a conflict between $\min \omega$ and $\max \{\eta = \mathbf{z}^T \mathbf{c}(\mathbf{u})\}$, this MMM problem could exist a (possibly infinite) number of Pareto optimal solutions.

The canonical duality theory is particularly useful in topology optimization for full-stress (or plastic limit) design. In this type of problems, it is much more convenient to use the stress as the unknown in analysis. Therefore, dual to (\mathcal{P}_{to}) the problem for full-stress design can be proposed as:

$$(\mathcal{P}_{to}^*) : \max \{T^*(\mathbf{z}, \boldsymbol{\sigma}) \mid \mathbf{z} \in \mathcal{Z}_a, \boldsymbol{\sigma} \in \mathcal{S}_a^+\}, \quad (109)$$

$$s.t. \boldsymbol{\sigma} \in \arg \max \{\Pi_h^d(\boldsymbol{\tau}, \mathbf{z}) \mid \boldsymbol{\tau} \in \mathcal{S}_a^+\}, \quad (110)$$

The top-level dual target $T^d(\mathbf{z}, \boldsymbol{\sigma})$ can be

$$T_1^d(\mathbf{z}, \boldsymbol{\sigma}) = \mathbf{z}^T \mathbf{c}^d(\boldsymbol{\sigma}), \quad (111)$$

$$T_2^d(\mathbf{z}, \boldsymbol{\sigma}) = \mathbf{z}^T \mathbf{c}^d(\boldsymbol{\sigma}) + \frac{1}{2} \mathbf{z}^T \mathbf{Q}^*(\boldsymbol{\sigma}) \mathbf{z}, \quad (112)$$

where $\mathbf{c}^d(\boldsymbol{\sigma}) \in \mathbb{R}_+^n$ is a positive vector such that each of its components $c_e^d(\boldsymbol{\sigma})$ is the pure complementary energy in the e -th element Ω_e . The feasible space \mathcal{S}_a^+ is a bounded convex set with an inequality constraint $\|\boldsymbol{\sigma}\|_g \leq \sigma_c$, where $\|\boldsymbol{\sigma}\|_g$ is a generalized norm which depends on the yield condition adopted, say either Trisca or von Mises criterion [11]. Corresponding to T_3 , we have

$$T_3^d(\alpha^*, \mathbf{z}, \boldsymbol{\sigma}) = \{\alpha^*, T_i^d(\mathbf{z}, \boldsymbol{\sigma})\}, \quad i = 1, 2, \quad (113)$$

where α^* could be $-\omega$, η , or other design parameters.

For linear elastic structures, the total potential energy is a quadratic function of \mathbf{u} :

$$\Pi_h(\mathbf{u}, \mathbf{z}) = \frac{1}{2} \mathbf{u}^T \mathbf{K}(\mathbf{z}) \mathbf{u} - \mathbf{u}^T \mathbf{f}, \quad (114)$$

where $\mathbf{K}(\mathbf{z}) = \{z_e \mathbf{K}_e\} \in \mathbb{R}^{n \times n}$ is the overall stiffness matrix, obtained by assembling the sub-matrix $z_e \mathbf{K}_e$ for each element Ω_e . Accordingly, $\mathbf{c}(\mathbf{u}) = \frac{1}{2} \{\mathbf{u}_e^T \mathbf{K}_e \mathbf{u}_e\}$ is the strain energy vector. In this case, the global optimal solution for the lower-level minimization problem (99) is simply governed by a linear equilibrium equation $\mathbf{K}(\mathbf{z}) \mathbf{u} = \mathbf{f}$. Then for $T = T_1$, the bilevel knapsack problem (\mathcal{P}_{to}) can be written in the single-level reduction:

$$(\mathcal{P}_{le}) : \min \left\{ \mathbf{f}^T \mathbf{u} - \mathbf{z}^T \mathbf{c}(\mathbf{u}) \mid \mathbf{K}(\mathbf{z}) \mathbf{u} = \mathbf{f}, \mathbf{v}^T \mathbf{z} \leq V_c, \mathbf{z} \in \{0, 1\}^n \right\}. \quad (115)$$

This knapsack-type problem makes a perfect sense in topology optimization, i.e., among all elements $\{\Omega_e\}$ with the given volume vector $\mathbf{v} = \{v_e\}$, one should keep only those who stored more strain energy density $\mathbf{c}(\mathbf{u})$. Based on the canonical dual solution to the knapsack problem, a canonical duality algorithm (CDT) is developed with successful applications.

In term of the stress, the full-stress design problem (\mathcal{P}_{fs}^*) for linear elastic structures can be simply given as:

$$(\mathcal{P}_{fs}^*) : \max \{ \mathbf{z}^T \mathbf{c}^d(\boldsymbol{\sigma}) \mid \mathbf{z} \in \mathcal{Z}_a, \boldsymbol{\sigma} \in \mathcal{S}_a^+ \}, \quad (116)$$

where $\mathbf{c}^d(\boldsymbol{\sigma}) = \{\frac{1}{2} \boldsymbol{\sigma}_e^T \mathbf{C}_e \boldsymbol{\sigma}_e\} \in \mathbb{R}_+^n$ is the stress energy vector, \mathbf{C}_e is the compliance matrix of the e -th element,

$$\mathcal{S}_a^+ = \{ \boldsymbol{\sigma} \in \mathbb{R}^p \mid \mathbf{D}^* \boldsymbol{\sigma} = \mathbf{f}, \|\boldsymbol{\sigma}\|_g \leq \sigma_c \}, \quad (117)$$

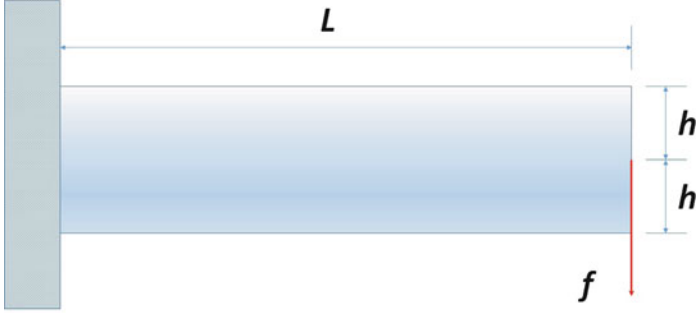


Fig. 2 The design domain for a long cantilever beam with external load

in which $\sigma_c > 0$ is a material constant, and $\mathbf{D}^* \in \mathbb{R}^{m \times p}$ is a balance operator depending on the polynomial interpolation in the mixed finite element method [10, 11].

Example Let us consider the 2-D classical long cantilever beam (see Figure 2). The correct topology optimization model for this benchmark problem should be (\mathcal{P}_{le}) [31]. We let $\omega = 0.4$ and the pre-given domain $\omega_0 = 1$ is discretized by $nex \times ney = 180 \times 60$ elements. Computational results obtained by the CDT and by the popular methods SIMP and BESO are summarized in Figure 3, where the $C = \mathbf{z}^T \mathbf{c}(\mathbf{u})$ is the total strain energy. The parameters used are $penal = 3$, $rmin = 1$ for BESO and $penal = 3$, $rmin = 1.5$, $ft = 1$ for SIMP. Clearly, the precise solid-void solution produced by the CDT method is much better than the approximate results produced by other methods. In order to look the strain energy distribution $\mathbf{c} = \{c_e(\mathbf{u})\}$ in the optimal structure, we let $nex \times ney = 80 \times 30$. Figure 4 shows clearly that the CDT can produce mechanically sound structure with homogeneous distribution of strain energy density. Detailed study on canonical duality theory for solving topology optimization problems is given recently in [30–32].

5 Symmetry, NP-Hardness, and Perturbation Methods

The concept of symmetry is closely related to the duality and, in certain sense, can be viewed as a *geometric duality*. Mathematically, symmetry means invariance under transformation. By the canonicity, the object $G(\mathbf{g})$ possesses naturally certain symmetry. If the subject $F(\chi) = 0$, then $\Pi(\chi) = G(\mathbf{D}\chi) = \Phi(\Lambda(\chi))$ and (\mathcal{P}_g) should have either a trivial solution or multiple solutions due to the symmetry. In this case, $\Pi^d(\zeta) = -\Phi^*(\zeta)$ is concave and, by the triality theory, its critical point $\bar{\zeta} \in \mathcal{S}_c^-$ is a global maximizer, and $\bar{\chi} = [\mathbf{G}(\bar{\zeta})]^+ \mathbf{f} = 0$ is the biggest local maximizer of $\Pi(\chi)$, while the global minimizers must be $\bar{\chi}(\bar{\zeta})$ for those $\bar{\zeta} \in \partial \mathcal{S}_c^+$ such that

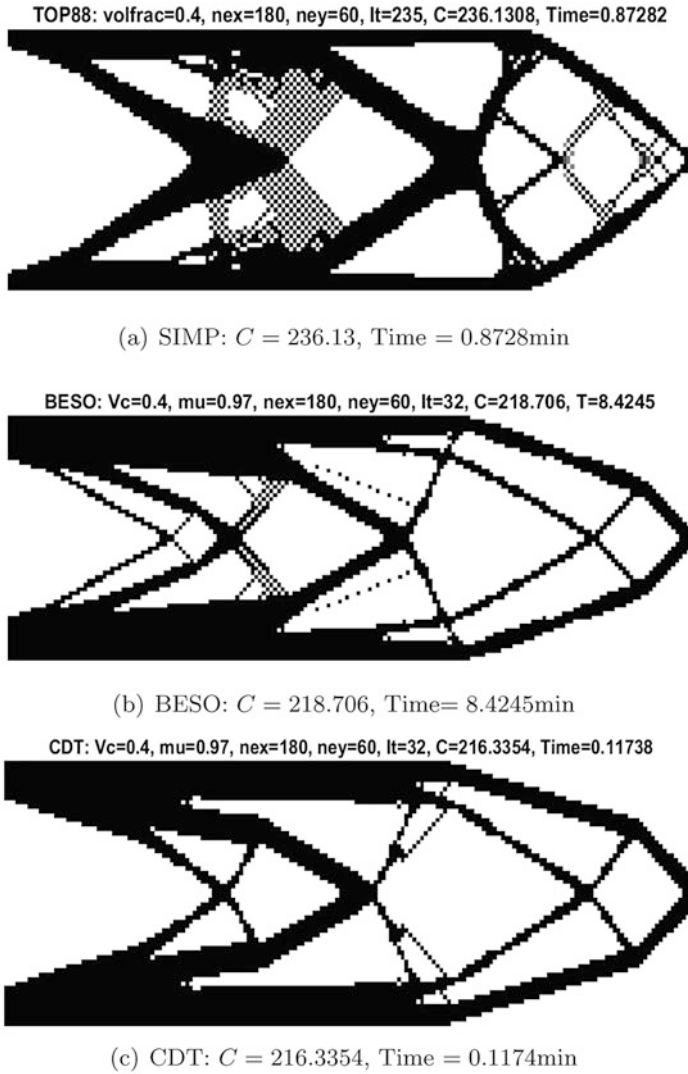


Fig. 3 Computational results by SIMP (a), BESO (b), and CDT (c) with $\omega = 0.4$.

$$\Pi^g(\bar{\zeta}) = \min\{-\Phi^*(\zeta) \mid \det \mathbf{G}(\zeta) = 0 \quad \forall \zeta \in \mathcal{S}_c\}. \quad (118)$$

Clearly, this nonconvex constrained concave minimization problem could be really NP-hard. Therefore, many well-known NP-hard problems in computer science and global optimization are not well-posed problems. Such as the *max-cut problem*, which is a special case of quadratic integer programming problem (\mathcal{P}_{qi}). Due to the symmetry $\mathbf{Q} = \mathbf{Q}^T$ and $\mathbf{f} = 0$, its canonical dual problem has multiple solutions

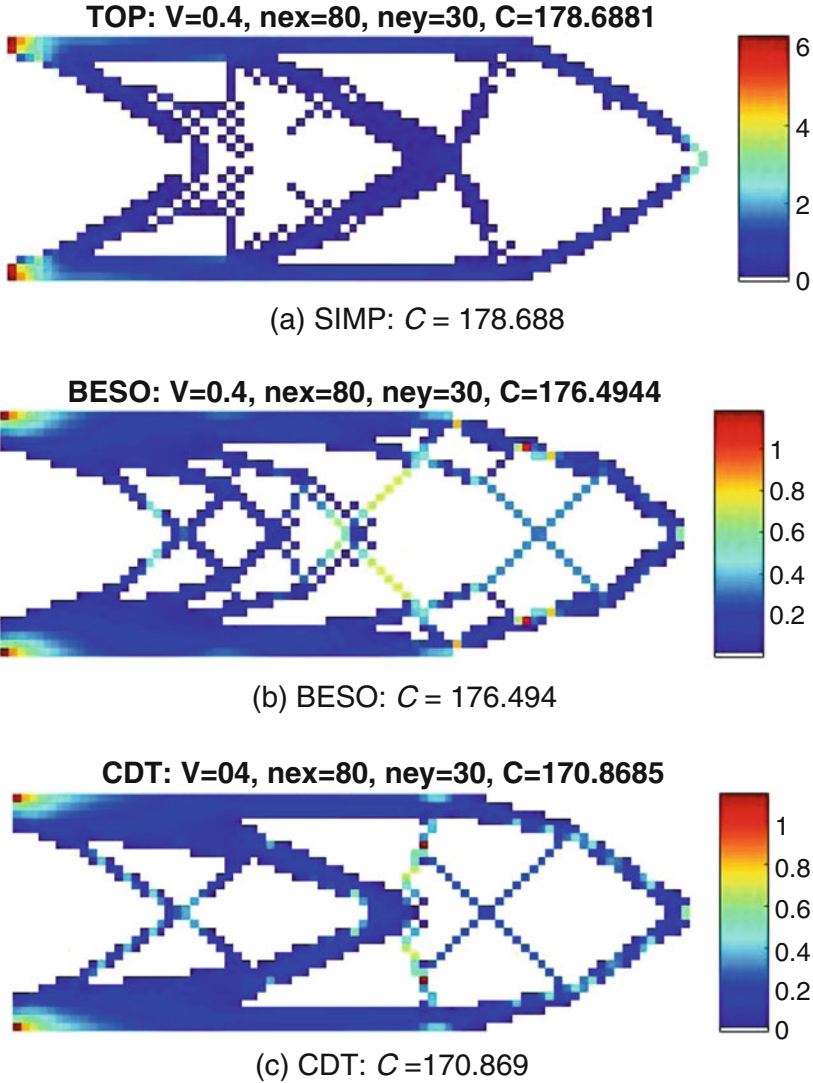


Fig. 4 Strain energy distributions by SIMP (a), BESO (b), and CDT (c) with $\omega = 0.4$.

on the boundary of \mathcal{S}_c^+ . The problem is considered as NP-complete even if $Q_{ij} = 1$ for all edges. Strictly speaking, this is not a real-world problem but only a perfect geometrical model. Without sufficient geometrical constraints in \mathcal{X}_a , the graph is not physically fixed and any rigid motion is possible. However, by adding a linear perturbation $\mathbf{f} \neq 0$, this problem can be solved efficiently by the canonical duality theory [85]. Also, it was proved by the author [25, 35] that the general quadratic integer problem (\mathcal{P}_{qi}) has a unique solution as long as the input $\mathbf{f} \neq 0$ is big enough.

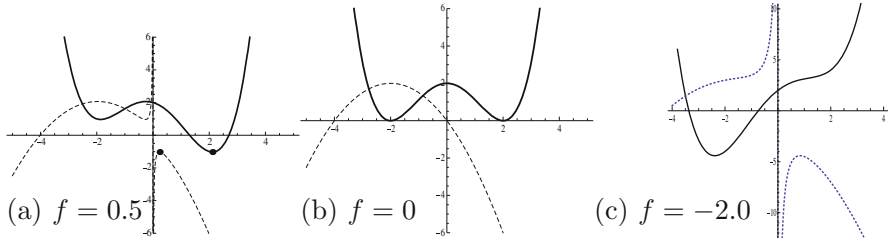


Fig. 5 Graphs of $\Pi(\mathbf{x})$ (solid) and $\Pi^g(\zeta)$ ($\alpha = 1, \lambda = 2$)

These results show that the subjective function plays an essential role for symmetry breaking which leads to a well-posed problem. To explain the theory and understand the NP-hard problems, let us consider a simple problem.

Example 1 (Nonconvex Minimization in \mathbb{R}^n)

$$\min \left\{ \Pi(\mathbf{x}) = \frac{1}{2}\alpha\left(\frac{1}{2}\|\mathbf{x}\|^2 - \lambda\right)^2 - \mathbf{x}^T \mathbf{f} \quad \forall \mathbf{x} \in \mathbb{R}^n \right\}, \quad (119)$$

where $\alpha, \lambda > 0$ are given parameters. Let $\Lambda(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 \in \mathbb{R}$, and the canonical dual function is $\Pi^d(\zeta) = -\frac{1}{2}\zeta^{-1}\|\mathbf{f}\|^2 - \lambda\zeta - \frac{1}{2}\alpha^{-1}\zeta^2$, which is defined on $\mathcal{S}_c = \{\zeta \in \mathbb{R} \mid \zeta \neq -\lambda, \zeta = 0 \text{ iff } \mathbf{f} = 0\}$. The criticality condition $\partial \Pi^d(\zeta) = 0$ leads to a canonical dual equation:

$$(\alpha^{-1}\zeta + \lambda)\zeta^2 = \frac{1}{2}\|\mathbf{f}\|^2. \quad (120)$$

This cubic equation has at most three real solutions satisfying $\zeta_1 \geq 0 \geq \zeta_2 \geq \zeta_3$, and, correspondingly, $\{\mathbf{x}_i = \mathbf{f}/\zeta_i\}$ are three critical points of $\Pi(\mathbf{x})$. By the fact that $\zeta_1 \in \mathcal{S}_a^+ = \{\zeta \in \mathbb{R} \mid \zeta \geq 0\}$, \mathbf{x}_1 is a global minimizer of $\Pi(\mathbf{x})$. While for $\zeta_2, \zeta_3 \in \mathcal{S}_a^- = \{\zeta \in \mathbb{R} \mid \zeta < 0\}$, \mathbf{x}_2 and \mathbf{x}_3 are local min (for $n = 1$) and local max of $\Pi(\mathbf{x})$, respectively (see Figure 5(a)).

If we let $\mathbf{f} = 0$, the graph of $\Pi(\mathbf{x})$ is symmetric (i.e., the so-called double-well potential or the Mexican hat for $n = 2$ [23]) with infinite number of global minimizers satisfying $\|\mathbf{x}\|^2 = 2\lambda$. In this case, the canonical dual $\Pi^g(\zeta) = -\frac{1}{2}\alpha^{-1}\zeta^2 - \lambda\zeta$ is strictly concave with only one critical point (local maximizer) $\zeta_3 = -\alpha\lambda < 0$. The corresponding solution $\mathbf{x}_3 = \mathbf{f}/\zeta_3 = 0$ is a local maximizer. By the canonical dual equation (120) we have $\zeta_1 = \zeta_2 = 0$ located on the boundary of \mathcal{S}_a^+ , which corresponding to the two global minimizers $x_{1,2} = \pm\sqrt{2\lambda}$ for $n = 1$, see Figure 1(b). If we let $f = -2$, then the graph of $\Pi(\mathbf{x})$ is quasi-convex with only one critical point and (120) has only one solution $\zeta_1 \in \mathcal{S}_c^+$ (see Figure 1(c)).

This simple example reveals an important truth, i.e., the symmetry is the key that leads to multiple solutions. Theoretically speaking, nothing is perfect in this real

world, a perfect symmetry is not allowed for any real-world problem. Thus, any real-world problem must be *well-posed* [29]. In reality, it is impossible to precisely model any real-world problem; although most of the NP-hard problems are artificial, they appear extensively not only in global optimization and computer science but also in chaotic dynamical systems, decision science, and philosophy, say, the well-known Buridan's ass paradox in its most simple version.

Example 2 (Paradox of Buridan's Ass and Perturbation) A donkey facing two identical hay piles starves to death because reason provides no grounds for choosing to eat one rather than the other.

The mathematical problem of this paradox was formulated in [31]:

$$\max\{c_1z_1 + c_2z_2 \mid c_1 = c_2 = c, z_1 + z_2 \leq 1, (z_1, z_2) \in \{0, 1\}^2\}. \quad (121)$$

Clearly, this is a linear knapsack problem in \mathbb{R}^2 . Due to the symmetries: $v_1 = v_2 = 1$ and $c_1 = c_2 = c$, the solution to (77) is $\tau_c = c$. Therefore, $\psi_i(\tau_c) = 0 \quad \forall i = 1, 2$ and by Theorem 2 this problem has multiple (two) solutions, which is NP-hard to this donkey.

In order to solve such NP-hard problems, the key idea is to break the symmetry. A linear perturbation method has been proposed by the author and his co-workers. This method is based on a simple truth, i.e., it is impossible to have the two identical hay piles. Thus, by adding a linear perturbation term $\epsilon\rho_1$ to the cost function to break the symmetry, then for $c = 2$, $\epsilon = 0.05$, the solution to (77) is $\tau_c = 2.0184$. So, the condition (75) holds for $i = 1, 2$ and by the canonical duality theory, the perturbed Buridan's ass problem has a unique solution $\mathbf{z} = (1, 0)$.

Perturbation method has been successfully applied for solving many challenging problems including hard cases of trust region method [7], NP-hard problems in integer programming [84, 85], and nonconvex constrained optimization in Euclidean geometry [62]. By the fact that the subjective function $F(\boldsymbol{\chi}) = \langle \boldsymbol{\chi}, \mathbf{f} \rangle$ plays a key role in real-world problems, the following conjecture was proposed recently [26, 28].

Conjecture 2 *For any given properly posed problem (\mathcal{P}_g) under the Assumption 1, there exists a constant $f_c > 0$ such that (\mathcal{P}^g) has a unique solution in \mathcal{S}_c^+ as long as $\|\mathbf{f}\| \geq f_c$.*

This conjecture shows that any properly posed problems are not NP-hard if the input $\|\mathbf{f}\|$ is big enough. Generally speaking, most NP-hard problems have multiple solutions located either on the boundary or the outside of \mathcal{S}_c^+ . Therefore, a quadratic perturbation method can be suggested as:

$$\begin{aligned} \Xi_{\delta_k}(\boldsymbol{\chi}, \boldsymbol{\varsigma}) &= \Xi(\boldsymbol{\chi}, \boldsymbol{\varsigma}) + \frac{1}{2}\delta_k\|\boldsymbol{\chi} - \boldsymbol{\chi}_k\|^2 \\ &= \frac{1}{2}\langle \boldsymbol{\chi}, \mathbf{G}_{\delta_k}(\boldsymbol{\varsigma})\boldsymbol{\chi} \rangle - \Phi^*(\boldsymbol{\varsigma}) - \langle \boldsymbol{\chi}, \mathbf{f}_{\delta_k} \rangle + \frac{1}{2}\delta_k\langle \boldsymbol{\chi}_k, \boldsymbol{\chi}_k \rangle, \end{aligned}$$

where $\delta_k > 0$, $\boldsymbol{\chi}_k$ ($k = 1, 2, \dots$) are perturbation parameters, $\mathbf{G}_{\delta_k}(\boldsymbol{\varsigma}) = \mathbf{G}(\boldsymbol{\varsigma}) + \delta_k \mathbf{I}$, and $\mathbf{f}_{\delta_k} = \mathbf{f} + \delta_k \boldsymbol{\chi}_k$. Thus, the original canonical dual feasible space \mathcal{S}_c^+ can be enlarged to $\mathcal{S}_{\delta_k}^+ = \{\boldsymbol{\varsigma} \in \mathcal{S}_c \mid \mathbf{G}_{\delta_k}(\boldsymbol{\varsigma}) \succ 0\}$ such that a perturbed canonical dual problem can be proposed as:

$$(\mathcal{P}_k^g) : \max \left\{ \min \{ \Xi_{\delta_k}(\boldsymbol{\chi}, \boldsymbol{\varsigma}) \mid \boldsymbol{\chi} \in \mathcal{X}_a \} \mid \boldsymbol{\varsigma} \in \mathcal{S}_{\delta_k}^+ \right\}. \quad (122)$$

Based on this problem, a canonical primal-dual algorithm has been developed with successful applications for solving sensor network optimization problems [70] and chaotic dynamics [54].

6 Connections with Popular Methods and Techniques

By the fact that the canonical duality-triality theory is a unified mathematical methodology with solid foundation in physics, it is naturally connected to many other powerful methods and techniques in different fields. This paper discusses only two well-known methods in optimization and a so-called composite minimization problem. Connections with other theories and methodologies can be found in [34, 56].

6.1 Relation with SDP Programming

Now, let us show the relation between the canonical duality theory and the popular semi-definite programming relaxation.

Theorem 8 *Suppose that $\Phi : \mathcal{E}_s \rightarrow \mathbb{R}$ is convex and $\bar{\boldsymbol{\varsigma}} \in \mathcal{E}_a^*$ is a solution of the problem:*

$$(\mathcal{P}^{sd}) : \min \{ g + \Phi^*(\boldsymbol{\varsigma}) \} \text{ s.t. } \begin{pmatrix} \mathbf{G}(\boldsymbol{\varsigma}) & \mathbf{f} \\ \mathbf{f}^T & 2g \end{pmatrix} \succeq 0 \quad \forall \boldsymbol{\varsigma} \in \mathcal{E}_a^*, \quad g \in \mathbb{R}, \quad (123)$$

then $\boldsymbol{\chi} = [\mathbf{G}(\boldsymbol{\varsigma})]^+ \mathbf{f}$ is a global minimum solution to the nonconvex problem (\mathcal{P}) .

Proof The problem (\mathcal{P}^d) can be equivalently written in the following problem (see [86]):

$$\min \left\{ g + \Phi^*(\boldsymbol{\varsigma}) \mid g \geq G_{ap}^\Delta(\boldsymbol{\varsigma}), \quad \mathbf{G}(\boldsymbol{\varsigma}) \succeq 0 \quad \forall \boldsymbol{\varsigma} \in \mathcal{E}_a^* \right\}. \quad (124)$$

Then, by using the Schur complement lemma, this problem is equivalent to (\mathcal{P}^{sd}) . The theorem is proved by the triality theory. \square

It was proved [35] that for the same problem (\mathcal{P}_{qi}) , if we use different geometrical operator:

$$\Lambda(\mathbf{x}) = \mathbf{xx}^T \in \mathcal{E}_a = \{\boldsymbol{\xi} \in \mathbb{R}^{n \times n} \mid \boldsymbol{\xi} = \boldsymbol{\xi}^T, \boldsymbol{\xi} \succeq 0, \\ \text{rank } \boldsymbol{\xi} = 1, \xi_{ii} = 1 \ \forall i = 1, \dots, n\},$$

and the associated canonical function $\Phi(\boldsymbol{\xi}) = \frac{1}{2}\langle \boldsymbol{\xi}; \mathbf{Q} \rangle + \{0 \text{ if } \boldsymbol{\xi} \in \mathcal{E}_a, +\infty \text{ otherwise}\}$, where $\langle \boldsymbol{\xi}; \boldsymbol{\zeta} \rangle = \text{tr}(\boldsymbol{\xi}^T \boldsymbol{\zeta})$, we should obtain the same canonical dual problem (\mathcal{P}_{qi}^d) . Particularly, if $\mathbf{f} = 0$, then (\mathcal{P}_{qi}) is a typical linear semi-definite programming:

$$\min \frac{1}{2}\langle \boldsymbol{\xi}; \mathbf{Q} \rangle \quad \text{s.t. } \boldsymbol{\xi} \in \mathcal{E}_a.$$

Since \mathcal{E}_a is not bounded and there is no input, this problem is not properly posed, which could have either no solution or multiple solutions for a given indefinite $\mathbf{Q} = \mathbf{Q}^T$.

The SDP programming has been used for solving a canonical dual problem in post-buckling analysis of a large deformed elastic beam [1].

6.2 Relation to Reformulation-Linearization/Convexification Technique

The *Reformulation-Linearization/Convexification Technique* (RLT) proposed by H. Sherali and C.H. Tuncbilek [75] is one well-known novel approach for efficiently solving general polynomial programming problems. The key idea of this technique is also to introduce a geometrically nonlinear operator $\boldsymbol{\xi} = \Lambda(\mathbf{x})$ such that the higher-order polynomial object $G(\mathbf{x})$ can be reduced to a lower-order polynomial $\Phi(\boldsymbol{\xi})$. Particularly, for the quadratic minimization problems with linear inequality constraints in \mathcal{X}_a :

$$(\mathcal{P}_q) : \min \left\{ \Pi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T \mathbf{Q} \mathbf{x} - \mathbf{x}^T \mathbf{f} \mid \mathbf{x} \in \mathcal{X}_a \right\}, \quad (125)$$

by choosing the quadratic transformation:

$$\boldsymbol{\xi} = \Lambda(\mathbf{x}) = \mathbf{x} \overrightarrow{\otimes} \mathbf{x} \in \mathcal{E}_a \subseteq \mathbb{R}^{n \times n}, \quad \text{i.e., } \boldsymbol{\xi} = \{\xi_{ij}\} = \{x_i x_j\}, \quad \forall 1 \leq i \leq j \leq n, \quad (126)$$

where $\overrightarrow{\otimes}$ represents the Kronecker product (avoiding symmetric terms, i.e., $\xi_{ij} = \xi_{ji}$), the quadratic object $G(\mathbf{g})$ can be reformulated as the following *first-level RLT linear relaxation*:

$$G(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} = \frac{1}{2} \sum_{k=1}^n q_{kk} \xi_{kk} + \sum_{k=1}^{n-1} \sum_{l=k+1}^n q_{kl} \xi_{kl} = \Phi(\xi). \quad (127)$$

The linear $\Phi(\xi)$ can be considered as a special canonical function since $\zeta = \partial\Phi(\xi)$ is a constant and $\Phi^*(\zeta) = \langle \xi; \zeta \rangle - \Phi(\xi) \equiv 0$ is uniquely defined. Thus, using $\Phi(\xi) = \langle \xi; \zeta \rangle$ to replace $G(\mathbf{x})$ and considering ξ as an independent variable, the problem (\mathcal{P}_q) can be relaxed by the following *RLT linear program*:

$$(\mathcal{P}_{RLT}) : \min \{ \Phi(\xi) - \langle \mathbf{x}, \mathbf{f} \rangle \mid \mathbf{x} \in \mathcal{X}_a, \xi \in \mathcal{E}_a \}. \quad (128)$$

Based on this RLT linear program, a branch and bound algorithm was designed [76]. It is proved that if $(\bar{\mathbf{x}}, \bar{\xi})$ solves (\mathcal{P}_{RLT}) , then its objective value yields a lower bound of (\mathcal{P}_q) and $\bar{\mathbf{x}}$ provides an upper bound for (\mathcal{P}_q) . Moreover, if $\bar{\xi} = \Lambda(\bar{\mathbf{x}}) = \bar{\mathbf{x}} \overrightarrow{\otimes} \bar{\mathbf{x}}$, then $\bar{\mathbf{x}}$ solves (\mathcal{P}_q) .

This technique has been significantly adapted along with supporting approximation procedures to solve a variety of more general nonconvex constrained optimization problems having polynomial or more general factorable objective and constraint functions [74].

By the fact that for any symmetric \mathbf{Q} , there exists $\mathbf{D} \in \mathbb{R}^{n \times m}$ such that $\mathbf{Q} = \mathbf{D}^T \mathbf{H} \mathbf{D}$ with $\mathbf{H} = \{h_{kk} = \pm 1, h_{kl} = 0 \ \forall k \neq l\} \in \mathbb{R}^{m \times m}$, the canonicity condition (127) can be simplified as:

$$G(\mathbf{D}\mathbf{x}) = \frac{1}{2} (\mathbf{D}\mathbf{x})^T \mathbf{H} (\mathbf{D}\mathbf{x}) = \frac{1}{2} \sum_{k=1}^m h_{kk} \xi_{kk} = \Phi(\xi), \quad (129)$$

$$\xi = \Lambda(\mathbf{x}) = (\mathbf{D}\mathbf{x}) \overrightarrow{\otimes} (\mathbf{D}\mathbf{x}) \in \mathbb{R}^{m \times m}. \quad (130)$$

Clearly, if the scale $m \ll n$, the problem (\mathcal{P}_{RLT}) will be much easier than the problems using the geometrically nonlinear operator $\xi = \mathbf{x} \overrightarrow{\otimes} \mathbf{x}$. Moreover, if we are using the Lagrange multiplier $\zeta \in \mathcal{E}_a^* = \{\zeta \in \mathbb{R}^{m \times m} \mid \langle \Lambda(\mathbf{x}); \zeta \rangle \geq 0 \ \forall \mathbf{x} \in \mathbb{R}^n\}$ to relax the ignored geometrical condition $\xi = \Lambda(\mathbf{x})$ in (\mathcal{P}_{RLT}) , the problem (\mathcal{P}_q) can be equivalently relaxed as:

$$(\mathcal{P}_\Upsilon) : \min_{\mathbf{x} \in \mathcal{X}_a} \min_{\xi \in \mathcal{E}_a} \max_{\zeta \in \mathcal{E}_a^*} \{ \Upsilon(\mathbf{x}, \xi, \zeta) = \Phi(\xi) + \langle \Lambda(\mathbf{x}) - \xi; \zeta \rangle - \langle \mathbf{x}, \mathbf{f} \rangle \}. \quad (131)$$

Thus, if $(\bar{\mathbf{x}}, \bar{\xi}, \bar{\zeta})$ is a solution to (\mathcal{P}_Υ) , then $\bar{\mathbf{x}}$ should be a solution to (\mathcal{P}_q) . By using the sequential canonical quadratic transformation $\Lambda(\mathbf{x}) = \Lambda_p(\dots(\Lambda_1(\mathbf{x})\dots)$ (see Chapter 4, [18]), this technique can be used for solving general global optimization problems.

6.3 Relation to Composite Minimization

The so-called composite minimization in optimization literature is given in the following form [57]:

$$\min_x h(c(x)), \quad (132)$$

where $c : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is called the inner function, and $h : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{-\infty, +\infty\}$ is called the outer function. Although there are some mathematical assumptions, such that $c(x)$ is smooth and $h(y)$ may be nonsmooth but is usually convex, this is another abstractly proposed problem. Therefore, this problem appears mainly from numerical approximation methods, for example, the least squares method for solving the fixed point problem (53).

In numerical analysis or *matrix completion* [5], the variable x is a $d \times n$ matrix $\mathbf{X} = \{\mathbf{x}_i\} = \{\mathbf{x}_i^\alpha\}$ ($\alpha = 1, \dots, d$, $i = 1, \dots, n$). In sensor network communication systems, the component $\mathbf{x}_i \in \mathbb{R}^d$ is the position of the i -th sensor and the well-studied sensor localization problem is to find the sensor locations $\{\mathbf{x}_i\}$ by solving the following nonlinear system [70]:

$$\|\mathbf{x}_i - \mathbf{x}_j\| = d_{ij} \quad \forall (i, j) \in \mathcal{A}_d, \quad \|\mathbf{x}_i - \mathbf{a}_k\| = e_{ik} \quad \forall (i, k) \in \mathcal{A}_e \quad (133)$$

where $\{d_{ij}\}$ and $\{e_{ik}\}$ are given distances, $\mathbf{a}_k \in \mathbb{R}^d$ ($k = 1, \dots, m$) are specified anchors, and \mathcal{A}_d and \mathcal{A}_e are two index sets. By the least squares method, this problem can be formulated as a fourth-order polynomial minimization:

$$\min \left\{ \Pi(\mathbf{X}) = \sum_{(i,j) \in \mathcal{A}_d} \frac{1}{2} w_{ij} (\|\mathbf{x}_i - \mathbf{x}_j\|^2 - d_{ij})^2 + \sum_{(i,k) \in \mathcal{A}_e} \frac{1}{2} \omega_{ik} (\|\mathbf{x}_i - \mathbf{a}_k\|^2 - e_{ik})^2 \right\},$$

where w_{ij} and ω_{ik} are given weights. Clearly, this is a composite minimization if we let

$$c(\mathbf{X}) = \{\mathbf{c}_{ij}(\mathbf{X}), \mathbf{c}_{ik}(\mathbf{X})\}, \quad \mathbf{c}_{ij} = \mathbf{x}_i - \mathbf{x}_j, \quad \mathbf{c}_{ik} = \mathbf{x}_i - \mathbf{a}_k, \quad (134)$$

$$h(c) = \sum_{(i,j) \in \mathcal{A}_d} \frac{1}{2} w_{ij} (\|\mathbf{c}_{ij}\|^2 - d_{ij})^2 + \sum_{(i,k) \in \mathcal{A}_e} \frac{1}{2} \omega_{ik} (\|\mathbf{c}_{ik}\| - e_{ik})^2. \quad (135)$$

In this case, the matrix-valued function $c(\mathbf{X}) = \mathbf{g}(\mathbf{X}) = \mathbf{D}\mathbf{X} = \{\mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{a}_k\}$ is the finite difference operator in numerical analysis and $h(c) = G(\mathbf{g})$ is a fourth-order nonconvex polynomial of the linear operator $\mathbf{g} = \mathbf{D}\mathbf{X}$.

We can also let

$$c(\mathbf{X}) = \{c_{ij}(\mathbf{X}), c_{ik}(\mathbf{X})\}, \quad c_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2 - d_{ij}, \quad c_{ik} = \|\mathbf{x}_i - \mathbf{a}_k\|^2 - e_{ik}, \quad (136)$$

$$h(c) = \sum_{(i,j) \in \mathcal{A}_d} \frac{1}{2} w_{ij} c_{ij}^2 + \sum_{(i,k) \in \mathcal{A}_e} \frac{1}{2} \omega_{ik} c_{ik}^2. \quad (137)$$

In this case, $\Pi(\mathbf{X}) = h(c(\mathbf{X}))$ is also a composite function but now $c(\mathbf{X}) = \boldsymbol{\xi}(\mathbf{X}) = \Lambda(\mathbf{X})$ is a nonlinear operator and $\Phi(\boldsymbol{\xi}) = h(\boldsymbol{\xi})$ is a convex function. Therefore, the composition:

$$\Pi(\mathbf{X}) = h(c(\mathbf{X})) = G(\mathbf{D}\mathbf{X}) = \Phi(\Lambda(\mathbf{X}))$$

is indeed a canonical transformation. The sensor localization problem is considered to be NP-hard by traditional theories and methods even if $d = 1$ [3]. From the point view of the canonical duality theory, this problem has usually multiple global minimizers due to the lacking of the subjective function. Therefore, by introducing a linear perturbation $F(\mathbf{X}) = \langle \mathbf{X}, \mathbf{T} \rangle = \text{tr}(\mathbf{X}^T \mathbf{T})$, the perturbed sensor localization problem $\min\{\Pi(\mathbf{X}) = \Phi(\Lambda(\mathbf{X})) - F(\mathbf{X})\}$ can be solved deterministically by the canonical duality theory in polynomial time [55, 68, 70].

Generally speaking, the composite function is a special case of the canonical transformation $G(\mathbf{g}) = \Phi \circ \Lambda(\mathbf{g})$ if $h(y) = \Phi(y)$ is convex, $x = \mathbf{g}$, and $\Lambda(x) = c(x)$ as the geometrical measure. It is an objective function if $c(x) = x^T x$. In this case, $h(c(x))$ is the so-called convex composite function. In real-world applications, $\mathbf{g}(\mathbf{x})$ could be again a composite function. For multi-scale systems, \mathbf{g} can be defined by (see [45]):

$$\mathbf{g}(\mathbf{x}) = (\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_k)\mathbf{x} = \{\mathbf{g}_i(\mathbf{x})\}, \quad \mathbf{g}_i(\mathbf{x}) = \mathbf{D}_i \mathbf{x}, \quad (138)$$

each \mathbf{g}_i is a geometrical measure with dimension different from other \mathbf{g}_j , $j \neq i$. Correspondingly:

$$G(\mathbf{D}\mathbf{x}) = \Phi(\Lambda(\mathbf{x})), \quad \Lambda(\mathbf{x}) = \Lambda_k \circ \Lambda_{k-1} \circ \dots \circ \Lambda_1(\mathbf{x}) \quad (139)$$

is called the *sequential canonical transformation* (see Chapter 4, [18]). Particularly, if every $\Lambda_i(\boldsymbol{\xi}_{i-1})$ is a convex polynomial function of $\boldsymbol{\xi}_{i-1} = \Lambda_{i-1}$ ($i = 1, \dots, k$, $\Lambda_0 = \mathbf{x}$), the composition $\Phi(\Lambda(\mathbf{x}))$ is the *canonical polynomial function*. The sequential canonical transformation for solving high-order polynomial minimization problems have been studied in [18, 46].

7 Conclusions

Based on the necessary conditions and basic laws in physics, a unified multi-scale global optimization problem is proposed in the canonical form:

$$\Pi(\boldsymbol{\chi}) = G(\mathbf{D}\boldsymbol{\chi}) - F(\boldsymbol{\chi}) = \Phi(\Lambda(\boldsymbol{\chi})) - \langle \boldsymbol{\chi}, \mathbf{f} \rangle. \quad (140)$$

The object G depends only on the model and $G(\mathbf{g}) \geq 0 \forall \mathbf{g} \in \mathcal{G}_a$ is necessary; G should be an objective function for physical systems, but it is not necessary for artificial systems (such as management/manufacturing processes and numerical simulations, etc.). The subject F depends on each properly posed problem and must satisfy $F(\boldsymbol{\chi}) \geq 0$ together with necessary geometrical constraints for the output $\boldsymbol{\chi} \in \mathcal{X}_a$ and equilibrium conditions for the input $\mathbf{f} \in \mathcal{X}_a^*$. The geometrical nonlinearity of $\Lambda(\boldsymbol{\chi})$ is necessary for nonconvexity in global optimization, bifurcation in nonlinear analysis, chaos in dynamics, and NP-hardness in computer science.

Developed from large deformation nonconvex analysis/mechanics, the canonical duality-triality is a precise mathematical theory with solid foundation in physics and natural root in philosophy, so it is naturally related to the traditional theories and powerful methods in global optimization and nonlinear analysis. By the fact that the canonical duality is a universal law of nature, this theory can be used not only to model real-world problems but also for solving a wide class of challenging problems in multi-scale complex systems. The conjectures proposed in this paper can be used for understanding and clarifying NP-hard problems.

It is author's hope that by reading this paper, the readers can have a clear understanding not only on the canonical duality-triality theory and its potential applications in multi-disciplinary fields, but also on the generalized duality-triality principle and its role in modeling/understanding real-world problems.

Acknowledgements This paper is based a series colloquiums/seminars and plenary/keynote lectures presented at institutions and international conferences during the past five years. The author expresses his sincere appreciation to his colleagues, co-workers, and students for their constant supports and excellent collaborations for developing this unconventional theory in multi-disciplinary fields. Valuable comments and advices from Professor Hanif Sherali at Virginia Tech and Professor Shu-Cherng Fang at North Carolina State University are highly appreciated. This research has been continuously supported by the US Air Force Office of Scientific Research under the grants FA9550-10-1-0487, FA2386-16-1-4082, and FA9550-17-1-0151.

References

1. Ali, E.J. and Gao, D.Y. (2017). Improved canonical dual finite element method and algorithm for post-buckling analysis of nonlinear Gao beam. In *Canonical Duality Theory*, D.Y. Gao, N. Ruan and V. Latorre (Eds), pages 277–289. Springer, 2017.
2. Anorld, VI (1998) On teaching mathematics, *Russian Math. Surveys*, 53 (1), 229–236.
3. Aspnes, J., Goldberg, D. and Yang, Y.R. On the computational complexity of sensor network localization, in: *Lecture Notes in Computer Science*, 3121, Springer-Verlag, 2004, pp. 3244.
4. Bader B W, Kolda T G (2006). Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Trans Math Software*, 32(4): 635–653.
5. Candés, E. and Recht, B. (2008). Exact matrix completion via convex optimization, Technical Report, California Institute of Technology, 2008.
6. Chen Y. and Gao, D.Y. (2016). Global solutions to nonconvex optimization of 4th-order polynomial and log-sum-exp functions, *J. Global Optimization*, 64(3), 417–431.
7. Chen Y. and Gao, D.Y. (2017). Global solutions to spherically constrained quadratic minimization via canonical duality theory, in *Canonical Duality Theory: Unified Methodology for Multidisciplinary Study*, DY Gao, V. Latorre and N. Ruan (eds), Springer, New York, pp 291–314.

8. Ciarlet, PG (2013). *Linear and Nonlinear Functional Analysis with Applications*, SIAM, Philadelphia.
9. Feynman R, Leighton R, and Sands M. *The Feynman Lectures on Physics*, Volume II, 1964.
10. Gao, DY (1988). Panpenalty finite element programming for limit analysis, *Computers & Structures*, 28, pp. 749–755.
11. Gao, DY (1996). Complementary finite element method for finite deformation nonsmooth mechanics. *J. Eng. Math.* 30, 339–353.
12. Gao, D.Y. (1997). Dual extremum principles in finite deformation theory with applications to post-buckling analysis of extended nonlinear beam theory. *Applied Mechanics Reviews*, 50 (11), S64-S71 (1997).
13. Gao, D.Y. (1998). Duality, triality and complementary extremum principles in nonconvex parametric variational problems with applications, *IMA J. Appl. Math.*, 61, 199–235.
14. Gao, D.Y. (1998). Bi-complementarity and duality: A framework in nonlinear equilibria with applications to the contact problems of elastoplastic beam theory, *J. Math. Anal. Appl.*, 221, 672–697.
15. Gao, D.Y. (1999). Pure complementary energy principle and triality theory in finite elasticity, *Mech. Res. Comm.* 26 (1), 31–37.
16. Gao, D.Y. (1999). *Duality-Mathematics*, Wiley Encyclopedia of Electrical and Electronics Engineering, 6, 68–77.
17. Gao, D.Y. (1999). General Analytic Solutions and Complementary Variational Principles for Large Deformation Nonsmooth Mechanics. *Meccanica* 34, 169–198.
18. Gao, D.Y. (2000). *Duality Principles in Nonconvex Systems: Theory, Methods and Applications*. Springer, New York/Boston, 454pp.
19. Gao, D.Y. (2000). Analytic solution and triality theory for nonconvex and nonsmooth variational problems with applications, *Nonlinear Analysis*, 42, 7, 1161–1193.
20. Gao, D. Y. *Canonical dual transformation method and generalized triality theory in nonsmooth global optimization*. *J. Glob. Optim.* 17(1/4), pp. 127–160 (2000).
21. Gao, D.Y. (2001). Complementarity, polarity and triality in nonsmooth, nonconvex and nonconservative Hamiltonian systems. *Phil. Trans. R. Soc. London A* 359, 2347–2367.
22. Gao, D.Y. (2003). Perfect duality theory and complete solutions to a class of global optimization problems. *Optimization* 52(4–5), 467–493
23. Gao, D.Y. (2003). Nonconvex semi-linear problems and canonical duality solutions. *Advances in Mechanics and Mathematics*, II, Springer, 261–311.
24. Gao, D.Y. (2007). Solutions and optimality to box constrained nonconvex minimization problems *J. Indust. Manage. Optim.*, 3(2), 293–304.
25. Gao, D.Y. (2009). Canonical duality theory: unified understanding and generalized solutions for global optimization. *Comput. & Chem. Eng.* 33, 1964–1972.
26. Gao, D.Y. (2014). Unified modeling and theory for global optimization. Plenary Lecture at *16th Baikal Int. Seminar on Methods of Optimization and Their Applications*, June 30 - July 6, 2014, Olkhon, Russia.
27. Gao, D.Y. (2016). Analytical solutions to general anti-plane shear problems in finite elasticity, *Continuum Mech. Thermodyn.* 28:175–194
28. Gao, D.Y. (2016). On unified modeling, theory, and method for solving multi-scale global optimization problems, in *Numerical Computations: Theory And Algorithms*, (Editors) Y. D. Sergeyev, D. E. Kvasov and M. S. Mukhametzhanov, AIP Conference Proceedings 1776, 020005.
29. Gao, D.Y. (2016). On unified modeling, canonical duality-triality theory, challenges and breakthrough in optimization, <https://arxiv.org/abs/1605.05534> .
30. Gao, D.Y. (2017). Canonical Duality Theory for Topology Optimization, *Canonical Duality-Triality: Unified Theory and Methodology for Multidisciplinary Study*, D.Y. Gao, N. Ruan and V. Latorre (Eds). Springer, New York, pp.263–276.
31. Gao, D.Y. (2018). On topology optimization and canonical duality method, *Computer Methods in Applied Mechanics and Engineering*, 341, 249–277.

32. Gao, D.Y. and Ali, E.J. (2018). A novel canonical duality theory for 3-D topology optimization, *Emerging Trends in Applied Mathematics and High-Performance Computing*, V.K. Singh, D.Y. Gao and A. Fisher (Eds). Springer, New York.
33. Gao, D.Y. and Hajilarov, E. Analytic solutions to 3-D finite deformation problems governed by St Venant–Kirchhoff material. In DY Gao, V. Latorre, and N. Ruan, editors, *Canonical Duality Theory: Unified Methodology for Multidisciplinary Study*, pages 69–88. Springer, New York, 2017.
34. Gao, D.Y., Latorre, V. and Ruan, N. (2017). *Canonical Duality Theory: Unified Methodology for Multidisciplinary Study*, Springer, New York, 377pp.
35. Gao, D.Y. and Ruan, N.: Solutions to quadratic minimization problems with box and integer constraints. *J. Glob. Optim.* 47, 463–484 (2010).
36. Gao, D.Y., Ogden, R.W. (2008). Multi-solutions to non-convex variational problems with implications for phase transitions and numerical computation. *Q. J. Mech. Appl. Math.* 61, 497–522.
37. Gao, D.Y., Ogden, R.W. (2008). Closed-form solutions, extremality and nonsmoothness criteria in a large deformation elasticity problem. *Zeits. Ang. Math. Physik* 59, 498–517.
38. Gao, DY, Ogden, RW, Stavroulakis, G (2001). *Nonsmooth and Nonconvex Mechanics: Modelling, Analysis and Numerical Methods*. Kluwer Academic Publishers.
39. Gao, DY, Ruan, N, and Latorre, V (2017). Canonical duality-triality theory: Bridge between nonconvex analysis/mechanics and global optimization in complex systems, in *Canonical Duality Theory: Unified Methodology for Multidisciplinary Study*, DY Gao, V. Latorre and N. Ruan (eds), Springer, New York, pp 1–48.
40. Gao, D.Y., Ruan, N., Sherali, H. (2009). Solutions and optimality criteria for nonconvex constrained global optimization problems with connections between canonical and Lagrangian duality. *J. Glob. Optim.* 45, 473–497.
41. Gao, D.Y., Ruan, N., Sherali, H. (2010). Canonical dual solutions for fixed cost quadratic programs, *Optimization and Optimal Control*, A. Chinchuluun et al. (eds.), Springer Optimization and Applications 39.
42. Gao, D.Y. and Sherali, H.D. (2009). Canonical Duality Theory: Connection between nonconvex mechanics and global optimization, *Advances in Appl. Math. and Global Optimization*, D.Y. Gao and H. Sherali (eds). Springer.
43. Gao, D.Y., Strang, G.: Geometric nonlinearity: Potential energy, complementary energy, and the gap function. *Quart. Appl. Math.* 47(3), 487–504 (1989).
44. Gao, D.Y. and Wu, C.Z. (2012). On the triality theory for a quartic polynomial optimization problem, *J. Ind. Manag. Optim.* 8(1), 229–242.
45. Gao, D.Y., Yu, H.F. (2008). Multi-scale modelling and canonical dual finite element method in phase transitions of solids. *Int. J. Solids Struct.* 45, 3660–3673.
46. Gao, T.K. (2013). Complete solutions to a class of eighth-order polynomial optimization problems, *IMA J Appl Math*, 80(1), 158–176.
47. Hiriart-Urruty, J.B. (1985). Generalized differentiability, duality and optimization for problems dealing with differences of convex functions. *Lecture Note Econ. Math. Syst.*, 256: 37–70.
48. Horst, R., Thoai, N.V. (1999). DC Programming: overview. *J. Opt. Theory Appl.*, 103: 1–43.
49. Jin, Z. and Gao, D.Y. (2017). On modeling and global solutions for d.c. optimization problems by canonical duality theory, *Applied Mathematics and Computation*, 296, 168–181
50. Karp, R.M. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103, New York: Plenum, 1972.
51. Lagrange, Joseph-Louis (1811). *Mecanique Analytique*. Courcier, (reissued by Cambridge Univ. Press, 2009).
52. Landau, L.D. and Lifshitz, E.M. (1976). *Mechanics*. Vol. 1 (3rd ed.). Butterworth-Heinemann.
53. Latorre, V. and Gao, D.Y. (2016). Canonical duality for solving general nonconvex constrained problems, *Optimization Letters*, 10(8):1763–1779. <http://link.springer.com/article/10.1007/s11590-015-0860-0/fulltext.html>
54. Latorre, V. and Gao, D.Y. (2016). Global optimal trajectory in chaos and NP-Hardness. *Int. J. Bifurcation and Chaos*, 26, 1650142 (2016) <https://doi.org/10.1142/S021812741650142X>

55. Latorre, V. and Gao, D.Y. (2018). Canonical duality theory for solving large-sized noisy sensor network localization problems. to appear in *IEEE Transactions on Cybernetics*
56. Latorre, V. and Gao, D.Y. (2018). On half-quadratic reformulation and canonical duality theory in image restoration. Submitted.
57. Lewis A. S. and Wright, S. J. (2016). A proximal method for composite minimization, *Mathematical Programming: Series A and B archive*, 158 (1–2): 501–546
58. Li, S.F. and Gupta, A. (2006). On dual configuration forces, *J. of Elasticity*, 84:13–31.
59. Liu, I.-S. (2005). Further remarks on Euclidean objectivity and the principle of material frame-indifference. *Continuum Mech. Thermodyn.*, 17, 125–133
60. Marsden, J.E. and Hughes, T.J.R. : *Mathematical Foundations of Elasticity*, Prentice-Hall, 1983.
61. Morales, D.M. and Gao, D.Y. (2015). Canonical duality theory and triality for solving general unconstrained global optimization problems, *Math. Mech. Complex Systems*, Vol. 3 (2015), No. 2, 139–161.
62. Morales, D.M. and Gao, D.Y. (2017). On minimum distance between two surfaces, in *Canonical Duality Theory: Unified Methodology for Multidisciplinary Study*, DY Gao, V. Latorre and N. Ruan (eds), Springer, New York, pp 359–372.
63. Moreau, J.J. (1968). La notion de sur-potentiel et les liaisons unilatérales en élastostatique, *C.R. Acad. Sc. Paris*, 267 A, 954–957.
64. Murdoch, A.I.(2005). On criticism of the nature of objectivity in classical continuum physics, *Continuum Mech. Thermodyn.*, 17(2):135–148
65. Oden, J.T. *An Introduction to Mathematical Modeling*. John Wiley & Sons, 2011.
66. Puchinger, J., Raidl, G.R., Pferschy, U. (2010). The multidimensional knapsack problem: structure and algorithms, *INFORMS J. Comput.* 22 (2): 250–265 .
67. Qi, L.Q., Chen, H. and Chen, Y. (2018) *Tensor Eigenvalues and Their Applications*, Springer, New York, 329pp
68. Ruan, N. and Gao, D.Y. (2014). Global optimal solutions to a general sensor network localization problem, *Performance Evaluations*, 75–76: 1–16.
69. Ruan, N. and Gao, D.Y. (2018). Global optimal solution to quadratic discrete programming problem with inequality constraints. In *Canonical Duality-Triality: Unified Theory and Methodology for Multidisciplinary Study*, D.Y. Gao, V. Latorre and N. Ruan (Eds). Springer, New York, pp.315–338. <http://arxiv.org/abs/1205.0856>
70. Ruan, N. and Gao, D.Y. (2014). Global optimal solutions to a general sensor network localization problem, *Performance Evaluations*, 75–76: 1–16.
71. Ruan, N. and Gao, D.Y.(2014). Canonical duality approach for nonlinear dynamical systems, *IMA J. Appl. Math.*, 79: 313–325.
72. Ruan, N. and Gao, D.Y. (2018). On Modelling and Complete Solutions to General Fixpoint Problems in Multi-Scale Systems with Applications, <https://arxiv.org/abs/1801.08651>
73. Santos, H.A.F.A. and Gao D.Y. (2011) Canonical dual finite element method for solving post-buckling problems of a large deformation elastic beam, *Int. J. Nonlinear Mechanics*, 47: 240–247.
74. Sherali, H. D. (2002). Tight Relaxations for Nonconvex Optimization Problems Using the Reformulation- Linearization/Convexification Technique (RLT), *Handbook of Global Optimization, Volume 2: Heuristic Approaches*, P. M. Pardalos and H. E. Romeijn, Eds., Kluwer Academic Publishers, 1–63.
75. Sherali, H.D. and Tuncbilek, C.H. (1992). A Global Optimization Algorithm for Polynomial Programming Problems Using a Reformulation-Linearization Technique, *J. Global Optim.*, Vol. 2, No. 1, pp. 101–112.
76. Sherali, H.D. and Tuncbilek, C.H. (1995). A reformulation-convexification approach for solving nonconvex quadratic programming problems. *J. Global Optimization*, 7:1–31.
77. Strang, G. (1986). *Introduction to Applied Mathematics*, Wellesley-Cambridge Press.
78. Strugariu, R. , Voisei, M.D. and Zalinescu, C. : Counter-examples in bi-duality, triality and tri-duality, *Discrete & Continuous Dynamical Systems - A*, 2011, 31(4): 1453–1468.

79. Pham Dinh Tao, Le Thi Hoai An(2014). Recent Advances in DC Programming and DCA. *Transactions on Computational Collective Intelligence*, 13: 1–37.
80. Truesdell, C. and Noll, W. (1965). *The Nonlinear Field Theories of Mechanics*, Springer-Verlag, 591pp.
81. Toland, J.F.(1979). A duality principle for non-convex optimisation and the calculus of variations. *Arch. Ration. Mech. Anal.*, 71: 41–61.
82. Tuy, H.(1995). D.C. optimization: Theory, methods and algorithms. In: Horst, R., Pardalos, P.M. (eds.) *Handbook of Global Optimization*, pp. 149–216. Kluwer Academic Publishers, Dordrecht.
83. Voisei, M.D. and Zălinescu , C.(2011). Some remarks concerning Gao-Strang’s complementary gap function, *Applicable Analysis*, Vol. 90, No. 6, 1111–1121.
84. Wang, Z.B., Fang, S.C., Gao, D.Y. and Xing, W.X. (2008). Global extremal conditions for multi-integer quadratic programming. *J Industrial and Management Optimization*, 4(2):213.
85. Wang, Z.B., Fang, S.C., Gao, D.Y. and Xing, W.X. (2012). Canonical dual approach to solving the maximum cut problem, *J. Glob. Optim.*, 54, 341–351.
86. Zhou, X.J., Gao, D.Y. and Yang, C.H. (2016) Global solutions to a class of CEC benchmark constrained optimization problems, *Optim Lett*, 10:457–472 <https://doi.org/10.1007/s11590-014-0784-0>

Numerical Investigation of Stochastic Neural Field Equations



Pedro M. Lima

1 Introduction

Neural field equations (NFE) are a powerful tool for analysing the dynamical behaviour of populations of neurons. The analysis of such dynamical mechanisms is crucially important for understanding a wide range of neurobiological phenomena [3]. In this work, we will be concerned with the NFE in the form:

$$\frac{\partial}{\partial t} V(x, t) = I(x, t) - \alpha V(x, t) + \int_{\Omega} K(|x - y|) S(V(y, t - \tau(x, y))) dy, \quad (1)$$
$$t \in [0, T], \quad x \in \Omega \subset \mathbb{R},$$

where $V(x, t)$ (the unknown function) denotes the membrane potential in point x at time t ; $I(x, t)$ represents the external sources of excitation; S is the dependence between the firing rate of the neurons and their membrane potentials (sigmoidal or Heaviside function); and $K(|x - y|)$ gives the connectivity between neurons at x and y , α is a constant (related to the decay rate of the potential); $\tau(x, y) > 0$ is a delay, depending on the spatial variables (it results from the finite propagation speed of nervous stimulus in the brain).

Equation (1) (without delay) was introduced first by Wilson and Cowan [18], and then by Amari [1], to describe excitatory and inhibitory interactions in populations of neurons.

P. M. Lima (✉)

Departamento de Matemática, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
e-mail: plima@math.ist.utl.pt

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_2

Intensive studies of Hopf bifurcations occurring in neural fields have been carried out in the last decade. In [3], the authors investigate the occurrence of spatially localised oscillations (or breathers) in two-dimensional neural fields with excitatory and inhibitory interactions; in [16], the authors obtain sufficient conditions for the stability of stationary solutions of neural field equations; the dependency of the stationary solutions of NFE with respect to the stiffness of the nonlinearity and the contrast of external inputs is studied in [17]. The effect of transmission delays on the stability and the bifurcations was analysed in [7] (in the case of a single delay) and in [2] (in the case of distributed delays).

Though the above-mentioned results were obtained analytically, numerical simulations play a fundamental role in studying brain dynamics. Thus, the availability of efficient numerical methods is an important ingredient for improving the understanding of neural processes. Concerning Equation (1), numerical approximations were obtained in [4]. The computational method applies quadrature rule in space to reduce the problem to a system of delay differential equations, which is then solved by a standard algorithm for this kind of equations. A more efficient approach was recently proposed in [9, 10], where the authors introduce a new approach to deal with the convolution kernel of the equation and use fast Fourier transforms to reduce significantly the computational effort required by numerical integration. Recently, a new numerical method for the approximation of two-dimensional neural fields has been introduced, based on an implicit second-order scheme for the integration in time and using Chebyshev interpolation to reduce the dimensions of the matrices [14]. Some applications of this algorithm to Neuroscience problems have been discussed in [15].

As in other sciences, in Neurobiology it is well known that better consistency with some phenomena can be provided if the effects of random processes in the system are taken into account. The conjugate role of noise and delays on the genesis of bifurcations and pattern formation was investigated in [8].

In recent work of Kühn and Riedler [12], the authors study the effect of additive noise in Neural Field Equations. With this purpose, they introduce the stochastic integro-differential equation:

$$dU_t(x) = \left(I(x, t) - \alpha U_t(x) + \int_{\Omega} K(|x - y|) S(U_t(y)) dy \right) dt + \epsilon dW_t(x), \quad (2)$$

where $t \in [0, T]$, $x \in \Omega \subset \mathbb{R}^n$, and W_t is a Q-Wiener process.

The main goal of the present work is to analyse the effect of noise in certain neural fields, which allow different types of stationary solutions. In this case, we consider the following modification of equation (2):

$$dU_t(x) = \left(I(x, t) - \alpha U_t(x) + \int_{\Omega} K(|x - y|) S(U_{t-\tau}(y)) dy \right) dt + \epsilon dW_t(x), \quad (3)$$

where, as in the deterministic case, τ is a delay, depending on the distance $|x - y|$. Equation (3) is completed with an initial condition of the form:

$$U_t(x) = U_0(x, t), \quad t \in [-\tau_{max}, 0], \quad x \in \Omega, \quad (4)$$

where $U_0(x, t)$ is some given stochastic process, and τ_{max} is the maximum value of the delay ($\tau_{max} = |\Omega|/v$), where v is the propagation speed of the signals. We assume that $U_t(x)$ satisfies periodic boundary conditions in space. We will consider domains of the form $\Omega = [-l, l]$, including the limit case when $l \rightarrow \infty$.

2 Numerical Approximation

To construct a numerical approximation of the solution of (2) in the one-dimensional case, we begin by expanding the solution $U_t(x)$ using the Karhunen-Loeve formula:

$$U_t(x) = \sum_{k=0}^{\infty} u_t^k v_k(x), \quad (5)$$

where v_k are the eigenfunctions of the covariance operator of the noise in (2), which form an orthogonal system (their explicit form is indicated below). To derive a formula for the coefficients u_t^k , we take the inner product of equation (2) with the basis functions v_i :

$$\begin{aligned} (dU_t, v_i) = & \left[(I(x, t), v_i) - \alpha(U_t, v_i) + \left(\int_{\Omega} K(|x - y|) S(U_{t-\tau}(y)) dy, v_i \right) \right] dt \\ & + \epsilon (dW_t, v_i). \end{aligned} \quad (6)$$

We expand dW_t as:

$$dW_t(x) = \sum_{k=0}^{\infty} v_k(x) \lambda_k d\beta_t^k, \quad (7)$$

where the functions β_t^k form a system of independent white noises in time and λ_k are the eigenvalues of the covariance operator of the noise. As an important particular case, we consider the one described in [12], p.7. In this case, the correlation function satisfies

$$E W_t(x) W_s(y) = \min(t, s) \frac{1}{2\xi} \exp\left(\frac{-\pi}{4} \frac{|x - y|^2}{\xi^2}\right),$$

where ξ is a parameter modeling the spatial correlation length. In this case, if $\xi \ll 2l$, the eigenvalues of the covariance operator satisfy

$$\lambda_k^2 = \exp\left(-\frac{\xi^2 k^2}{4\pi}\right).$$

By substituting (5) into (6) and taking into account (7) and the orthogonality of the system v_k , we obtain

$$du_t^i = \left[(I(x, t), v_i) - \alpha u_t^i + (KS)^i(\bar{u}_{t-\tau}) \right] dt + \epsilon \lambda_i d\beta_t^i, \quad (8)$$

where $(KS)^i(\bar{u}_t)$ denotes the nonlinear term of the system:

$$(KS)^i(\bar{u}_t) = \int_{\Omega} v_i(x) \left(\int_{\Omega} K(|x-y|) S \left(\sum_{k=1}^{\infty} u_{t-\tau}^k v_k(y) \right) dy \right) dx. \quad (9)$$

When using the Galerkin method, we define an approximate solution by truncating the series expansion (5):

$$U_t^N(x) = \sum_{k=0}^{N-1} u_t^{k,N} v_k(x). \quad (10)$$

Then, the coefficients $u_t^{k,N}$ satisfy the following nonlinear system of stochastic delay differential equations:

$$du_t^{i,N} = \left[(I(x, t), v_i) - \alpha u_t^{i,N} + (KS)^{i,N}(\bar{u}_{t-\tau}) \right] dt + \epsilon \lambda_i d\beta_t^i, \quad (11)$$

where $(KS)^{i,N}(\bar{u}_t)$ is given by:

$$(KS)^{i,N}(\bar{u}_{t-\tau}) = h^2 \sum_{l=1}^N v_i(x_l) \left(\sum_{j=1}^N K(|x_l - x_j|) S \left(\sum_{k=1}^N u_{t-\tau}^k v_k(x_j) \right) \right) \quad (12)$$

$i = 0, \dots, N-1$. In this case, we are introducing in $[-L, L]$ a set of $N+1$ equidistant grid points $x_j = -l + j * h$, $j = 0, \dots, N$, where $h = 2l/N$, and using the rectangular rule to evaluate the integral in (9).

Since the problem has been reduced to the system (11) (a system of nonlinear stochastic delay differential equations), we can apply the Euler-Maruyama method to the solution of this system. For the discretisation in time, we introduce on the interval $[0, T]$ a uniform mesh with step size h_t , such that $t_j = jh_t$, $j = 0, 1, \dots, n$. Then, the solution $u_t^{k,N}$ of (11) will be approximated by a vector $(u_1^{k,N}, u_2^{k,N}, \dots, u_n^{k,N})$, where:

$$u_j^{k,N} \approx u_{t_j}^{k,N}.$$

In these notations, the Euler-Maruyama method may be written as:

$$u_{j+1}^{i,N} = u_j^{i,N} + h_t \left[(I(x_i, t_j), v_i) - \alpha u_{j+1}^{i,N} + (KS)^{i,N}(\bar{u}_{t_j-\tau}) \right] + \sqrt{h_t} \epsilon \lambda_i w_i, \quad (13)$$

where w_i is a random variable with normal distribution ($w_i = N(0, 1)$), $j = 0, \dots, n$, $i = 0, \dots, N - 1$. In the right-hand side of (13), we have written $\alpha u_{j+1}^{i,N}$, meaning that we are using a *semi-implicit* version of the Euler-Maruyama method.

Further, we can rewrite equations (13) in the form:

$$u_{j+1}^{i,N} = \frac{u_j^{i,N} + h_t \left[(I(x_i, t_j), v_i) + (KS)^{i,N}(\bar{u}_{t_j}) \right] + \sqrt{h_t} \epsilon \lambda_i w_i}{1 + \alpha h_t}. \quad (14)$$

The meaning of the inner product in (14) will be explained below. In order to compute $u_{t_j-\tau}^k$, we should take into account that $\tau = \frac{|x_{k_1} - x_{k_2}|}{v}$ (the time spent by the signal to travel between x_{k_1} and x_{k_2}). In general, τ may not be a multiple of h_t . Let d and δ_t be the integer and the fractional part of $\frac{\tau}{h_t}$. In this case, we have

$$t_{j-d-1} \leq t_j - \tau \leq t_{j-d}$$

and

$$h_t \delta_t = \tau - d h_t.$$

The needed value of the solution $u_{t_j-\tau}$ in (13) is then approximated by:

$$u_{t_j-\tau} \approx \begin{cases} u(t_{j-d}), & \text{if } \delta_t < 0.5, \\ u(t_{j-d-1}), & \text{if } \delta_t \geq 0.5. \end{cases} \quad (15)$$

Concerning the choice of the basis functions, we consider a set of orthogonal functions, similar to the one described in [12], page 7. More precisely, we define

$$v_k(x) = \exp(ikx), \quad k = 0, 1, \dots, N. \quad (16)$$

Note that with this choice of the basis functions the inner product in (14) and the sums in (12) can be interpreted as the discrete Fourier transform (DFT). In particular, the set of inner products $(I(x, t_j), v_i)$, $i = 1, \dots, N$, may be seen as the DFT of the vector I_N , which contains the values of the function $I(x, t)$ at the grid points $x_k = -l + kh$, $k = 1, \dots, N$. In this case, t is fixed ($t = t_j$). Therefore, these inner products can be evaluated efficiently by the fast Fourier transform (FFT).

3 Numerical Examples

We have applied our algorithm to analyse the effect of noise on the formation of multi-bump solutions in dynamic neural fields, in the presence of space-dependent external stimuli. In [13], the authors have investigated the formation of regions of high activity (bumps) in neural fields, which can be switched ‘on’ and ‘off’ by transient stimuli. The stability analysis of such patterns in the deterministic case was carried out in [5] and in [6]. The effects of noise on such stationary pulse solutions were studied in [11]. In this case, the firing rate function $S(x)$ is the Heaviside function; the connectivity kernel is given by:

$$K(x) = 2 \exp(-0.08x) (0.08 \sin(\pi x/10) + \cos(\pi x/10)).$$

In [6], it was proved that such oscillatory connectivity kernels support the formation of stationary stable multi-bump solutions, induced by external inputs. In this case, the external input has the form:

$$I(x) = -3.39967 + 8 \exp\left(-\frac{x^2}{18}\right).$$

This specific example was considered in [5], p. 37. The parameters of the numerical approximation are $N = 100, h = 1, l = 50; n = 200, h_t = 0.02$. We start by considering the deterministic case. Using our code with $\epsilon = 0$, we are able to reproduce three kinds of stationary solutions, which are displayed in Figures 1 and 2.

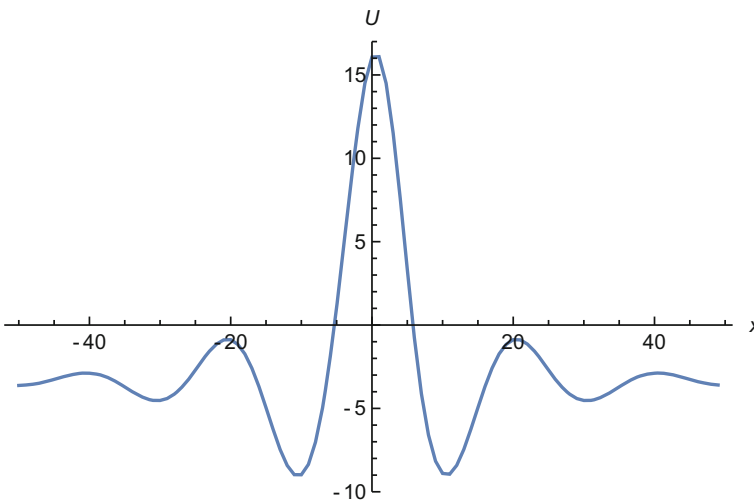


Fig. 1 Deterministic case: stationary one-bump solution

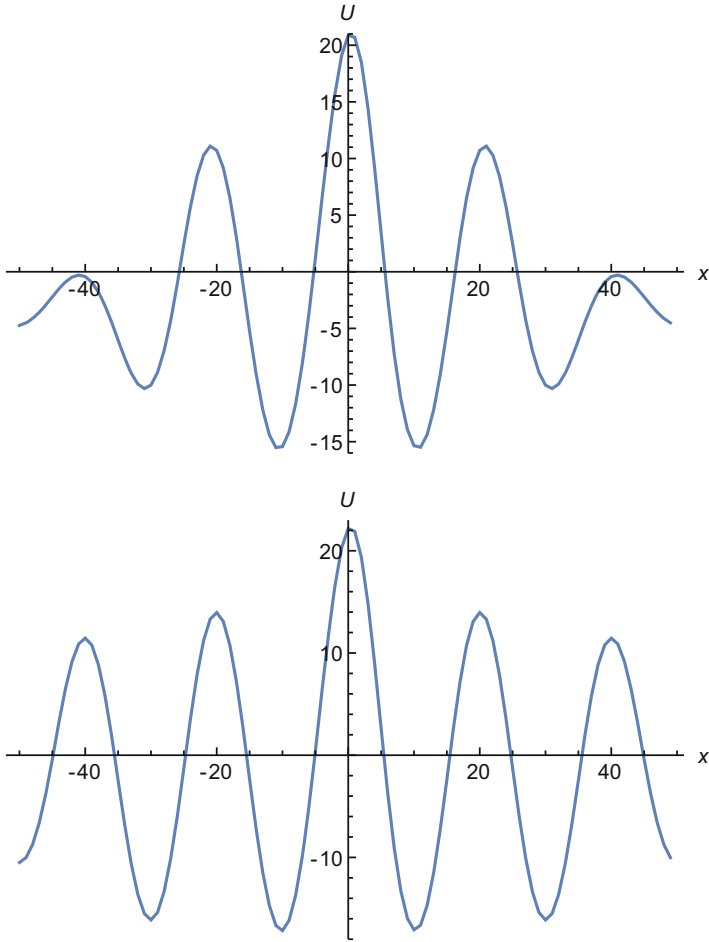


Fig. 2 Deterministic case: three-bump (up) and five-bump (down) solutions

As the **first experiment** with the **stochastic case**, we have performed a simulation with 100 paths, with level of noise $\epsilon = 0.01$ starting with the initial condition $U_0(x, t) \equiv 0$, over the time interval $t \in [0, 4]$. Our aim was to investigate the evolution of the paths of the stochastic equation and their relation with the stationary multi-bump solutions observed in the deterministic case. We remark that each solution $V(x, t)$ of the deterministic equation (1), as t tends to infinity, tends to a certain stationary solution $u(x)$, which is characterised by a certain maximal value $u_{max} = \max_{x \in [-l, l]} u(x)$ (which is usually located at the origin) and a minimal value $u_{min} = \min_{x \in [-l, l]} u(x)$ (which is attained at two symmetric points). Usually, a solution may have other local minima and maxima, whose absolute values do not exceed those of u_{min} and u_{max} . The greater is the number of bumps, the higher is the value of u_{max} and the lower is the value of u_{min} . These properties can be observed from the data displayed in Table 1.

Table 1 Properties of stationary solutions of the deterministic equation

	u_{min}	u_{max}
One-bump solution	-8.97	16.10
Three-bump solution	-15.52	20.88
Five-bump solution	-17.16	22.18

Note that u_{max} and u_{min} are, respectively, the limits, as $t \rightarrow \infty$, of $u_{max}(t)$ and $u_{min}(t)$, that is, $u_{max}(t) = \max_{x \in [-l, l]} V(x, t)$ and $u_{min}(t) = \min_{x \in [-l, l]} V(x, t)$. We have used these properties to analyse the paths of the stochastic equation. As we will see below, in certain cases, when these paths are rated according to the values of u_{min} and u_{max} , they can be clearly divided into classes, each class having typical values of these parameters, close to the ones of a certain stationary solution of the deterministic equation.

Moreover, when analysing the paths of the stochastic equation, we are often interested in evaluating their average and dispersion, and how these characteristics change with time. The maxima and minima of the paths give us a convenient way to analyse this evolution. Let $u(s, x, t)$ denote the approximate value of $U_t(x)$, given by the s -th path. Assuming that n_p is the number of paths, use the following notations:

$$U_{max,max}(t) = \max_{s \in \{1, \dots, n_p\}} \max_{i \in \{1, \dots, N\}} u(s, x_i, t);$$

$$U_{min,max}(t) = \min_{s \in \{1, \dots, n_p\}} \max_{i \in \{1, \dots, N\}} u(s, x_i, t);$$

$$U_{max,min}(t) = \max_{s \in \{1, \dots, n_p\}} \min_{i \in \{1, \dots, N\}} u(s, x_i, t);$$

$$U_{min,min}(t) = \min_{s \in \{1, \dots, n_p\}} \min_{i \in \{1, \dots, N\}} u(s, x_i, t);$$

We consider the following approximations of mathematical expectations:

$$E(u(x, t)) \approx \frac{1}{n_p} \sum_{s=1}^{n_p} u(s, x, t);$$

$$E\left(\max_{x \in [-l, l]} u(x, t)\right) \approx E_{max}(t) = \frac{1}{n_p} \sum_{s=1}^{n_p} \max_{i \in \{1, \dots, N\}} u(s, x_i, t);$$

$$E\left(\min_{x \in [-l, l]} u(x, t)\right) \approx E_{min}(t) = \frac{1}{n_p} \sum_{s=1}^{n_p} \min_{i \in \{1, \dots, N\}} u(s, x_i, t);$$

In Figure 3, we can see the time evolution of the solution maximum (up) and the time evolution of the solution minimum (down). In the first case, the graphs of $U_{max,max}(t)$, $U_{min,max}(t)$, and $E_{max}(t)$ are displayed; the second graphic shows

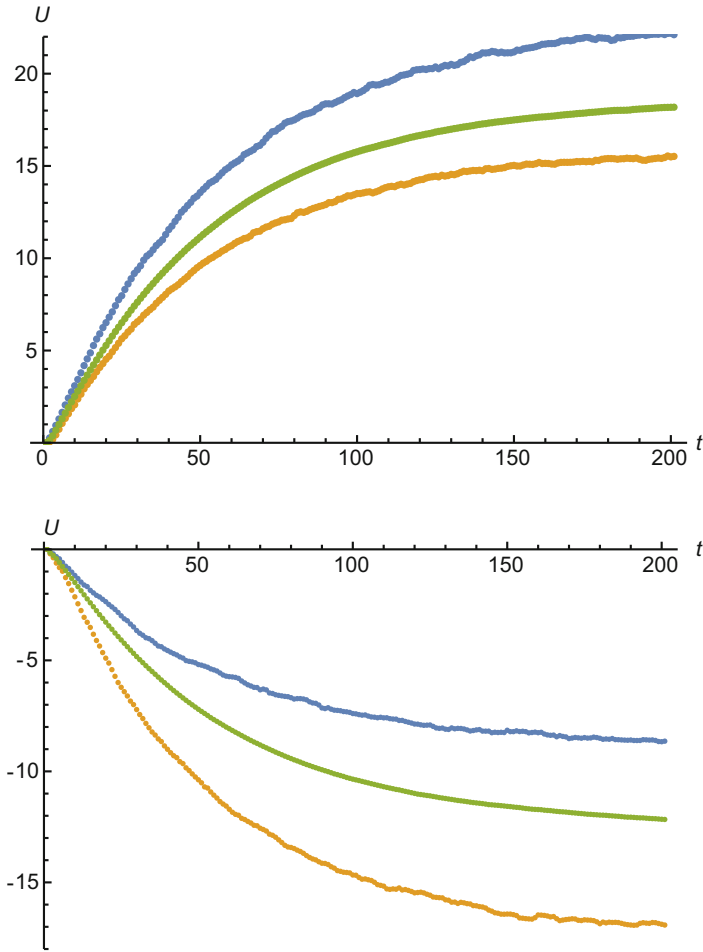


Fig. 3 Starting from the zero solution. Up: evolution of solution maximum— $U_{max,max}$ (blue), $U_{min,max}$ (yellow), and E_{max} (green). Down: evolution of solution minimum— $U_{max,min}$ (blue), $U_{min,min}$ (yellow), and E_{min} (green).

values of $U_{max,min}(t)$, $U_{min,min}(t)$, and $E_{min}(t)$. These figures show that the average value of the maximum increases with time, while the average value of the minimum decreases; as a result, the average amplitude of the oscillations (height of the bumps) increases with time. Moreover, as it could be expected the dispersion (maximal difference between values of different paths) also increases with time. As it happens with $u_{min}(t)$ and $u_{max}(t)$ in the deterministic case, the values $E_{min}(t)$ and $E_{max}(t)$ stabilise as t increases. For $j = 200(t_j = 4)$, their values are close to the ones of u_{min} and u_{max} , in the case of a deterministic one-bump solution.

Figure 4 shows the distribution of u_{max} (up) and u_{min} (down), at $t = 4$, for the different paths. We see that the maxima are concentrated on the range $[15.8, 16.6]$, which corresponds to the stationary one-bump solution, and $[20, 21.2]$, which

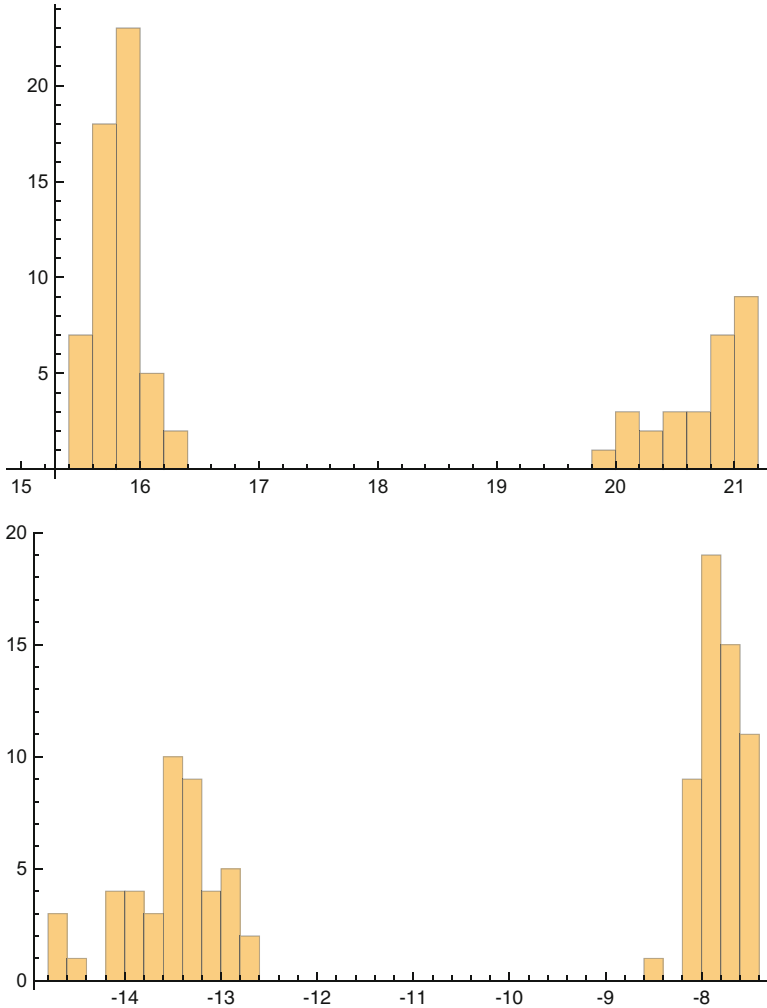


Fig. 4 Histograms of distribution of u_{max} (up) and u_{min} (down), at $t_j = 4$ ($j = 200$), in the case $U_0(x, t) \equiv 0$, with $\epsilon = 0.01$.

corresponds approximately to three-bump and five-bump solutions (see Table 1). The minima are concentrated on the interval $[-8.2, -7.4]$, which corresponds to the stationary one-bump solution, and $[-14., -12.5]$, which corresponds approximately to three-bump solutions of the deterministic equation. This suggests that under the effect of not very strong noise, the most probable values of the solution of the stochastic equation are close to the ones of some of the stationary solutions of the deterministic equation. By other words, if the level noise is not too high, there is a high probability that with time each path of the stochastic equation will approach

one of the known stationary solutions of the deterministic equation. Moreover, the probability that it will approach a one-bump solution is significantly higher than the probability of tending to a third-bump or five-bump one.

As a **second numerical experiment**, we have performed a simulation with 100 paths, with level of noise $\epsilon = 0.01$, over the time interval $t \in [0, 4]$, taking as initial condition $U_0(x, t)$ a stationary one-bump solution (of the type represented in Figure 1).

In Figure 5, we again observe the evolution of $U_{max,max}(t)$, $U_{min,max}(t)$, and $E_{max}(t)$ (up); the values of $U_{max,min}(t)$, $U_{min,min}(t)$, and $E_{min}(t)$ (down). As in the previous case (Figure 3), we observe that the average amplitude of the solutions'

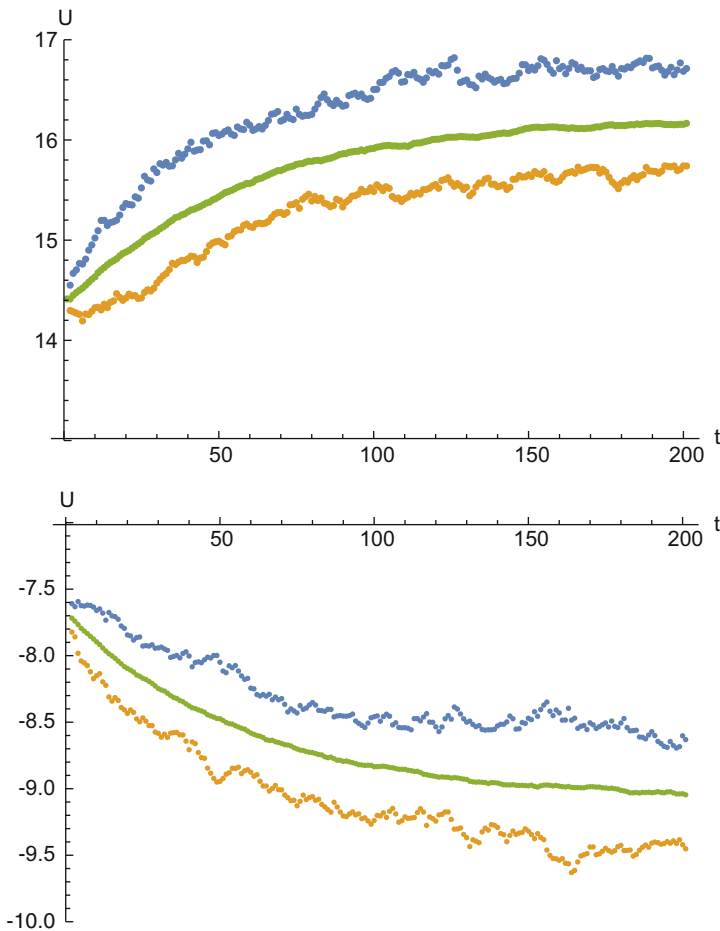


Fig. 5 Starting from a one-bump solution, with $\epsilon = 0.01$. Up: evolution of solution maximum— $U_{max,max}$ (blue), $U_{min,max}$ (yellow), and E_{max} (green). Down: evolution of solution minimum— $U_{max,min}$ (blue), $U_{min,min}$ (yellow), and E_{min} (green).

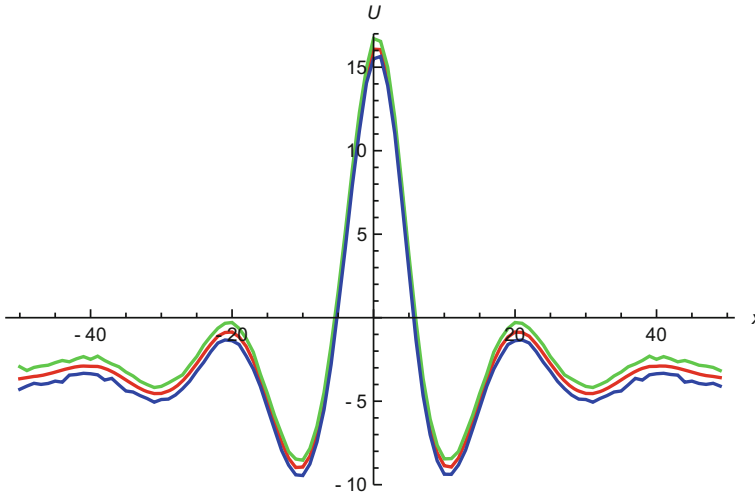


Fig. 6 Graph of $E(u(x, 4))$ (red line), $\max_{s \in \{1, \dots, 100\}} u(s, x, 4)$ (green line), and $\min_{s \in \{1, \dots, 100\}} u(s, x, 4)$ (blue line), in the case $\epsilon = 0.01$.

oscillations increases with time; however for this level of noise ($\epsilon = 0.01$), the amplitude of oscillations tends to stabilise and the mean value of the stochastic solution remains close to the one-bump deterministic solution.

In Figure 6, the graph of the average solution $E(u(x, t))$ at $t = 4$ is displayed, between the graphs containing the minimal and maximal values (of all the paths); these graphs correspond to the simulation starting with the stationary one-bump solution and with noise level $\epsilon = 0.01$.

In Figure 7, we can see the distribution (at $t = 4$) of the values of u_{max} (up) and u_{min} (down), for the same initial values and noise level. We see that the maxima are concentrated on the range $[15.8, 16.6]$ and the minima are concentrated on the range $[-9.4, -8.3]$. This means that with level noise $\epsilon = 0.01$ when the initial value of the simulation is a deterministic one-bump solution, the most probable values of the stochastic solution are concentrated near the ones of this stationary solution.

As a **third numerical experiment**, we have performed a simulation with 100 paths, with noise level $\epsilon = 0.05$, over the time interval $t \in [0, 4]$, taking again as initial condition $U_0(x, t)$ a stationary one-bump solution.

The evolution of the maximum and minimum of the solutions is displayed in Figure 8; in each case, we can see the graphs of the average value of the 100 paths ($E_{max}(t)$ and $E_{min}(t)$, respectively). In Figure 9, the graph of the average solution at $t = 4$ is plotted, between the graphs containing the minimal and maximal values (of all the paths). As in the previous cases, we observe that the average amplitude of the solutions' oscillations increases with time; moreover, for this level of noise ($\epsilon = 0.05$), the amplitude of oscillations increases so much that some paths behave like three-bump solutions or five-bump solutions.

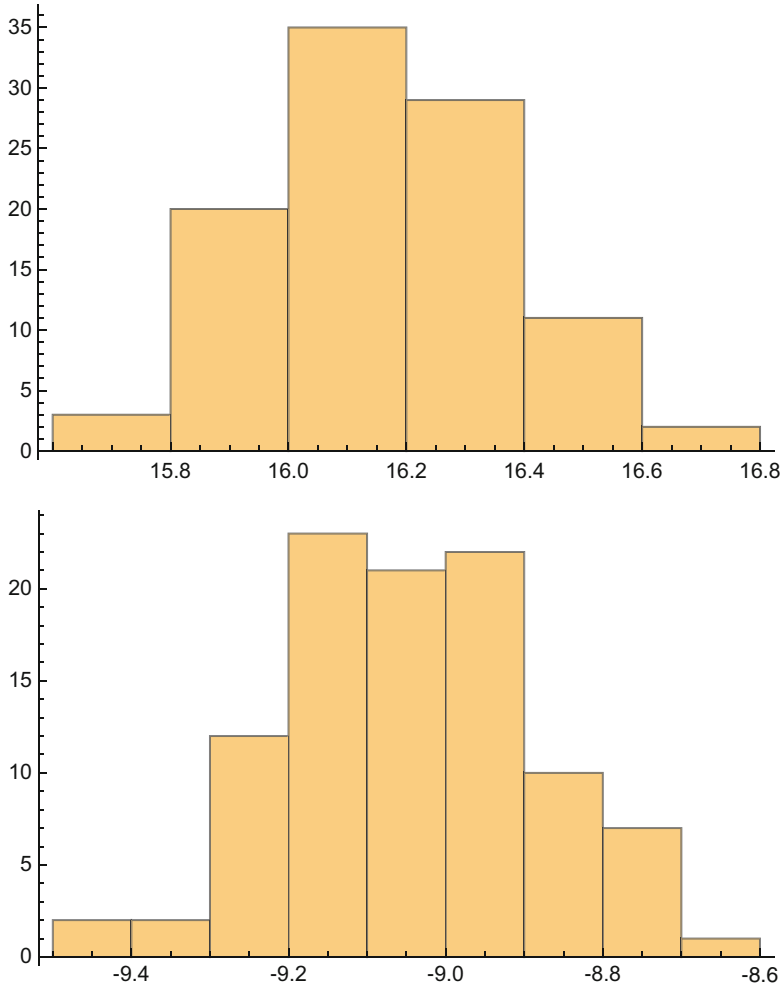


Fig. 7 Histograms of distribution of u_{max} (up) and u_{min} (down), at $t = 4$ (for the case where $U_0(x, t)$ is a deterministic one-bump solution, with $\epsilon = 0.01$).

In Figure 10, the distribution (at $t = 4$) of the maxima and minima (respectively) of the different paths is displayed. We see that most of the maxima are on the range $[16, 25]$ and most of the minima are on the range $[-19, -9]$. This reflects the fact that with noise level $\epsilon = 0.05$ many paths starting from a stationary one-bump solution can after some time take values that are characteristic of three-bump or five-bump deterministic solutions. Moreover, the probability that the solution at $t = 4$ remains as a one-bump solution is much smaller than the probability of being transformed into a solution with a higher number of bumps.

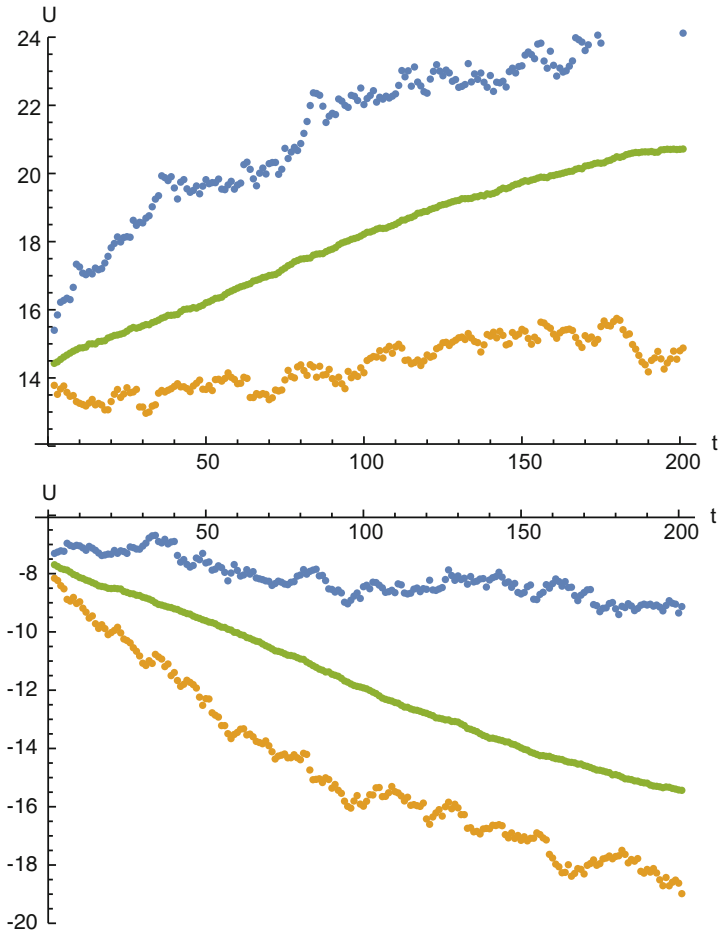


Fig. 8 Starting from a one-bump solution, with $\epsilon = 0.05$. Up: evolution of solution maximum— $U_{max,max}$ (blue), $U_{min,max}$ (yellow), E_{max} (green). Down: evolution of solution minimum: $U_{max,min}$ (blue), $U_{min,min}$ (yellow), E_{min} (green).

The numerical algorithm was implemented in Mathematica [19] and the computations were performed in a PC with a 1.7-Ghz processor and 8 Gb of installed memory (RAM). The computation of the numerical examples presented in this section (with 100 paths) takes about 1 hour.

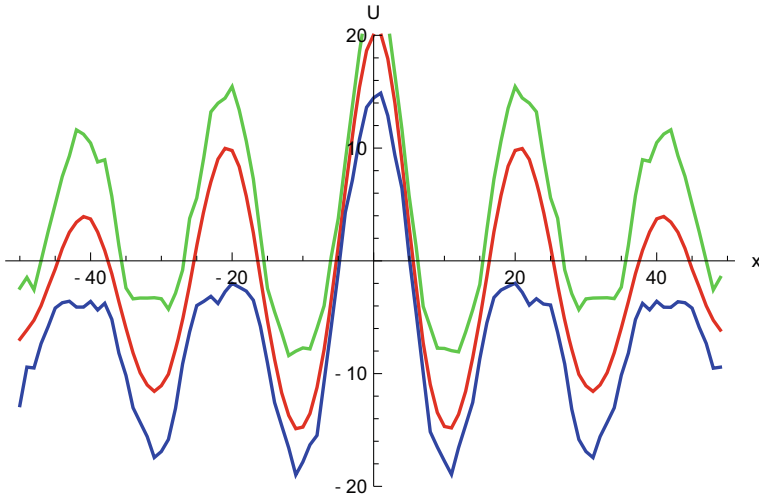


Fig. 9 Graph of $E(u(x, 4))$ (red line), $\max_{s \in \{1, \dots, 100\}} u(s, x, 4)$ (green line), and $\min_{s \in \{1, \dots, 100\}} u(s, x, 4)$ (blue line), in the case $\epsilon = 0.05$.

4 Conclusions

In this paper, we have introduced a new numerical algorithm for the approximation of the stochastic neural field equation with delay. This numerical algorithm uses the Galerkin method and is inspired in the Kühn and Riedler’s approach [12]. The choice of the basis functions and grid points allows the use of the fast Fourier transform to perform summations, which improves significantly the efficiency of the algorithm.

To test the algorithm, we have applied it to the numerical solution of a neural field, in the presence of external stimuli, where stationary one-bump and multi-bump solutions are known to exist in the deterministic case. The numerical results suggest that for a low level of noise the trajectories of the stochastic equation are concentrated near the stationary solutions of the deterministic one. In particular, if the initial condition is the null function, in the deterministic case the solution tends with time to a one-bump stationary solution. But in the presence of not very strong noise, the trajectories of the stochastic equation split into several classes, each of them close to a different stationary solution of the deterministic equation.

In conclusion, we can say that our results are consistent with the study of Kilpatrick and Ermentrout [11], where the authors conclude that upon breaking the translation symmetry of a neural field, by introducing spatially heterogeneous input or synapses, bumps in the stochastic neural field can become temporarily fixed to a finite number of locations.

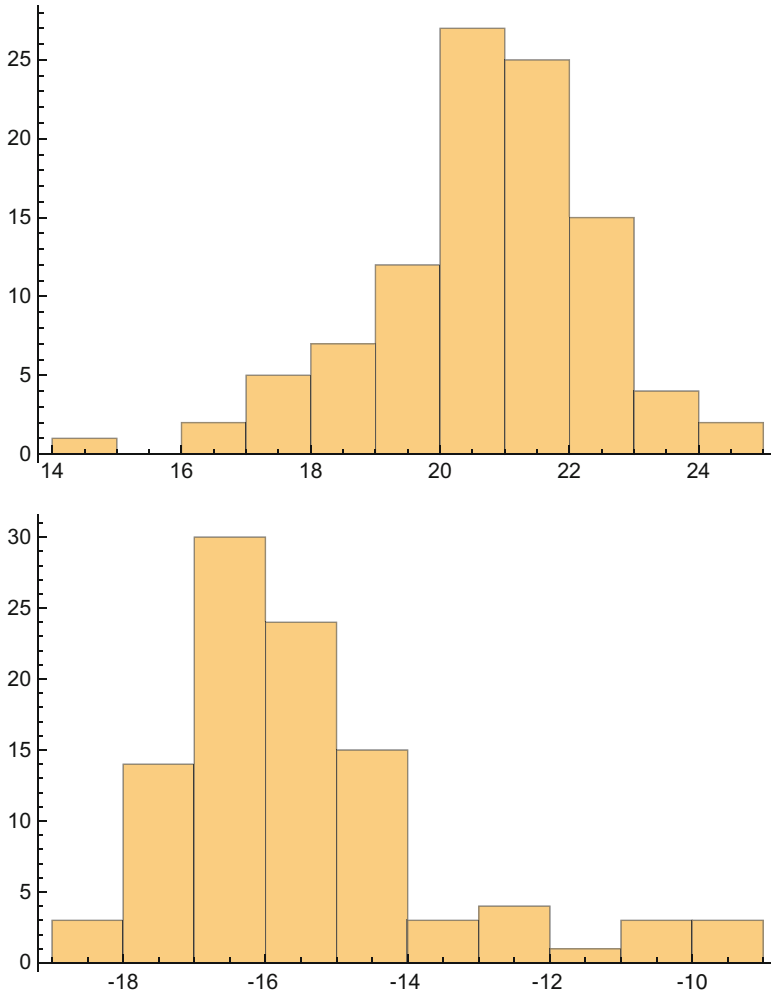


Fig. 10 Histograms of distribution of u_{max} (up) and u_{min} (down), at $t = 4$ (in the case where $U_0(x, t)$ is a deterministic one-bump solution, with $\epsilon = 0.05$).

Acknowledgements We acknowledge support from Fundação para a Ciência e a Tecnologia (the Portuguese Foundation for Science and Technology) through the grant SFRH/BSAB/135130/2017 and POCI-01-0145-FEDER-031393.

References

1. Amari, S.L.: Dynamics of pattern formation in lateral-inhibition type neural fields, *Biol. Cybernet.* **27** (2), 77–87 (1977).
2. Attay, F.M. and Hutt, A.: Neural fields with distributed transmission speeds and long-range feedback delays, *SIAM Journal of Applied Dynamical Systems* **5**(4), 670–698 (2006).

3. Folias, S.E., and Bressloff, P.C.: Breathers in Two-dimensional neural media, *Phys. Rev. Let.* **95** 208107 (2005).
4. Faye, G. and Faugeras, O.: Some theoretical and numerical results for delayed neural field equations, *Physica D* **239** 561–578 (2010).
5. Ferreira, F: Multi-bump solutions in dynamic neural fields: analysis and applications, PhD thesis, University of Minho, 2014. <http://hdl.handle.net/1822/34416>.
6. Ferreira, F., Erlhagen, W. and Bicho, E.: Multi-bump solutions in a neural field model with external inputs, *Physica D (Nonlinear Phenomena)*, **326** 32–51 (2016).
7. Van Gils, S.A., Janssens, S.G., Kuznetsov Y.A. and Visser, S.: On local bifurcations in neural field models with transmission delays, *Journal of Mathematical Biology* **66** (4–5), 837–887 (2013).
8. Hutt, A and Lefebvre, J. Stochastic center manifold analysis in scalar nonlinear systems involving distributed delays and additive noise, *Markov Processes and Related Fields*, **22**, 555–572 (2016).
9. Hutt, A. and Rougier, N.: Activity spread and breathers induced by finite transmission speeds in two-dimensional neuronal fields, *Physical Review E* **82** 055701 (2010).
10. Hutt, A. and Rougier, N.: Numerical simulations of one- and two-dimensional neural fields involving space-dependent delays. In: S. Coombes et al. (Eds.), *Neural Fields Theory and Applications*, 175–183, Springer (2014).
11. Kilpatrick, Z. and Ermentrout, B.: Wandering bumps in stochastic neural fields, *SIAM J. Appl. Dyn. Syst.*, **12** 61–94 (2013).
12. C. Kühn and M.G. Riedler, Large deviations for nonlocal stochastic neural fields, *J. Math. Neurosci.* **4** (1) 1–33 (2014).
13. Laing, C.R., Troy, W.C., Gutkin, B. and Ermentrout, G.B.: Multiple bumps in a neuronal model of working memory, *SIAM J. Appl. Math.* **63** 62–97 (2002).
14. Lima, P.M. and Buckwar, E.: Numerical solution of the neural field equation in the two-dimensional case, *SIAM Journal of Scientific Computing*, **37**, B962–B979 (2015).
15. Lima, P.M. and Buckwar, E.: Numerical investigation of the two-dimensional neural field equation with delay. In: *Proceedings of the Second International Conference on Mathematics and Computers in Sciences and in Industry*, 131–137, IEEE Conference Publications (2015).
16. Veltz, R. and Faugeras, O.: Stability of stationary solutions of neural field equations with propagation delays, *J. Math. Neurosci.* **1**:1,1–28 (2011).
17. Veltz, R. and Faugeras, O.: Local/Global analysis of the stationary solutions of some neural field equations, *SIAM J. Appl. Dyn. Syst.* **9** 954–998 (2010).
18. Wilson, H.R. and Cowan, J.D.: Excitatory and inhibitory interactions in localized populations of model neurons, *Bipophys. J.*, **12** 1–24 (1972).
19. Wolfram, S.: *The Mathematica Book*. Wolfram Media (2003).

Nonstationary Signal Decomposition for Dummies



Antonio Cicone

1 Introduction

This work is an introductory survey on nonstationary signals and some of the most advanced techniques available in the literature for their decomposition. It is intended for an audience of people working with nonstationary signals who have never tried anything more sophisticated than Fourier or wavelet transform, and for everyone who is simply curious about the subject.

The idea is, starting from simple and basic concepts, to draw a brief and self-contained picture of the subject and then to show how to use two modern algorithms for signal decomposition.

Nonstationary signals are ubiquitous in real life. We can consider, for instance, a stock market index, the ECG of a pregnant woman, or the terrestrial magnetic field measured by a magnetometer. For all these signals there are two kind of problems we may want to address:

- Q1** What are the active frequencies at each instant of time?
- Q2** How to decompose such signals into simpler components?

We will see how these two questions, which may sound apparently unrelated, are indeed two sides of the same medal: understanding the nonstationary behavior of the signal.

A. Cicone (✉)

Istituto Nazionale di Alta Matematica, Città Universitaria, P.le Aldo Moro 5, 00185 Rome, Italy

Department of Information Engineering, Computer Science and Mathematics, Università degli Studi dell'Aquila, via Vetoio n.1, 67100 L'Aquila, Italy

Gran Sasso Science Institute, Via Michele Jacobucci, 2, 67100 L'Aquila, Italy

e-mail: antonio.cicone@univaq.it

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,

https://doi.org/10.1007/978-3-030-02487-1_3

Before addressing the previous two questions it is important to review what are the *instantaneous frequency* and the *time–frequency representation* of a signal.

The idea of instantaneous frequency is natural. All of us know the basic sine function $y = A \sin(2\pi\phi t)$ where A is the amplitude, $1/\phi$ the period, and ϕ the frequency. What if these quantities vary over time? We end up having what is called a amplitude modulated and frequency modulated (AM FM) signal $z = A(t) \sin(2\pi)\phi(t)t$. The intuition suggests to consider $\phi(t)$ as the instantaneous frequency of z at the instant of time t . The question now is how to compute such quantity when a closed form formula of the signal is not known.

Ideally we would like to generalize the definition of stationary frequency as reciprocal of the period to the nonstationary case. However, it becomes immediately clear that such generalization is, in most cases, not feasible. In Figure 1 it is shown an example of a nonstationary AM FM signal with a frequency which increases over time. How to compute rigorously its instantaneous frequency? We cannot simply rely on the periods which cannot be computed exactly at each instant of time. One possible approach, well known and broadly adopted, is based on the evaluation of the Hilbert transform of the signal [20, equation (3.4)]. Another way, recently proposed in [6, equation (34)], is based instead on the computation of the signal derivative. Both approaches have their own advantages as well as limitations. On the one hand, the one based on the Hilbert transform, since it relies on integration, is inherently more stable, but at the same time the integration implies that not only local information are used for the instantaneous frequency computation. On the

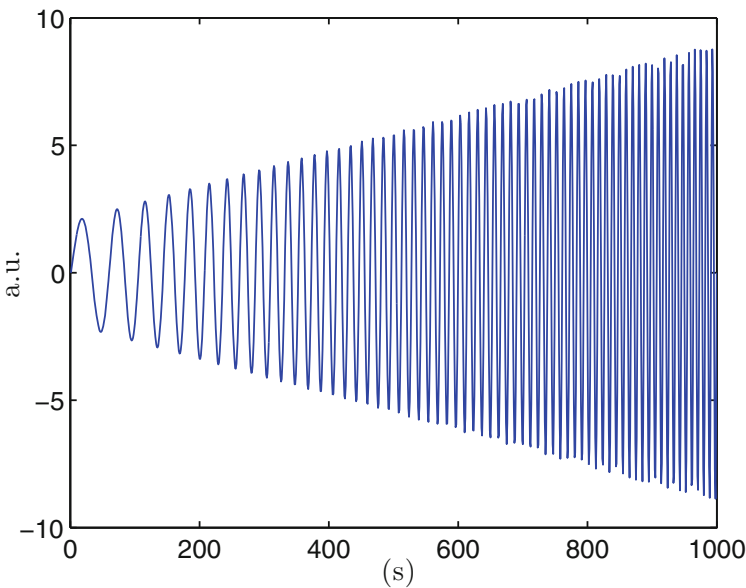


Fig. 1 Example of a nonstationary signal with instantaneous frequency which increases over time

other hand, the method based on derivation is less stable, but at the same time it is completely local since only instantaneous information is used in its computation.

Once such information is made available we can represent it in a plot where the horizontal axis corresponds to the time and the vertical one to the frequencies. This is what we call a time–frequency representation.

The aforementioned methods allow to compute the instantaneous frequency of a signal whose instantaneous frequency is unique at each instant of time. What about signals which contain two or more instantaneous frequencies? In this case these two methods will not work.

This is where the questions Q1 and Q2 arise. In fact we can address the problem of studying a nonstationary signal in two ways. On the one hand, we can try to compute directly its time–frequency representation, addressing Q1. On the other hand, we can tackle Q2 by first decomposing the signal into simple components each of which contains, ideally, a unique instantaneous frequency at each instant of time and then applying the previously mentioned techniques to compute the instantaneous frequencies component by component.

Historically researchers have first tackled Q1. The first technique used was the standard Fourier transform [2] which is, however, inherently unable to capture any nonstationarity in the signal. For this reason the so-called short time Fourier transform was developed [9]. This last method allowed to produce a meaningful time–frequency representations of nonstationary signals. The continuous wavelet transform [10] allowed to improve the accuracy of the time–frequency representation, in fact, instead of relying on an orthonormal basis of sines and cosines functions, it uses dilation and translations of a mother wavelet. To further sharpen the time–frequency representation the so-called synchrosqueezing transform was proposed in [11, 12], which proved to be a special kind of a broader class of techniques named reassignment methods [1, 13].

All the aforementioned algorithms are linear: they treat signals as linear combinations of elements in a preselected basis, sinusoidal or wavelet. There are also quadratic methods which are based instead on energy and power distributions, like the spectrogram, the scalogram, and the Wigner–Ville distribution. We refer the interested reader to the book [13].

It is important to remind at this point the so-called Heisenberg uncertainty principle [9]. Based on this principle the accuracy that it can be achieved in producing a time–frequency representation is limited. In particular the idea is that either we achieve a good accuracy in assigning the frequency, but then the corresponding time horizon is not well identified or, vice versa, we have a good accuracy in identifying the time horizon of a nonstationary event in a signal, but we lose accuracy in assigning the frequency [14].

In 1998, meanwhile new methods for a more accurate than ever time–frequency representations were proposed, Huang and his research group at NASA devised the so-called empirical mode decomposition (EMD) algorithm [20]. This method is the first technique ever developed able to address Q2 without any kind of a priori information and without making any a priori selection of a basis to be used. Furthermore, in some sense, it allows to bypass the Heisenberg uncertainty

principle. In fact, decomposing directly a signal without any a priori information on its instantaneous frequencies allows to address Q2 avoiding the limitations induced by the Heisenberg uncertainty principle.

Few years later it became clear that this method is unstable. In particular its sensitivity to noise was unraveled. For this reason Huang and his group devised the so-called ensemble empirical mode decomposition (EEMD) algorithm [26] which allows to overcome the sensitivity of the original EMD algorithm.

The publication of the EMD first and the EEMD after and their success inspired many other researchers to work on alternative methods for the decomposition of a signal into a few simple and meaningful components. All the alternative methods proposed so far are based on the minimization of some functional, like the sparse time–frequency representation algorithm [17, 18], and they require to make some assumption on the signal under study. Only one of them, called iterative filtering (IF) method [22], and its generalization, the adaptive local iterative filtering (ALIF) algorithm [6], are based on iterations like EMD and EEMD, and therefore, no assumptions are required on the kind of signal we want to decompose.

The rest of this paper focuses on decomposition methods based on iterations for 1D signals. In particular we review the EEMD method, Section 2, and the IF algorithm, Section 3. The paper ends with concluding remarks and an outlook to the main open problems in the field.

We point out here that both EEMD and IF have been generalized to 2D. See [27] and [4, 7] for further details.

2 The Ensemble Empirical Mode Decomposition Algorithm

As we mentioned in the introduction, the goal of the EMD and EEMD methods is the decomposition of a signal into simple components, called intrinsic mode functions (IMFs), each of which with a unique instantaneous frequency at every instant of time.

We start reviewing the EMD algorithm [20], whose pseudocode is given in Algorithm 1.

Algorithm 1 Empirical Mode Decomposition IMF = EMD(f)

```

IMF = {}
while the number of extrema of  $f \geq 2$  do
   $s_1 = s$ 
  while the stopping criterion is not satisfied do
    compute the moving average  $M(s_m(x))$ 
     $s_{m+1}(x) = s_m(x) - M(s_m(x))$ 
     $m = m + 1$ 
  end while
  IMF = IMF  $\cup$   $\{s_m\}$ 
   $s = s - s_m$ 
end while
IMF = IMF  $\cup$   $\{s\}$ 

```

The key idea behind this method is what the authors called the sifting process: given a signal s we capture its highest frequency oscillations by subtracting its moving average $M(s)$ from the signal itself. To do so we need to compute somehow the function $M(s)$. Huang and his research group proposed to compute first the upper and lower envelopes connecting the maxima and minima of the signal, respectively, by means of cubic splines. Then the moving average is computed as mean between these two curves point by point.

Since the derived moving average is only an approximation of the exact one, the idea is to iterate the aforementioned calculation applying it to the new signal generated after the subtraction. Assuming $s_1 = s$ we compute

$$s_{m+1} = s_m - M(s_m) \quad \forall m \geq 1 \quad (1)$$

In the end we expect the method to converge to an IMF or, using a stopping criterion, we discontinue the calculations at a certain \tilde{m} when we are close in some sense to an IMF.

Then, using the approach just described, we compute $\text{IMF}_1 = s_{\tilde{m}}$ and we subtract it from the signal under study. The remainder $r = s - \text{IMF}_1$ can be treated as a new signal to which we apply again the sifting process. In the end we decompose the original signal into several IMFs and a remainder that cannot be decomposed anymore because it does not contain any oscillations.

Everything works fine except that the EMD process just described proved to be sensitive to small perturbations. In particular if we perturb a given signal with white noise, even if small amplitude compared with the signal itself, the EMD may end up providing a completely different decomposition.

To address this issue Huang and his collaborators proposed [26] to add to the given signal hundreds of different white noise realizations and to decompose separately each outcome of this addition. The final decomposition is then computed as the average of all these decompositions.

The EEMD Matlab implementation can be download from <http://rcada.ncu.edu.tw/>.

Besides the signal there are other two inputs that we need to pass to the EEMD Matlab function:

- Nstd which represents the ratio between the standard deviation of the added noise and that of the signal under study.
- NE Number of noise realizations in the ensemble, id est (i.e.) the number of noise realizations to be added to the given signal.

The authors of the EEMD suggest in [26] to set Nstd to 0.2. They also point out that if the data is dominated by high-frequency signals, the noise amplitude may be smaller, and when the data is dominated by low-frequency signals, the noise amplitude may be increased.

Furthermore they suggest to use ensembles of a few hundreds perturbations of a single signal. Based on our experience setting NE to one hundred is enough in many cases.

2.1 Numerical Examples

Following what has been done in [19] we apply the EEMD algorithm, available at <http://rcada.ncu.edu.tw/>, to two well-known geophysical nonstationary signals: the Vostok temperature derived from ice core signal [23, 25] and the length-of-day data (LOD) [15].

We start with the Vostok temperature dataset. We focus, for simplicity, on the last 50 thousand years sample values, left panel of Figure 2. Applying the EEMD method with $N_{std} = 0.2$ and $NE = 100$ we obtain the decomposition shown in the middle and right panels of Figure 2.

For the LOD dataset, instead, we consider the data from the beginning of 1983 to the end of 1986. The signal is shown in the left panel of Figure 3.

If we apply the EEMD method again with N_{std} set to 0.2, and NE set to 100, we obtain the decomposition shown in the middle and right panels of Figure 3.

In the following we describe the alternative method, iterative filtering, and we apply it to the very same datasets. We provide in there a detailed description of the derived components physical meaning.

For further details on these two datasets and the meaning of their decompositions we refer the interested reader to [19].

3 The Iterative Filtering Method

Inspired by the EMD algorithm, Lin et al proposed in 2009 an alternative method called iterative filtering [22]. The structure is the very same as in EMD with the only difference that now the moving average $M(s_m(x))$ is computed as a local average of the values of the signal. This is achieved by integration of the signal itself weighted using an a priori chosen mask w_m with nonzero values concentrated on a finite interval $[-l_m, l_m]$

$$M(s_m(x)) = \int_{-l_m}^{l_m} s_m(x+t)w_m(t)dt \quad (2)$$

where the subscript m stands for the step number in the iteration.

How to choose the mask function w_m ? Based on the theoretical results [4–6, 8, 24], to guarantee the convergence and stability of the algorithm it is sufficient to consider a mask function which is generated as follows. We convolve with itself a function fulfilling the following properties: compactly supported (it is zero outside a closed and finite set), nonnegative valued, with integral equal to 1, and even (symmetric with respect to the vertical axis). This in turn guarantees also the physical meaningfulness of the decomposition. In [6] Fokker–Plank filters were proposed. They depend on two parameters, α and β , which can be tuned to produce infinitely many filters. They have the extra property of being infinitely smooth on

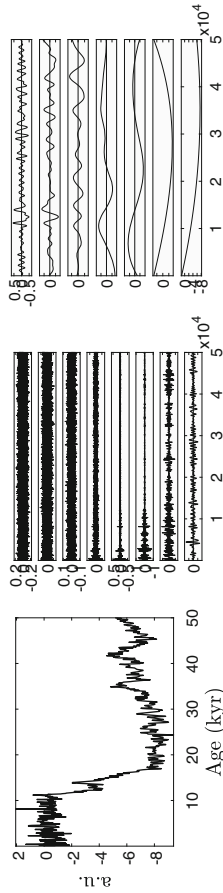


Fig. 2 Left panel, Vostok temperature dataset of the last 50 thousand years. Middle and right panel EEMD decomposition

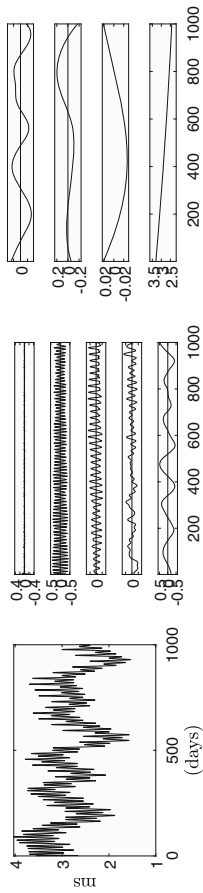


Fig. 3 Left panel, length-of-day data from the beginning of 1983 to the end of 1986. Middle and right panel, EEMD decomposition

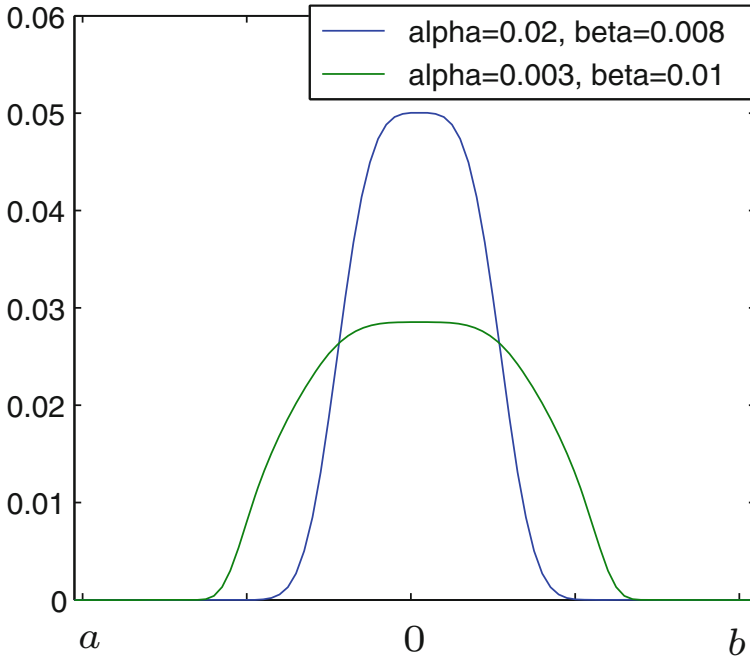


Fig. 4 The Fokker–Planck filters associated with $\alpha = 0.02$, $\beta = 0.008$ and $\alpha = 0.003$, $\beta = 0.01$, respectively

the entire real line. In Figure 4 two examples of such filters are shown. The only drawback of Fokker–Planck filters is that they are not known in an explicit form. However, it is possible to compute them up to machine precision using numerical methods. In the following we use the filter associated with $\alpha = 0.005$ and $\beta = 0.09$. This specific Fokker–Planck filter is available, together with a Matlab version of iterative filtering, online.¹

The pseudocode of the IF algorithm is given in Algorithm 2 where $w_m(t)$ is the chosen mask function whose support is in $[-l_m, l_m]$, and l_m is called *mask length* which represents the half support length.

The current Matlab release of IF is `IF_v6` and it requires as input the signal we want to decompose and, optionally, a variable, generated using an extra function called `Settings_IF`, which contains all the tuning parameters needed in the main algorithm. This implementation allows also the user to pass a third input, a vector containing a priori determined mask length values, forcing the method to skip their computation.

As outputs the algorithm returns a matrix and a vector. The matrix contains as rows the IMFs, that we recall are the simple components in which the signal is

¹www.cicone.com, GitHub and Mathworks.

Algorithm 2 Iterative Filtering $\text{IMF} = \text{IF}(s)$

```

IMF = {}
while the number of extrema of  $s \geq N \geq 2$  do
   $s_1 = s$ 
  while the stopping criterion is not satisfied do
    compute the filter length  $l_m$  for  $s_m(x)$ 
     $s_{m+1}(x) = s_m(x) - \int_{-l_m}^{l_m} s_m(x+t)w_m(t)dt$ 
     $m = m + 1$ 
  end while
   $\text{IMF} = \text{IMF} \cup \{s_m\}$ 
   $s = s - s_m$ 
end while
 $\text{IMF} = \text{IMF} \cup \{s\}$ 

```

decomposed. The vector contains the values of mask length used to decompose each IMF.

What are the tuning parameters that can be set in IF by means of the Matlab function `Settings_IF`? Let us review them one by one.

delta	In the IF algorithm we need to use a stopping criterion to discontinue the calculation for each IMF, as pointed out in line four of the pseudocode. The one currently implemented, called <code>delta</code> , corresponds to the ratio between the norm 2 of the moving average curve and the norm 2 of the signal. Since the moving average curve converges towards a zero function we discontinue the calculations when the aforementioned ratio is small enough. Default value is set to 0.001.
ExtPoints	The algorithm iterates until the remainder has at most <code>ExtPoints</code> number of extrema. This number corresponds to the value N in line two of the pseudocode. Default value is equal to 3.
NIMFs	Maximal number of IMFs allowed in the decomposition, excluding the remainder. Default value is set to 1.
extensionType	The IF algorithm requires, as any other method in signal processing, to make an assumption on how the signal extends outside the boundaries. Three options are given: constant, periodical, and reflection. In the constant case the signal extends outside the boundaries with the last values achieved at the boundaries. Whereas periodical implies that the signal is assumed to repeat infinitely many times outside the boundaries. Finally reflection means that we extend the signal assuming symmetry with respect to the vertical lines passing through the boundary points.
MaxInner	Maximum number of iterations allowed for the computation of each IMF. Default value is set to 200.
alpha	Parameter used for the mask length computation. In particular the algorithm measures the distances between subsequent extrema in

the signal under study. It is up to the user to decide which value to be used in the decomposition. In fact, if `alpha` is set to 0, the mask length is proportional to the minimum distance between two subsequent extrema. If it is set to 1, then it is proportional to the maximum distance. If we set it to `ave`, the mask length equals to, as suggested in [22], the roundoff value of the ratio $\frac{2 * \bar{X}_i * L}{N_e}$, where L is the length of the signal, \bar{X}_i is defined in the following, and N_e represents the number of extrema in the signal. Finally if we set it to `Almost_min`, the mask length is selected equal to the roundoff value of $2 * \bar{X}_i * P_{30}$, where P_{30} represents the 30 percentile of the vector containing all the distances between subsequent extrema of the signal. Default value set to `ave`.

We point out that the smaller is the mask length the higher is the number of IMFs produced in the decomposition and the finer is the separation of two frequencies. Based on our experience the option `Almost_min` allows to obtain a decomposition with a high number of components, but with IMFs which have unique instantaneous frequencies all the time. Higher values in the parameter `alpha` may lead, in some cases, to components which show an intermittency in the frequency.

`Xi` This last parameter allows to tune the mask length. Depending on the chosen filter, in fact, we may need to extend the filter length more or less due to its shape. In particular if we select a mask shape with big values in the center, but which goes quickly to zero we may need to select `Xi` bigger than 2. Whereas if the mask shape is almost constant everywhere on its support, then `Xi` may be selected closer to 1. Based on the numerical analysis conducted in [5] we know that enlarging the mask function support implies squeezing its Fourier transform and in turn it allows to tune the sampled frequencies that enter in the extracted IMF. Suggested values range in the interval [1.1, 3]. Default value is equal to 1.6.

3.1 Numerical Examples

We apply the `IF_v6` to the Vostok temperature derived from ice core signal [23, 25] and the length-of-day data (LOD) [15].

Regarding the Vostok temperature we set `NIMFs` to 100, for `alpha` we use the default value `ave`. If we use the default value `Xi`= 1.6, the method decomposes the low-frequency component in several IMFs. Therefore we increase `Xi` up to 3 using the script `Settings_IF('IF.NIMFs', 100, 'IF.Xi', 3);` We obtain a decomposition which contains 5 IMFs and a remainder, left panel of Figure 5. The first three components are related to high-frequency oscillatory patterns, whereas the fourth, fifth, and last component in the decomposition correspond to three

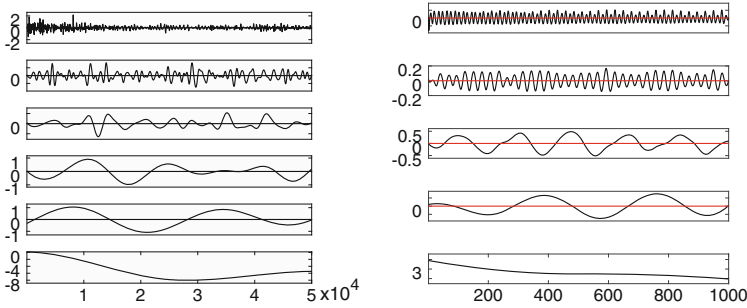


Fig. 5 Left panel, Vostok temperature IF decomposition. Right panel, length-of-day IF decomposition

Milankovitch cycles which are related to the Earth’s eccentricity, axial tilt, and precession [19, figure 11].

For the LOD data we use the `Almost_min` option for the parameter `alpha` and we set `NIMFs` to 100. Also in this example we have to increase `Xi` up to 3.

We use the script `Settings_IF('IF.NIMFs',100,'IF.alpha','Almost_min','IF.Xi',3);`. We obtain a decomposition which contains 4 IMFs and a remainder, right panel of Figure 5. The first IMF has quasiregular extrema with an average period around 14 days which was related to semimonthly tides [21]. The second IMF has an average period of about 28 days, which was linked to monthly tides [19]. The third and fourth IMFs are semiannual and annual components, respectively. The causes of these cycles in the length of day have been attributed to both the semiannual and annual cycles of the atmospheric circulation and to other factors, such as tidal strength change related to the revolution of the Earth around the Sun [16]. All the IMFs produced by the technique have physical meaning and are all and only the oscillatory patterns we expected to find in the given signal [19].

4 Conclusions and Outlook

In this paper we provide the reader with a brief overview of methods for the decomposition of nonstationary signals. We address more in detail the question of how to decompose such signals into simpler components using iterative methods which do not require any a priori assumptions.

From a research point of view, while the EEMD is still lacking a rigorous mathematical analysis, the IF algorithm has been extensively studied and analyzed in 1D and higher dimensions [3–6, 8]. However, it is important to point out that the iterative filtering method may fail to meaningfully decompose signals whose instantaneous frequencies vary consistently over time, for instance, in a chirp. For this reason the adaptive local iterative filtering (ALIF) method has been proposed in [6]. It generalizes the IF algorithm allowing for meaningful decompositions of any kind of nonstationary signals even with wide and quick changes in the instantaneous

frequencies, like chirps. However, ALIF mathematical understanding is far from being complete [8]. More research has to be done in this direction.

Acknowledgements The author’s research was supported by Istituto Nazionale di Alta Matematica (INdAM) “INdAM Fellowships in Mathematics and/or Applications cofunded by Marie Curie Actions,” PCOFUND-GA-2009-245492 INdAM-COFUND Marie Skłodowska Curie Integration Grants.

The author is deeply grateful to Haomin Zhou, a great researcher and a wonderful person. He contributed substantially to this work and to the author career with many suggestions and pieces of advice he gave to the author over the years.

References

1. Auger, F., Flandrin, P., Lin, Y. T., McLaughlin, S., Meignen, S., Oberlin, T., Wu, H.-T.: Time–frequency reassignment and synchrosqueezing: An overview. *IEEE Signal Processing Magazine*, 30, 32–41 (2013)
2. Bracewell, R. N., Bracewell, R. N.: *The Fourier transform and its applications*, McGraw-Hill, New York (1986)
3. Cicone, A., Dell’Acqua, P.: Study of boundary conditions in the Iterative Filtering method for the decomposition of nonstationary signals. Preprint. ArXiv 1811.07610
4. Cicone, A., Zhou, H.: Multidimensional iterative filtering method for the decomposition of high–dimensional non–stationary signals. *Numer. Math. Theory Methods Appl.*, 10, 278–298 (2017). <https://doi.org/10.4208/nmtma.2017.s05>
5. Cicone, A., Zhou, H.: Numerical Analysis for Iterative Filtering with New Efficient Implementations Based on FFT. Submitted. ArXiv 1802.01359
6. Cicone, A., Liu, J., Zhou, H.: Adaptive local iterative filtering for signal decomposition and instantaneous frequency analysis. *Appl. Comput. Harmon. Anal.*, 41, 384–411 (2016). <https://doi.org/10.1016/j.acha.2016.03.001>
7. Cicone, A., Liu, J., Zhou, H.: Hyperspectral chemical plume detection algorithms based on multidimensional iterative filtering decomposition. *Phil. Trans. R. Soc. A: Math. Phys. Eng. Sci.*, 374, 20150196 (2016). <https://doi.org/10.1098/rsta.2015.0196>
8. Cicone, A., Garoni, C., Serra-Capizzano, S.: Spectral and convergence analysis of the Discrete ALIF method. Submitted. <http://www.it.uu.se/research/publications/reports/2017-018/>
9. Cohen, L.: *Time–frequency Analysis*. Prentice Hall (1995)
10. Daubechies, I.: *Ten lectures on wavelets*. SIAM (1992)
11. Daubechies, I., Maes, S.: A nonlinear squeezing of the continuous wavelet transform based on auditory nerve models. *Wavelets in Medicine and Biology*, 527–546 (1996).
12. Daubechies, I., Lu, J., Wu, H.-T.: Synchrosqueezed wavelet transforms: An empirical mode decomposition–like tool. *Appl. Comput. Harmon. Anal.*, 30, 243–261 (2011)
13. Flandrin, P.: *Time–frequency/time–scale analysis*. Academic press (1998)
14. Flandrin, P., Chassande-Mottin, E., Auger, F.: Uncertainty and spectrogram geometry. *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 794–798 (2012)
15. Gross, R. S.: Combinations of Earth–orientation measurements: SPACE97, COMB97, and POLE97. *Journal of Geodesy*, 73, 627–637 (2000)
16. Höpfner, J.: Seasonal variations in length of day and atmospheric angular momentum. *Geophys. J. Int.*, 135, 407–437 (1998). <https://doi.org/10.1046/j.1365-246X.1998.00648.x>
17. Hou, T.Y., Shi, Z.: Adaptive data analysis via sparse time–frequency representation. *Adv. in Adap. Data Anal.*, 3, 1–28 (2011)
18. Hou, T.Y., Yan, M.P., Wu, Z.: A variant of the EMD method for multi–scale data. *Adv. in Adap. Data Anal.*, 1, 483–516 (2009)

19. Huang, N. E., Wu, Z.: A review on Hilbert–Huang transform: Method and its applications to geophysical studies. *Reviews of Geophysics*, 46 (2008)
20. Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., Yen, N. C., Tung, C. C., Liu, H. H.: The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proc. R. Soc. A: Math. Phys. Eng. Sci.*, 454, 903–995 (1998)
21. Huang, N. E., Wu, M. L., Long, S. R., Shen, S. S., Qu, W. D., Gloersen, P., Fan, K. L.: A confidence limit for the position empirical mode decomposition and Hilbert spectral analysis. *Proc. R. Soc. London, Ser. A*, 459, 2317–2345 (2003)
22. Lin, L., Wang, Y., Zhou, H.: Iterative filtering as an alternative algorithm for empirical mode decomposition. *Adv. Adapt. Data Anal.*, 1, 543–560 (2009)
23. Petit, J. R., Jouzel, J., Raynaud, D., Barkov, N. I., Barnola, J. M., Basile, I., Bender, M., Chappellaz, J., Davis, M., Delaygue, G. et al.: Climate and atmospheric history of the past 420,000 years from the Vostok ice core, Antarctica. *Nature*, 399, 429–436 (1999). <https://doi.org/10.1038/20859>.
24. Piersanti, M., Materassi, M., Cicone, A., Spogli, L., Zhou, H., Ezquer R. G.: Adaptive Local Iterative Filtering: a promising technique for the analysis of non-stationary signals. *Journal of Geophysical Research – Space Physics*. <https://doi.org/10.1002/2017JA024153>
25. Saltzman, E. S., Petit, J. R., Basile, I., Leruyet, A., Raynaud, D., Lorius, C., Jouzel, J., Stievenard, M., Lipenkov, V. Y., Barkov, N. I., et al.: Four climate cycles in Vostok ice core. *Nature*, 387, 359–360 (1997). <https://doi.org/10.1038/387359a0>.
26. Wu Z., Huang, N. E.: Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Adv. Adapt. Data Anal.*, 1, 1–41 (2009)
27. Wu, Z., Huang, N. E., Chen, X.: The Multi-Dimensional Ensemble Empirical Mode Decomposition Method. *Advances in Adaptive Data Analysis*, 1, 339–372 (2009)

Modeling the Socio-Economic Waste Generation Factors Using Artificial Neural Network: A Case Study of Gurugram (Haryana State, India)



Ajay Satija, Dipti Singh, and Vinai K. Singh

1 Introduction

Solid waste management process includes managing the waste generation, collection with segregation, transportation with proper treatment (compost formation, recycling, incineration, plasma gasification), and final disposal [8]. Improper solid waste management causes serious environmental issues [17, 19, 21, 28]. Hence proper Municipal solid waste management is required. Basically fastest growing population, changing lifestyle, tourism, geographical conditions, public unawareness, and industrialization are the socio-economic factors responsible for municipal solid waste generation in developed as well as developing countries [15, 25]. The major socio-economic factors affecting the waste generation are population growth, rate of urbanization, literacy rate, and per capita income. The population growth effects waste generation [9, 12, 18, 31]. Numbers of members in household and corresponding waste generation are strongly correlated with each other [29]. Indian population growth rate in census year 2011 was 17.6 [5]. Waste generation rate in developed countries is higher than India. Rapid growing urbanization also effects waste generation [2]. Per capita income and amount of waste generation are also highly correlated. Hence, economic prosperity and population density are highly correlated to waste generation. In the previous studies the time series models have illustrated various trends in waste generation in successful manner [10]. Some

A. Satija (✉) · V. K. Singh
Department of Applied Science and Humanities, Inderprastha Engineering College, Ghaziabad,
Uttar Pradesh, India
e-mail: aajaysatija@rediffmail.com

D. Singh
School of Vocational Studies & Applied Sciences, Gautam Buddha University, Gautam Budh
Nagar, Uttar Pradesh, India

models are based on autoregressive as well as exponential smoothing techniques [23]. Such models give only seasonal variations of waste but do not tell about the factors responsible for its generation and how to minimize this [20, 27]. On the contrary, the ANN models have learning capability to correlate the dependent and independent variables. In various case studies such models have been applied to correlate socio-economic factors as independent variables to the municipal solid waste generation as dependent variable. The ANN models have been extensively used since last 20 years to solve the various daily life practical problems [14]. Adamovic et al. [1] have developed structural-break general regression neural network model to observe the effect of economic crisis on waste generation prediction. Bayer et al. [3] have developed ANN model as well as regression model to analyze the leaching behavior of solidified waste. Jahandideh et al. [13] have developed ANN and multiple regression models to predict the medical waste of 50 hospitals of Fars province (Iran). Lomeling and Kenyi have developed ANN and ARMA (1,1) models to predict the weekly waste generated in Juba Town, South Sudan [16]. Back propagation learning algorithm has been used in ANN models to give optimized predicted results [7]. But, there are shortcomings of such ANN models. In complex ANN model the training process will take more time. Large numbers of data points have been required for such networks. The multiple regression models show very clear approach between predictors and response variable. In regression analysis, the predictor variables which are not very useful to fit the model have to be eliminated by evaluating the optimum value of adjusted R^2 . But in ANN models such type of methodology has not been considered. The ANN model is just like a black box. It is difficult to interpret the structure of ANN model. The problem of overfitting and underfitting of the data generally arises in ANN models. In this study the major socio-economic factors such as population, urban population, literate population, and per capita income have been analyzed which are responsible for municipal solid waste (MSW) generation in Gurugram district (Haryana State, India). There are two main objectives of the present study (1) to predict the collected MSW of Gurugram district for five years (January 2017-December 2021) and (2) to observe the socio-economic factors effect individually and collectively on waste collection of Gurugram district. The present study will be helpful for the authorities of Municipal Corporation of Gurugram for better future planning and management.

2 Materials and Methods

2.1 Case Study Area

The study area of present research work is Gurugram. Gurugram is the fastest growing metropolitan city of National Capital Region (India). The population of Gurugram is 1,514,432 people [6]. The people of the district are blessed with six seasons and better groundwater level.

There are 35 wards in the district. These wards are subdivided into blocks. There are large numbers of educational institutes in the city. India's top ranked business schools, technical institutes, and reputed private universities are situated in the city.

Gurugram become third highest in per capita income in India. There are headquarters of world famous companies such as Coca-Cola, Pepsi, BMW, and Maruti in the city. Faridabad district is in east, Rewari district is in west, National capital Delhi is in north, and Mewat district is in its south direction.

2.2 *Material and Methods*

The socio-economic data of population growth, urban population %, and literacy rate have been assembled from the Census of India. Per capita income data has been assembled from Department of Economic and Statistical Analysis, Haryana (India). The municipal solid waste data has been collected from landfill site of Bandhwari village Gurugram (Haryana, India).

2.2.1 **Population (POP)**

Gurugram is a well-known industrialized city. People are migrating from other places to the city in search of job opportunities, education as well as to get advanced medical facilities. The population density is 1204 people per km². Population wise Gurugram occupies fourth place in the Haryana state.

The population growth rate data has been obtained from District Census Handbook, Gurgaon (formally named) Census of India 2011, Haryana, series-07, part XII-A. The population data of previous years has been generated from the specified population growth rate from the Handbook.

Table 1 shows that the population growth rate of the district has been continuously increasing since last three decades. Population, the urban population, and the literate population of census years 2021 have been predicted by graphical method of population forecasting [22]. Population of census years 2021 would be 2373670. Table 1 presents the population growth rate of Gurugram district.

Table 1 Population growth of Gurugram district

Census Year	Population growth %	Population
1951		201699
1961	28.7	259587
1971	34.1	348106
1981	35.5	471683
1991	28.6	606585
2001	44.2	874695
2011	73.1	1514432

Source [4, 6]

Table 2 Urban population of Gurugram district

Census Year	Urban Population%	Urban Population
1951	10.62	21420
1961	13.07	33928
1971	13.81	48073
1981	18.61	87780
1991	20.30	123137
2001	35.41	309703
2011	68.82	1042232

Table 3 Literacy rate % of Gurugram district

Census Year	Literacy rate %	Literate Population
1971	27.20	94685
1981	34.66	163485
1991	52.61	319124
2001	78.50	343135
2011	84.70	1111116

Source: [4, 6, 30]

2.2.2 Urban Population (URB)

Urbanization drastically effects waste generation. Gurugram is the second highly urbanized district after Faridabad in Haryana state, India. Gurugram is the IT (information technology) industries hub. There are plenty of job opportunities in the city.

The Delhi metro train provides transport facilities to the people. The urban population % of the district of previous census years has been illustrated in Table 2. The urban population data has been generated with the help of Table 1. The urban population in census years 2021 would be approximately 1996409. Source: [4, 6, 26]

2.2.3 Literate Population (LIT)

The literacy rate of Gurugram district is highest among all districts in the Haryana state. Literacy rate creates public awareness to minimize the waste. Well-equipped, trained sanitary staff always helps to minimize the waste. The illiterate people create obstacles in the waste management system. They generate waste, spread it in anywhere in public area, and create various environmental issues. Waste management in such areas is quite difficult and expensive. Literate population for year 2021 would be approximately 1924206 (Table 3).

2.2.4 Per Capita Income (PCI)

It has been observed that Gurugram district generates highest revenue for the Haryana state. Gurugram is also called millennium city. The city is best place for

Table 4 Per capita income of the Gurugram district

Financial Year	Per Capita Income (Rs/-)
2004–05	81478
2005–06	165878
2006–07	181730
2007–08	199095
2008–09	206817
2011–12	305233
2012–13	333168
2013–14	355343
2014–15	388278
2015–16	415959
2016–17	443641
2017–2018	471323

Source: [11]

Table 5 MSW collected (Kg) (April 2010–January 2017)

Year	Municipal solid waste (Kg)
2010	84167945
2011	133604460
2012	171003215
2013	144729705
2014	173499200
2015	189410855
2016	221569015
January 2017	20750290

real estate investors in Northern India. Industrial growth and expensive residential complexes, sky-scraping shopping mall create an eminent picture of the millennium city.

This is a well-planned city. High class Malls such as Ambience Mall, Sahara Mall, DLF Mall, and Central Mall make city as shopper’s paradise. Table 4 presents the per capita income of the district. Per capita income data from the years 2005–2010 has been obtained from final report (based on Economic profile of NCR (National Capital Region), 2015) submitted by Apex Cluster Development Services Pvt. Limited to National Capital Region Planning Board.

PCI data from 2011–2015 has been arranged from Department of Economic and Statistical Analysis, Haryana. PCI data for year 2016–2018 has been estimated on average basis from last four financial years

2.2.5 Municipal Solid Waste (MSW) Data

The MSW data records have been collected from April 2010 to January 2017 (82 months). MSW presents fluctuations due to seasonal variations, people migration, and festivals. Municipal Corporation of Gurugram (MCG) is the main governing body responsible for waste management in the city. Table 5 presents the MSW

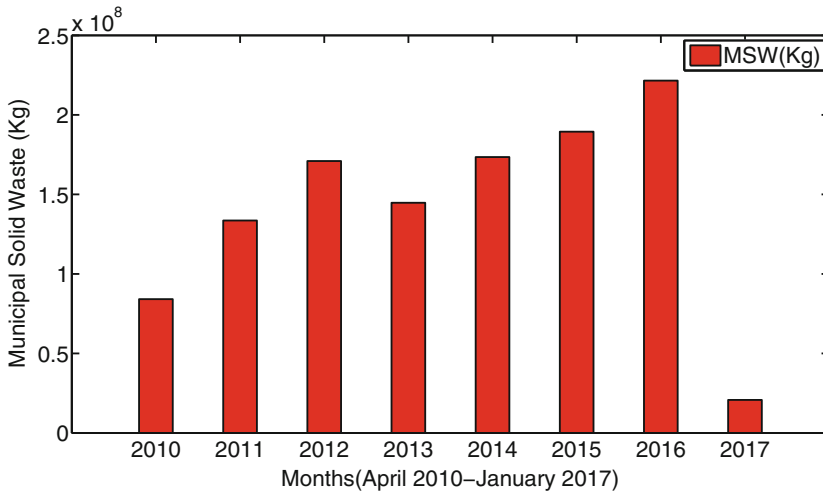


Fig. 1 Municipal solid waste (Kg) disposed at landfill site

Table 6 Statistical analysis of proposed socio-economic factors and MSW

Variable	Number of Months	Mean	Std Dev	Min	Max	Skewness
Population	82	1739265	171106	1453165	2032843	0.03
Urban Population	82	1310225	202198	967511	1650409	-0.01
Literate Population	82	1395762	209595	1025893	1727739	-0.11
Per Capita Income	82	376913	55594	278887	471323	0.01
Municipal solid waste	82	13854441	3771103	4164135	21910590	-0.54

collected and disposed in the landfill site of Bandhwari village, Gurugram. Figure 1 presents year-wise MSW collected from April 2010 to January 2017.

2.2.6 Statistical Analysis of Data

Table 6 presents the statistical analysis of socio-economic factors and MSW. The statistical analysis presents that the socio-economic factors and MSW data are normalized because the standard deviation in each case is less than half of the mean and skewness factor in case lies between -1 and 1.

Table 7 presents the coefficient of correlation and p-value of input variables of ANN model to the output variables.

2.2.7 Proposed Artificial Neural Network (ANN) Model

The ANN models have high learning capability from previous data. Such models have forecasting as well as extrapolating behavior. In this study MATLAB software

Table 7 The coefficient of correlation R, coefficient of determination R^2 of collected MSW with other socio-economic factors

Value	Population and MSW	Urban Population and MSW	Literacy Rate and MSW	Per Capita Income and MSW
(R value)	0.747	0.747	0.745	0.750
R^2 Value	55.8%	55.8%	55.4%	56.25%
p-value	0	0	0	0

has been used to develop ANN model to predict the MSW. Here neural network tool (nntool) has been used to create, train, test, validate, and simulate the neural network model. The ANN architecture has been denoted by I-H-O. Here I, H, and O presents input layer, hidden layer, and output layer neurons. The brief working of nntool has been described in the following steps:

Step 1. The variables population, urban population, literate population, and per capita income of Gurugram district have been used as input layer neurons. MSW (Monthly) has been used as output layer neuron. Here the data records of 82 months (April 2010–January 2017) of these variables have been compiled. Then the data of input-output neurons has been normalized between -1 and 1 by using the formula

$$X_{nor} = \frac{2 * (X - X_{min})}{X_{max} - X_{min}} - 1 \tag{1}$$

Here X_{nor} , X_{min} , and X_{max} present the normalized value, minimum, and maximum value of each variable say X .

Step 2. First read the normalized input(s) as well as target data points of 82 months through data network manager of nntool.

Step 3. Create ANN network by selecting network type, training function, number of layers, and number of neurons. The network will undergo supervised learning. Levenberg-Marquardt back propagation learning algorithm has been employed to adjust the weights and biases in the network so that the difference between observed and predicted MSW can be minimized. Hyperbolic tan sigmoid transfer function has been used between input and hidden layer neurons. The pure linear transfer function has been used between hidden and output layer neurons.

Step 4. Train and retrain the network by setting training parameters. The input-output data of 82 data samples (100%) has been randomly subdivided into 58 training data samples (70%), 12 validation data samples (15%), and 12 testing data samples (15%). Then note down the best validation performance and highest value of the coefficient of correlation between observed and predicted MSW.

Step 5. The input parameters population, urban population, literate population, and per capita income have been individually as well as collectively trained with

output variable MSW to observe their effect on MSW. Now simulate the network for the next 59 months extrapolated data values (February 2017–December 2021) of population, urban population, literate population, and per capita income to predict the MSW for next 5 years.

3 Results

The results of the present study have been compiled in Tables 8 and 9. One of the main objectives of study is to identify the effect of single, double, and multiple input variables on the output variable MSW. It is assumed that sanitation worker of MCG collects 70% of generated waste, respectively [24].

Initially, population variable has been used to predict the MSW. Various network structures have been tried by changing hidden layer neurons to predict best ANN model. The ANN model 1-6-1 has been observed as best ANN model. Here 1, 6, and 1 show input layer neuron population, number of hidden layer neurons, and MSW as output layer neuron, respectively.

The mean squared error of the model is 0.0474 and the coefficient of correlation between observed and predicted MSW is 0.8569. The expected collected and generated MSW due to this variable would be 1313237.06 and 1876052.94 Metric tons,

Table 8 Validation of proposed ANN models by mean squared error and coefficient of correlation R

Model No.	Input Variable(s)	Best ANN Model Structure	MSE of Total Network	Regression R			
				Training	Validation	Testing	All
1	POP	1-6-1	0.0474	0.8599	0.8047	0.9215	0.8569
2	URB	1-7-1	0.0485	0.8528	0.8750	0.8692	0.8564
3	LIT	1-7-1	0.0440	0.8561	0.9586	0.8986	0.8705
4	PCI	1-7-1	0.0364	0.8945	0.8046	0.9718	0.8924
5	POP, URB	2-7-1	0.0489	0.8230	0.9080	0.9356	0.8525
6	POP, LIT	2-7-1	0.0501	0.8542	0.8860	0.8806	0.8496
7	POP, PCI	2-7-1	0.0512	0.8282	0.9020	0.8940	0.8450
8	URB, LIT	2-8-1	0.0424	0.8480	0.9288	0.9640	0.8743
9	URB, PCI	2-8-1	0.0492	0.8359	0.9142	0.9123	0.8510
10	LIT, PCI	2-9-1	0.0440	0.8650	0.9057	0.8754	0.8709
11	POP, URB, LIT	3-10-1	0.0511	0.8339	0.8954	0.8595	0.8455
12	URB, LIT, PCI	3-9-1	0.0522	0.8168	0.9068	0.9231	0.8422
13	POP, LIT, PCI	3-10-1	0.0410	0.8634	0.9403	0.9077	0.8786
14	POP, URB, LIT, PCI	4-12-1	0.0294	0.9132	0.9065	0.9307	0.9150

Table 9 Expected collected and generated waste (Metric tons) by proposed ANN models

Model No.	Variables Used Input Variable(s) and output variable MSW	Expected Collected MSW within Years 2017–2021	MSW collected in Year 2021	Uncollected MSW within Years 2017–2021	Total expected MSW generated from year 2017 to 2021
1	POP	1313237.06	262919.84	562815.88	1876052.94
2	URB	1311584.53	262926.77	562107.65	1873692.18
3	LIT	1240532.39	248934.17	531656.73	1772189.12
4	PCI	1314109.73	262925.55	563189.88	1877299.61
5	POP, URB	1314059.53	262926.92	563168.37	1877227.90
6	POP, LIT	1306299.24	262924.08	559842.53	1866141.77
7	POP, PCI	1303999.57	262927.07	558856.95	1862856.52
8	URB, LIT	1302134.60	262912.50	558057.68	1860192.28
9	URB, PCI	1291066.93	262431.65	553314.39	1844381.32
10	LIT, PCI	1267575.91	21305.65	543246.81	1810822.72
11	POP, URB, LIT	1298316.00	262517.10	556421.14	1854737.14
12	URB, LIT, PCI	1290055.09	261070.81	552880.75	1842935.84
13	POP, LIT, PCI	1289331.37	262891.82	552570.58	1841901.95
14	POP, URB, LIT, PCI	1247096.43	260572.84	534469.89	1781566.32

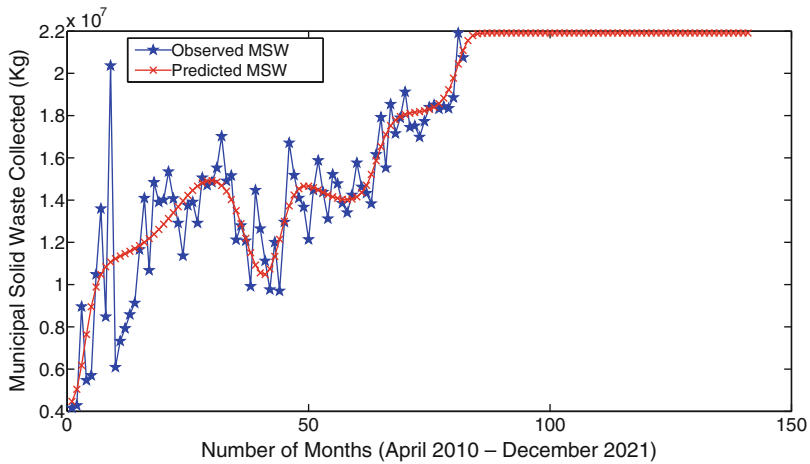


Fig. 2 Comparison between observed and predicted MSW of Gurugram using population as predictor variable ((1-6-1) model no. 1)

respectively, for the period (January 2017–December 2021). Figure 2 presents the comparison between observed and predicted MSW of Gurugram using population as predictor variable.

It can be determined in Figure 2 that the noise (error) is associated with proposed network. Hence only population factor is not just sufficient to explain the variability in response variable MSW. The forecasting behavior shown by this variable is linear for upcoming 5 years (2017–2021). It predicts the trend.

Figure 2 shows that the curve is not good fitted from 1 December 2010 to 1 November 2011. This error may be occurred due to other unseen factors. The model number 2 displays the urban population has been used as predictor variable individually to predict the MSW. By changing hidden layer neurons the ANN model 1-7-1 has been observed as best ANN model. The mean squared error between observed and predicted MSW of the proposed model is 0.0485 and the coefficient of correlation between observed and predicted MSW is 0.8564. The expected collected and generated MSW due to urban population would be 1311584.53 and 1873692.18 Metric tons, respectively, for the period (2017–2021).

Figure 3 presents the comparison between observed and predicted MSW of Gurugram using urban population as predictor variable. This is observed from Figure 3 that the noise (error) is also associated with proposed network.

In model number 3 the literate population has been used to predict the MSW individually. The ANN model 1-7-1 has been examined as best ANN model by changing hidden layer neurons.

Figure 4 presents the comparison between observed and predicted MSW of Gurugram using literate population as predictor variable. The MSE and coefficient of correlation between observed and predicted MSW are 0.0440 and 0.8705, respectively. The expected waste collected and generated for period January 2017–December 2021 would be approximately 1240532.39 and 1772189.12 Metric ton, respectively.

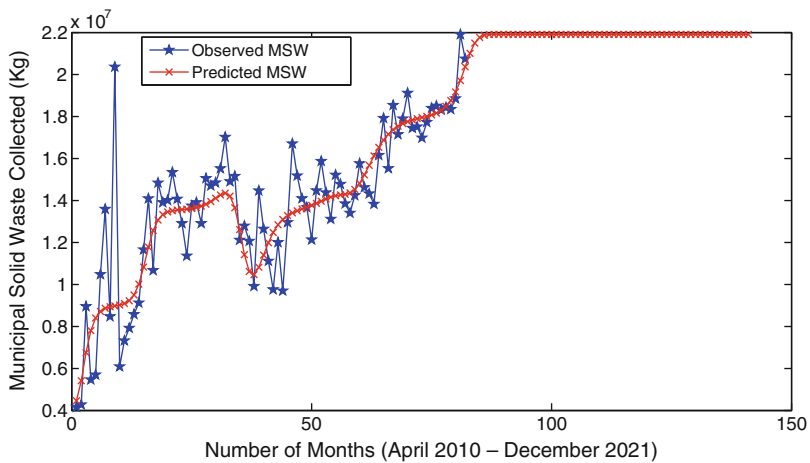


Fig. 3 Comparison between observed and predicted MSW of Gurugram using urban population as predictor variable ((1-7-1) model no. 2)

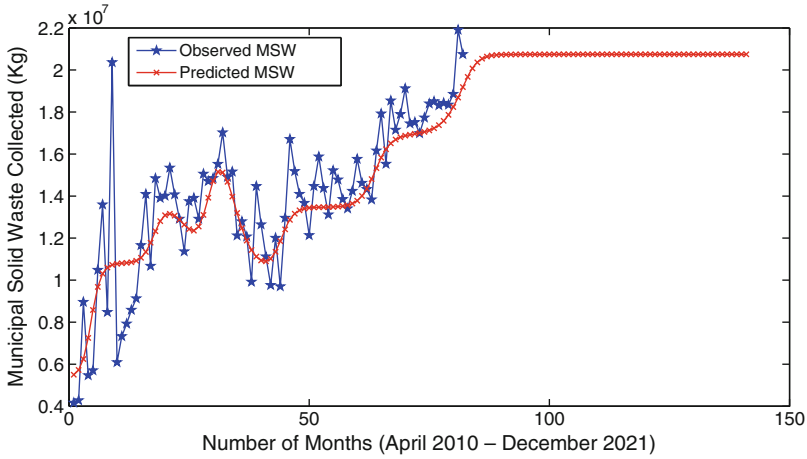


Fig. 4 Comparison between observed and predicted MSW of Gurugram using literate population as predictor variable ((1-7-1) model no. 3)

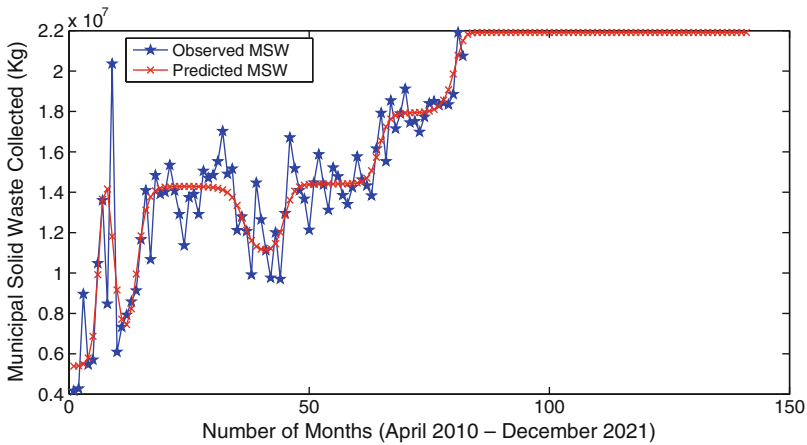


Fig. 5 Comparison between observed and predicted MSW of Gurugram using per capita income as predictor variable ((1-7-1) model number 4)

In model number 4 the variable per capita income has been used to predict the MSW.

The ANN model 1-7-1 has been observed as best ANN model by changing hidden layer neurons. The MSE and coefficient of correlation between observed and predicted MSW are 0.0364 and 0.8924, respectively. The expected collected and generated waste for period January 2017–December 2021 would be approximately 1314109.73 and 1877299.61 Metric ton, respectively.

Figure 5 presents the comparison between observed and predicted MSW of Gurugram using per capita income as predictor variable. Per capita income has

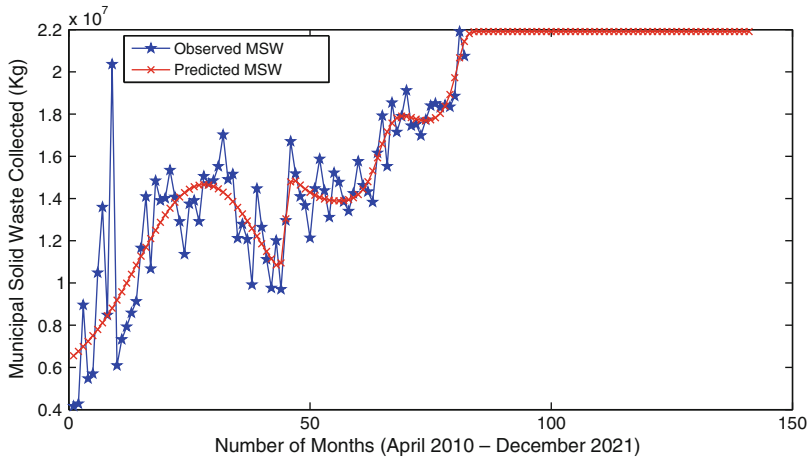


Fig. 6 Comparison between observed and predicted MSW of Gurugram using population and urban population as predictor variable ((2-7-1) model no. 5)

shown highest predictive behavior than the rest three variables individually because the coefficient of correlation between observed and predicted MSW is high 0.8924.

The model number 5 presents that the variables population and urban population have been used to predict the MSW. The ANN model 2-7-1 has been found as best ANN model. In the ANN Model 2-7-1, 2 shows population and urban population as input layer neurons, 7 shows the hidden layer neurons are 7, and 1 shows MSW as output layer neuron. The MSE and coefficient of correlation between observed and predicted MSW are 0.0489 and 0.8525, respectively. The expected collected and generated waste for period 2017–2021 would be approximately 1314059.53 and 1877227.9 Metric ton, respectively. Figure 6 presents the comparison between observed and predicted MSW of Gurugram using population and urban population as predictor variables collectively.

In model number 6 the variables population and literate population collectively have been used to predict the MSW. Figure 7 presents the comparison between observed and predicted MSW of Gurugram using population and literate population as predictor variables

The ANN model 2-7-1 has been observed as best ANN model. The MSE and coefficient of correlation between observed and predicted MSW are 0.0501 and 0.8496, respectively. The expected waste collected and generated for period January 2017–December 2021 would be approximately 1306299.24 and 1866141.77 Metric tons, respectively. The model number 7 shows that the variables population and per capita income have been used to predict the MSW.

The ANN model 2-7-1 has been observed as best ANN model. The MSE and coefficient of correlation between observed and predicted MSW are 0.0512 and 0.8450, respectively. The expected waste collected and generated for period January

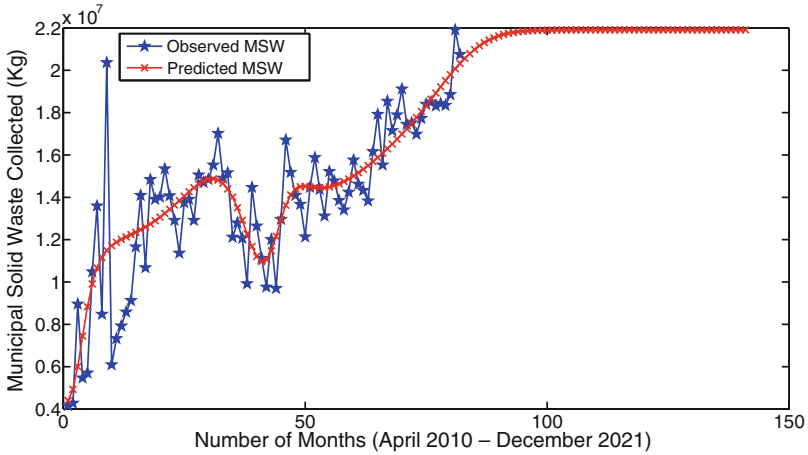


Fig. 7 Comparison between observed and predicted MSW of Gurugram using population and literate population as predictor variables ((2-7-1) model no. 6)

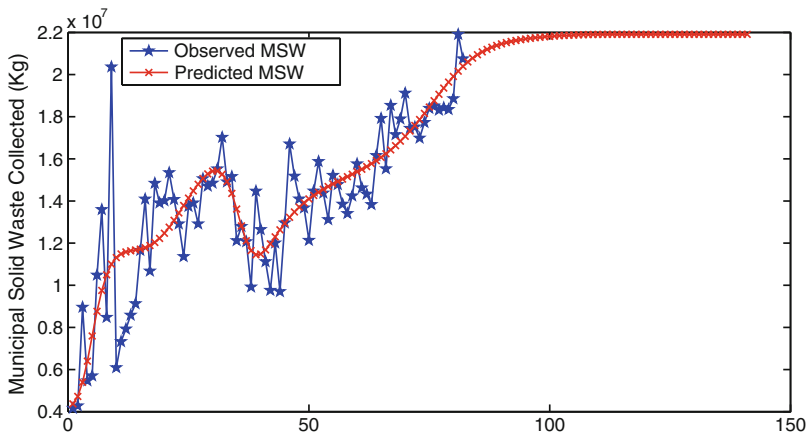


Fig. 8 Comparison between observed and predicted MSW of Gurugram using population and per capita income as predictor variables ((2-7-1) model no. 7)

2017–December 2021 would be approximately 1303999.57 and 1862856.52 Metric tons, respectively.

Figure 8 illustrates the comparison between observed and predicted MSW of Gurugram using population and per capita income as predictor variables collectively.

The variables urban population and literate population collectively have been used to predict the MSW in model number 8. The ANN model 2-8-1 has been observed as best ANN model by changing hidden layer neurons. The MSE and coefficient of correlation between observed and predicted MSW are 0.0424 and 0.8743, respectively.

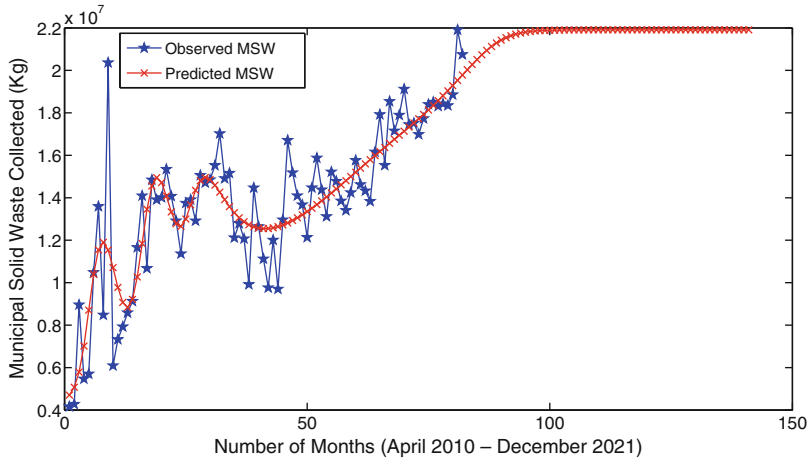


Fig. 9 Comparison between observed and predicted MSW of Gurugram using urban population and literate population as predictor variables ((2-8-1) model no. 8)

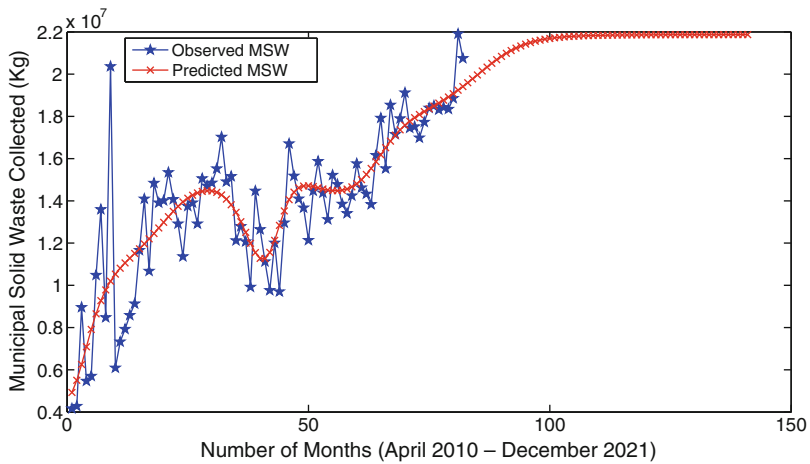


Fig. 10 Comparison between observed and predicted MSW of Gurugram using urban population and per capita income as predictor variables ((2-8-1) model no. 9)

The expected waste collected and generated for period January 2017–December 2021 would be approximately 1302134.60 and 1860192.28 Metric tons, respectively. Figure 9 presents the comparison between observed and predicted MSW of Gurugram using urban population and literate population income as predictor variables.

The model number 9 shows that the variables urban population and per capita income have been used to predict the MSW.

Figure 10 presents the comparison between observed and predicted MSW of Gurugram using urban population and per capita income as predictor variables.

The ANN model 2-8-1 has been observed as best ANN model. The MSE and coefficient of correlation between observed and predicted MSW are 0.0492 and 0.8510, respectively. The expected waste collected and generated for period January 2017–December 2021 would be approximately 1291066.93 and 1844381.32 Metric tons, respectively.

The model number 10 shows that the variables literate population and per capita income have been used to predict the MSW.

The ANN model 2-9-1 has been observed as best ANN model by changing hidden layer neurons. The MSE and coefficient of correlation between observed and predicted MSW are 0.0440 and 0.8709, respectively. The expected waste collected and generated for period January 2017–December 2021 would be approximately 1267575.91 and 1810822.72 Metric tons, respectively.

Figure 11 illustrates the comparison between observed and predicted MSW of Gurugram using literate population and per capita income as predictor variables. Hence, model no. 8 and model no. 10 have shown high predictive results.

Model number 11 illustrates that the variables population, urban population, and literate population have been used to predict the MSW. The ANN model 3-10-1 has been observed as best ANN model. Here 3 shows the variables population, urban population, and literate population as input layer neurons, 10, hidden layer neurons, and 1, MSW as output layer neuron. The MSE and coefficient of correlation between observed and predicted MSW are 0.0511 and 0.8455, respectively. The expected waste collected and generated for period January 2017–December 2021 would be approximately 1298316.00 and 1854737.14 Metric tons, respectively. Figure 12 presents the comparison between observed and predicted MSW of Gurugram using population, urban population, and literate population as predictor variables.

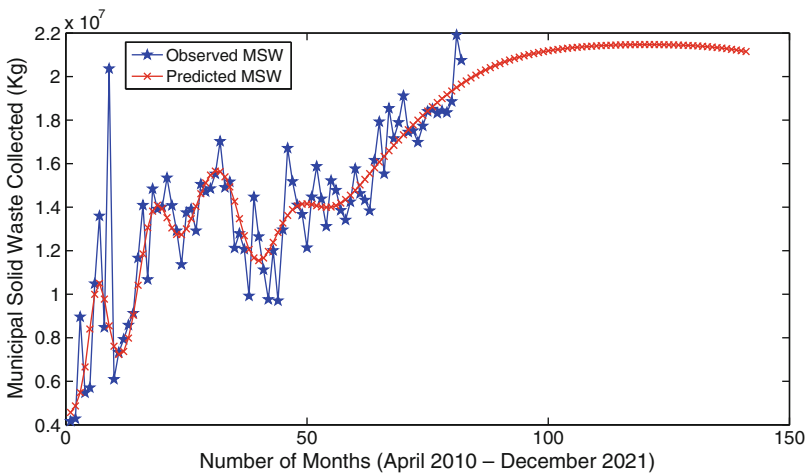


Fig. 11 Comparison between observed and predicted MSW of Gurugram using literate population and per capita income as predictor variables ((2-9-1) model no. 10)

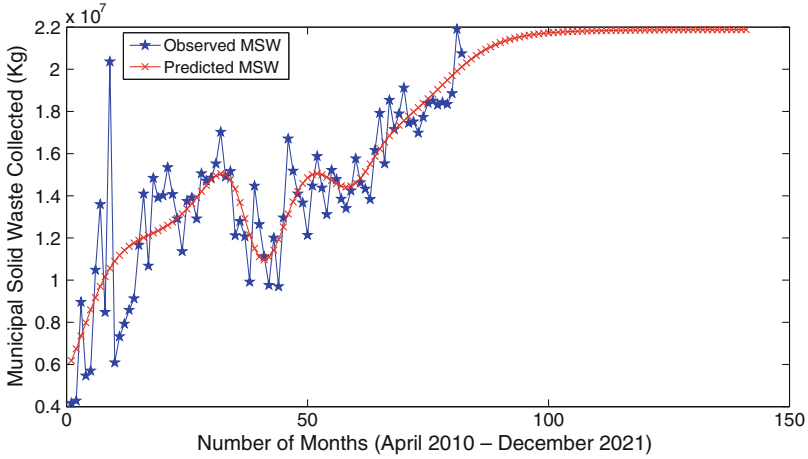


Fig. 12 Comparison between observed and predicted MSW of Gurugram using population, urban population, and literate population as predictor variables ((3-10-1) model no. 11)

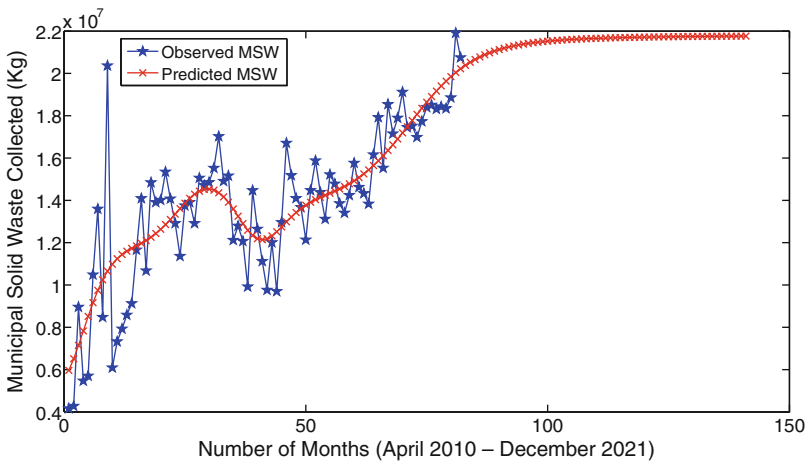


Fig. 13 Comparison between observed and predicted MSW of Gurugram using urban population, literate population, and per capita income as predictor variables ((3-9-1) model no. 12)

The variables urban population, literate population, and per capita income have been used to predict the MSW in model number 12.

The ANN model 3-9-1 has been observed as best ANN model by changing hidden layer neurons. The MSE and coefficient of correlation between observed and predicted MSW are 0.0522 and 0.8422, respectively. The expected waste collected and generated for period January 2017–December 2021 would be approximately 1290055.09 and 1842935.84 Metric tons, respectively. Figure 13 presents the comparison between observed and predicted MSW of Gurugram using urban population, literate population, and per capita income as predictor variables.

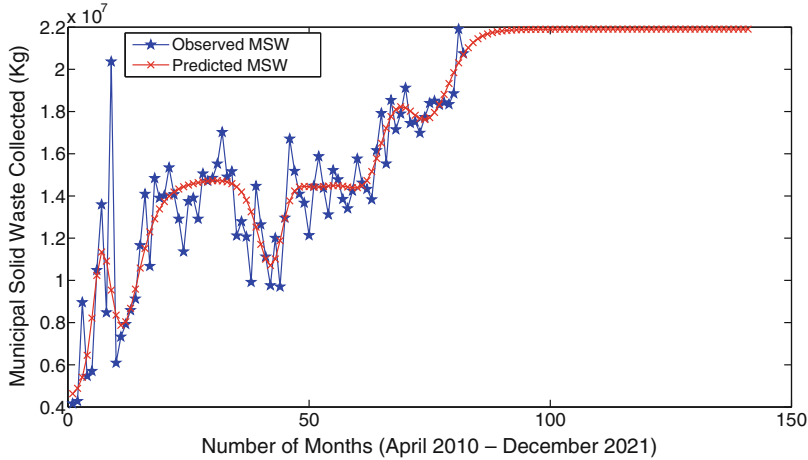


Fig. 14 Comparison between observed and predicted MSW of Gurugram using population, literate population, and per capita income as predictor variables ((3-10-1) model no. 13)

In model number 13, the variables population, literate population, and per capita income have been used to predict the MSW. The ANN model 3-10-1 has been observed as best ANN model by changing hidden layer neurons. The MSE and coefficient of correlation between observed and predicted MSW are 0.0410 and 0.8786, respectively. The expected waste collected and generated for period January 2017–December 2021 would be approximately 1289331.37 and 1841901.95 Metric tons, respectively. Figure 14 presents the comparison between observed and predicted MSW of Gurugram using population, literate population, and per capita income as predictor variables. Model no. 13 has shown high predictive results than model no. 11 and 12. In model number 14 all the variables population, urban population, literate population, and per capita income have been included to predict the MSW. The MSE and coefficient of correlation between observed and predicted MSW are 0.0294 and 0.9150, respectively. Hence Model 14 is the best predictive model. The expected waste collected and generated for period January 2017–December 2021 would be approximately 1247096.43 and 1781566.32 Metric tons, respectively. Figure 15 presents the comparison between observed and predicted MSW of Gurugram using population, urban population, literate population, and per capita income as predictor variables. It was necessary to include hidden layer neurons because if no hidden layer gets selected, then the ANN model will become linear model and the quality of prediction get reduced.

All above results have been compiled in Tables 8 and 9.

The physical composition of waste samples of Gurugram has been obtained from the article “Gurugram: A Framework For Sustainable Development (2017)” published by Centre for Science and Environment [24]. The percentage of physical composition of waste has been shown in Figure 16.

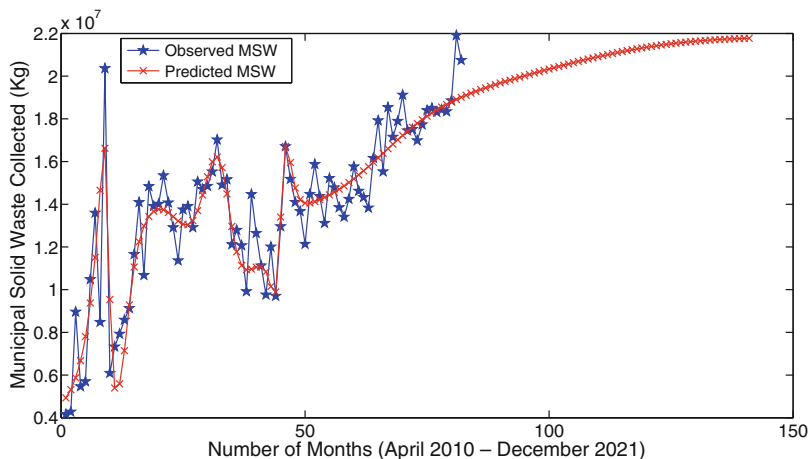


Fig. 15 Comparison between observed and predicted MSW of Gurugram using population, urban population, literate population, and per capita income as predictor variables (4-12-1) model no. 14)

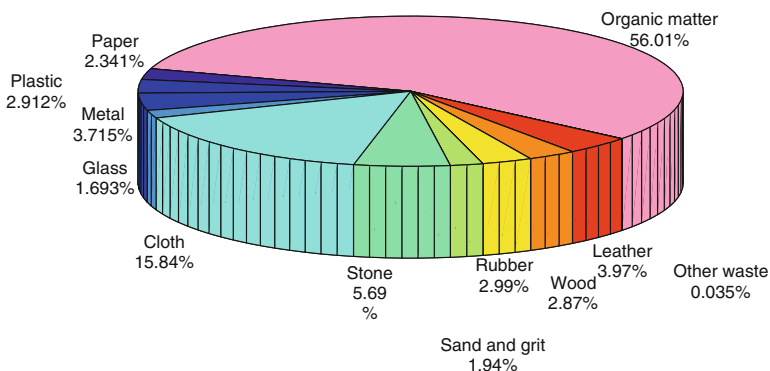


Fig. 16 Physical composition of waste samples of Gurugram [24]

The expected physical composition of waste produced (in Metric tons) has been shown in Figure 17 for period 2017–2021 (by using model no. 14 results).

It is found from the study [24] that about 50–52% of the waste is bio-degradable, 12–15% of the waste is recyclable, and 30–35% of the waste is inert. The predictive results obtained from model 14 have shown that the amount of bio-degradable, recyclable, and inert waste generated would lie at most in range (890783.16, 926414.49), (213787.95, 267234.94), and (534469.89, 623548.21) Metric tons, respectively. Hence 50–52% waste generated could be used in aerobic compost formation or in anaerobic biodegradation where biogas can be recovered and can be used as energy resource. Rest 12–15% waste can be

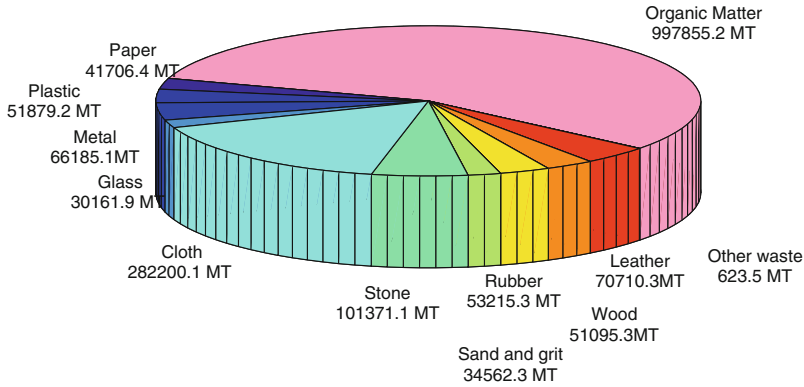


Fig. 17 Expected physical composition of waste produced (in Metric tons) (within period 2017–2021)

reused by recycling. Hence at most 62–67% (1104571–1193649.44 Metric tons) waste can be minimized in the period 2017–2021. However 30–35% inert waste can be used to fill up low lying areas or disposed of in well-designed landfill site.

4 Conclusion

Various ANN models have been developed to observe the effect of socio-economic factors such as population, urban population, literate population, and per capita income individually and collectively on collection and generation of municipal solid waste of Gurugram district, Haryana state, India. The ANN model (4-12-1) based on all these factors has shown the least mean squared error (0.0294) and high coefficient of correlation (0.9150) between observed and predicted MSW. The proposed model (4-12-1) has shown the better predictive results. The model (4-12-1) outcomes predict that 1247096.43 and 1781566.329 Metric tons waste would be collected as well as generated in period 2017–2021. It is also observed that the amount of bio-degradable, recyclable, and inert waste generated would lie in range (890783.16, 926414.49), (213787.95, 267234.94), and (534469.89, 623548.21) Metric tons, respectively. About 62–67% (1104571–1193649.44 Metric tons) waste can be minimized in the period 2017–2021 through compost formation and recycling. It is expected that the proposed research outcomes will be helpful for the authorities of Municipal Corporation of Gurugram for better planning and future management.

Acknowledgements The authors are very thankful to Dr. David Lomeling (Department of Agricultural Sciences, College of Natural Resources and Environmental Studies (CNRES), University of Juba, Juba South Sudan) for his valuable and informative suggestions in completion of this study.

References

- Adamovic, V.M., Antanasijevic, D.Z., Ristic, M.D., Peric-Grujic, A. A., Pocajt, V.V.: Prediction of municipal solid waste generation using artificial neural network approach enhanced by structural break analysis. *Environ Sci Pollut Res*, 24, 299–311, (2016). <https://doi.org/10.1007/s11356-016-7767-x>
- Asnani, P.U.: India Infrastructure Report 2006: 8. Solid Waste Management (2006). http://www.iitk.ac.in/3inetwork/html/reports/IIR2006/Solid_Waste.pdf. Accessed 2 January 2018
- Bayar, S., Demir, I., Engin, G.O.: Modeling leaching behavior of solidified wastes using back-propagation neural networks. *Ecotoxicology and Environmental Safety* 72, 843–850 (2009). <https://doi.org/10.1016/j.ecoenv.2007.10.019>
- Census of India 2011, Haryana, Series 7, Part XII-A, District Census Handbook Gurgaon. Directorate of Census Operations, Haryana http://www.censusindia.gov.in/2011census/dchb/DCHB_A/06/0618_PART_A_DCHB_GURGAON.pdf. (2011). Accessed 2 January 2018
- Census of India, 2011. Ministry of Home Affairs, Government of India, New Delhi, India. <http://censusindia.gov.in/2011-common/censusdataonline.html> (2011). Accessed 2 January 2018
- Census of India, 2011. <http://www.census2011.co.in/census/district/225-gurgaon.html> (2011). Accessed 2 January 2018.
- Chau, K.W., Wu, C.L.: A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *J Hydroinformatics* 12, 458–473 (2010). <https://doi.org/10.2166/hydro.2010.032>
- Chen, X., Geng, Y., Fujita, T.: An overview of municipal solid waste management in China. *Waste Management* 30, 716–724 (2010). <https://doi.org/10.1016/j.wasman.2009.10.011>
- Chiemchaisri, C., Juanga, J. P., Visvanathan, C.: Municipal solid waste management in Thailand and disposal emission inventory. *Environ Monit Assess* 135, 13–20 (2007). <https://doi.org/10.1007/s10661-007-9707-1>
- Denafas, G., Ruzgas, T., Martuzevicius, D., Shmarin, S., Hoffmann, M., Mykhalenko, V., Ogorodnik, S., Romanov, M., Neguliaeva, E., Chusov, A., Turkadze, T., Bochoidez, I., Ludwig, C.: Seasonal variation of municipal solid waste generation and composition in four East European cities. *Resour. Conserv. Recycl.* 89, 22–30 (2014).
- Economic Profile of NCR, Final Report. http://ncrpb.nic.in/pdf_files/FinalReportofstudyofeconomicprofile_17122015.pdf.(2015). Accessed 06.08.2017
- Grossmann, D., Hudson, J. F., Marks, D. H.: Waste generation models for solid waste collection. *Journal of the Environmental Engineering Division* 100, 1219–1230 (1974)
- Jahandideh, S., Jahandideh, S., Asadabadi, E.B., Askarian, M., Movahedi, M.M., Hosseini, S., Jahandideh, M.: The use of artificial neural networks and multiple linear regression to predict rate of medical waste generation. *Waste Manage* 29, 2874–2879 (2009)
- Kalogirou, S.A.: Artificial intelligence for the modeling and control of combustion processes: a review. *Progress in Energy and Combustion Science* 29, 515–566 (2003).
- Liu, C., Wu, X. W.: Factors influencing municipal solid waste generation in China: a multiple statistical analysis study. *Waste Management and Research* 29, 371–378 (2010). <https://doi.org/10.1177/0734242X10380114>
- Lomeling, D., Kenyi, S.W.: Forecasting solid waste generation in Juba Town, South Sudan using Artificial Neural Networks (ANNs) and Autoregressive Moving Averages (ARMA). *Journal of Environment and Waste Management* 4, 211–223, (2017)
- Louis, G.E.: A historical context of municipal solid waste management in the United States. *Waste Management and Research* 22, 306–322 (2004).
- Medina, M.: The effect of income on municipal solid waste generation rates for countries of varying levels of economic development: A model. *Journal of Resource Management and Technology*, 24,149–155 (1997)
- Meyers, G.D., Mcleod, G., Anbarci, M. A.: An international waste convention: measures for achieving sustainable development. *Waste Management and Research* 24, 505–513 (2006)

20. Noori, R., Abdoli, M.A., Farokhnia, A., Abbasi, M.: Results uncertainty of solid waste generation forecasting by hybrid of wavelet transform-ANFIS and wavelet transform-neural network. *Expert Syst. Appl.* 36, 9991–9999 (2009)
21. Pfammatter, R., Schertenleib, R.: Nongovernmental refuse collection in low-income urban areas (Lessons learned from selected schemes in Asia, Africa and Latin America. SANDEC Report No.1/96). *Water and Sanitation in Developing Countries*. Dübendorf, Switzerland: EAWAG/SANDEC (1996)
22. Population Forecasting. NPTEL IIT Kharagpur web courses. <http://nptel.ac.in/courses/105105048/M5L5.pdf>. Accessed 2 January 2018.
23. Rimaityte, I., Ruzgas, T., Denafas, G., Racys, V., Martuzevicius, D.: Application and evaluation of forecasting methods for municipal solid waste generation in an Eastern-European city. *Waste Management and Research* 30, 89–98 (2012)
24. Roychowdhury, A.: Gurugram a framework for sustainable development. Centre for Science and Environment, New Delhi, India. <http://www.cseindia.org/userfiles/gurugram-a-framework-for-sustainable-development-update.pdf> (2017). Accessed 2 January 2018
25. Saeed, M.O., Hassan, M.N., Mujeebu, M.A.: Assessment of municipal solid waste generation and recyclable materials potential in Kuala Lumpur, Malaysia. *Waste Management* 29, 2209–2213 (2009)
26. Sangwan, R.S.: Urbanization in Haryana during post-independence period: trends and patterns. *Radix International Journal of Research in Social Science* 2, 1–17 (2013)
27. Shan, C.S.: Projecting municipal solid waste: the case of Hong Kong SAR. *Resour. Conserv. Recycl.* 54, 759–768 (2010)
28. Sinha, A., Enayetullah, M. I.: Community based solid waste management: The Asian experience. Dhaka. *Waste Concern & USAID* (2000)
29. Suthar, S., Singh, P.: Household solid waste generation and composition in different family size and socio-economic groups: a case study. *Sustain. Cities Soc.* 14, 56–63 (2015)
30. Tilak, J.B.G: Educational Planning at Grassroots. In: Nangia, S.B. (eds) APH Publishing Corporation, New Delhi, India (2008)
31. Wertz, K. L.: Economic factors influencing household's production of refuse. *Journal of Environmental Management and Economics* 2, 263–272(1976)

Regularization of Highly Ill-Conditioned RBF Asymmetric Collocation Systems in Fractional Models



K. S. Prashanthi and G. Chandhini

1 Introduction

The fractional models are an efficient tool to describe any complex phenomenon in various fields such as the description of material hereditariness [23], mechanical and population models [6, 25], electrical circuits [13], and visco-elasto-plastic model [17]. Literature throws light on numerous efforts by mathematicians and scientists to solve these fractional models which are expressed in terms of fractional differential equations (FDE). Some of the initial or boundary fractional differential equations may be solved analytically by Laplace or Fourier transform methods [24]. Many semi-analytical approaches based on Adomian decomposition [22], homotopy perturbation [16], q-homotopy analysis, and variational iteration [29] have also been proposed to solve linear and nonlinear FDEs, and their analytical solutions are represented in the form of convergent series with easily computable components. But, if the obtained series is an unknown series, then the evaluation of solution is computationally complex. Also, the computation of fractional integrals and fractional derivatives of most of the functions by an analytical or semi-analytical method is not possible always. Hence we resort to numerical (approximation) methods and radial basis function method is one of the popular meshless methods used to solve linear as well as nonlinear FDEs.

Radial basis functions (RBFs) are considered as an effective tool in scattered data interpolation, solving differential equations, neural network, image processing, etc. In 1990 Kansa [18] introduced RBF based collocation method for both boundary and initial boundary value problems. Its simple characteristics like grid-free nature, higher order of convergence, and ease of extension to higher dimension lead it to

K. S. Prashanthi (✉) · G. Chandhini
National Institute of Technology Karnataka, Surathkal 575 025, India
e-mail: ksprashanthi@gmail.com; chandhini@nitk.edu.in

© Springer Nature Switzerland AG 2019
V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_5

105

use in complex phenomena arising in different areas. It has been observed that for many problems RBF accuracy increases with “flat” limit and flatness of the RBF depends on the shape parameter of the RBF. But, unfortunately, at the “flat” limit or as the number of nodes increases [3], the discretization leads to an ill-conditioned system which results in an unreliable solution. This motivated many efforts in the direction of reduction or removal of the ill-conditioning of the system.

Fornberg and Wright [11] have developed an algorithm named Contour-Padé to gain the stability of small shape parameter and therefore refuses the uncertainty principle described by Schaback [27]. But this algorithm is suitable for the case of flat RBFs with a small number of nodes. Also, Fornberg and Piret [10] introduced the RBF-QR algorithm which is computationally stable for flat RBFs interpolation and is faster and easier to implement rather than Contour-Padé and also it can be implemented for a large number of nodes. The key idea behind the RBF-QR method is to replace, in the case of small ε , the extremely ill-conditioned RBF basis with a well-conditioned one that spans exactly the same space. Then the coefficient (interpolation) matrix is factorized into a product as $Q \times E \times R$, where Q is unitary, E is diagonal, and R is upper triangular. Also, this algorithm has been generalized for node distributions in 2-D or 3-D [9, 19]. In [8], authors proposed a highly accurate least-squares approximation based on the early truncation of the kernel expansion. This method establishes the general connection between the RBF-QR algorithm and Mercer or Hilbert-Schmidt series expansions of a positive definite kernel. In [14], a new approach was introduced to avoid the inherent ill-condition in the computation of RBF-FD weights which was based on the semi-analytical computation of the Laurent series of the inverse RBF interpolation matrix. But the computational cost of the algorithm to obtain the Laurent series of the inverse grows exponentially with the order of the singularity.

Some of the works in the literature have shown that the regularization techniques can be efficient in handling ill-conditioning problems in RBF and Tikhonov regularization is one of the most popular regularization methods considered for this purpose. In [20], Lin introduced zeroth order Bessel function as a new RBF in collocation method. Arghand and Amirfakhrian [1] proposed a numerical scheme based on the fundamental solution and radial basis functions (RBFs) whereas Zhang and Li [30] used only RBF method to solve inverse heat equation. In all these works Tikhonov regularization (TR) along with generalized cross-validation (GCV) criterion is applied to tackle the ill-conditioning issue.

The main focus of the present work is to extend Tikhonov regularization (TR) for stabilizing the linear system obtained after discretizing linear fractional differential models using Kansa’s asymmetric collocation scheme. Then the proposed algorithm is applied to some important fractional models. To proceed further, consider linear fractional differential equation of the form

$${}^c D_x^{p\alpha} u + \sum_{l=0}^{p-1} a_l(x) {}^c D_x^{l\alpha} u = f(x), \quad x \in [a, b] \quad (1)$$

with appropriate initial and boundary conditions. Here, $p \in \mathbb{N}$, $0 < \alpha \leq 1$. a_l , ($l = 0, 1, \dots, (p - 1)$) and f are continuous functions defined on $[a, b]$ and ${}^c D_x^\alpha$ is the Caputo differential operator. In the present work p and α are considered such that $1 < p\alpha \leq 2$.

2 Preliminaries

In this section, we introduce the notations, definitions, and some of the fundamental assumptions that are considered in the present work [24].

2.1 Fractional Integrals and Derivatives

Definition 1 Let $\alpha \in \mathbb{R}_+$ and $u \in L_1[a, b]$. The operator ${}_a I_x^\alpha$, defined on $L_1[a, b]$ by

$${}_a I_x^\alpha u(x) = \frac{1}{\Gamma(\alpha)} \int_a^x (x - s)^{\alpha-1} u(s) ds \quad (2)$$

for $a \leq x \leq b$ is called the Riemann-Liouville (R-L) fractional integral operator of order α . For $\alpha = 0$ we set ${}_a I_x^0 = I$, the identity operator.

Definition 2 Let $A^m[a, b]$ denote the set of functions which have continuous derivatives up to order $(m - 1)$ on $[a, b]$ such that $u^{(m-1)}$ is absolutely continuous. The Caputo fractional derivative of $u \in A^m[a, b]$ of order $\alpha \in (m - 1, m]$ is defined as

$${}^c D_x^\alpha u(x) = {}_a I_x^{m-\alpha} \frac{d^m u(x)}{dx^m} = \frac{1}{\Gamma(m - \alpha)} \int_a^x (x - s)^{m-\alpha-1} \frac{d^m u(s)}{ds^m} ds \quad (3)$$

2.2 RBF Approximation

The radial basis functions are very successful in approximating multidimensional scattered data [4]. A brief overview of RBF interpolation is given below.

Given a finite set of n distinct points, $\{x_i, i = 1, 2, \dots, n\}$ in \mathbb{R} , along with the corresponding values $\{u(x_i)\}, i = 1, 2, \dots, n$ then the RBF interpolant $S(x)$ corresponding to $u(x)$ can be expressed as a linear combination of single univariate function which is called as radial basis function, i.e.,

$$S(x) = \sum_{j=1}^n \lambda_j \phi(|x - x_j|), \quad x \in [a, b]. \quad (4)$$

such that it satisfies the interpolation condition

$$S(x_i) = u(x_i), \quad i = 1, 2, \dots, n. \tag{5}$$

where $\{\phi(|x - x_j|)\}_{j=1}^n$ are radial basis function about x_j 's. These n conditions lead to a linear system that is to be solved for λ_j 's.

Some of the well-known radial functions are Gaussian (GA - $e^{-(\epsilon r)^2}$), multi-quadratic (MQ - $\sqrt{1 + (\epsilon r)^2}$), thin plate splines (TPS - $r^2 \log r$), and so on, where $r = |x - x_j|$. It can be seen that GA and MQ depend on a parameter say ϵ , which determines the shape of these basis functions. They have significant effect on the accuracy, however, solutions become unstable as $\epsilon \rightarrow 0$. For many problems, these small values of ' ϵ ' provide better accuracy. Hence, it is important to stabilize the resultant linear system after discretization by RBFs. In the following section, we discuss the discretization of the problem (1) using Kansa's RBF collocation method and regularization of the resulting collocation system.

3 Methodology

3.1 A Fractional RBF Approximation

To derive the scheme consider the governing equation (1) along with the boundary conditions

$$u(a) = u_a, \quad u(b) = u_b. \tag{6}$$

Extension of the scheme to initial boundary value problems and higher dimensional problems is straightforward.

Assume that $u^*(x)$ represents the solution which can be expressed in terms of RBFs as follows:

$$u^*(x) = \sum_{j=1}^n \lambda_j \phi(|x - x_j|), \quad x \in [a, b]. \tag{7}$$

where $x_j, j = 1, 2, \dots, n$, are collocation points distributed in the given interval $[a, b]$. Assuming that ϕ is sufficiently smooth, (7) is substituted in both governing equation (1) and its boundary conditions (6) at each node $x_i, i = 1, 2, \dots, n$. These equations lead to an $n \times n$ linear system,

$$A\bar{\lambda} = F \tag{8}$$

The components of these matrices are:

$$A_{ij} = \begin{cases} \phi|x_i - x_j|, & \text{if } x_i = a \text{ or } x_i = b \\ ({}^c D_a^{\rho\alpha} \phi)(|x_i - x_j|) + \sum_{l=0}^{p-1} a_l(x) ({}^c D_a^{l\alpha} \phi)(|x_i - x_j|), & \text{if } x_i \in (a, b) \end{cases}$$

$$F_i = \begin{cases} u_a & \text{if } x_i = a \\ f(x_i), & \text{if } x_i \in (a, b) \\ u_b & \text{if } x_i = b \end{cases}$$

where $i, j = 1, 2, \dots, n$ and $\bar{\lambda} = [\lambda_1, \lambda_2, \dots, \lambda_n]^T$ is unknown vector to be determined. The fractional derivatives present in equations require the computation of integrals which are obtained by Gauss-Jacobi quadrature rule. The following subsection discusses Tikhonov regularization (TR) method and an effective approach to implement the algorithm to stabilize RBF system (8).

3.2 Tikhonov Regularization

Regularizations simply means replacing the current system with a nearby system that is less sensitive to perturbation and Tikhonov regularization (TR) [2] is one of the efficient regularization methods. In its simplest form, TR replaces the linear system (8) by a regularized system

$$(A^T A + \mu^2 I) \bar{\lambda} = A^T F \quad (9)$$

where μ^2 is known as regularization parameter which determines the amount of regularization. For a fixed $\mu^2 > 0$, linear system (9) has a unique solution

$$\bar{\lambda}_\mu = (A^T A + \mu^2 I)^{-1} A^T F \quad (10)$$

The solution, $\bar{\lambda}_\mu$, as in (10) satisfies the minimization problem

$$\min\{\|A\bar{\lambda} - F\|^2 + \mu^2 \|\bar{\lambda}\|^2\}, \quad (11)$$

where $\|\cdot\|$ denotes the Euclidean norm. It is important to choose μ carefully to make Tikhonov regularization effective. There are various algorithms such as L-curve and generalized cross-validation (GCV) [21] to efficiently find the value of μ . In the present work, generalized cross-validation method is extended to determine μ . GCV method estimates the optimal value of μ by minimizing the function

$$G(\mu) = \frac{\|A\bar{\lambda} - F\|^2}{(\text{trace}(I - AA^*))^2} \quad (12)$$

where $A^* = (A^T A + \mu^2 I)^{-1} A^T$.

3.3 Variable Shape Parameter

It has been shown by many works that to get stable and accurate results, variable shape parameters (VSP) can be used instead of constant shape parameter (CSP) [12, 26]. The VSP assigns different values to shape parameter corresponding to each collocation points. This results in distinct entries in collocation matrix which in turn reduces the ill-conditioning of the system and improves accuracy. There are many ways to generate variable values, either randomly or using exponential, linear, and trigonometric functions. In the present work we have considered trigonometric shape parameter, defined by $\varepsilon_j = \varepsilon_{min} + (\varepsilon_{max} - \varepsilon_{min})\sin(j)$, $j = 1, 2, \dots, n$ where ε_{min} and ε_{max} are, respectively, the minimum and maximum values for ε .

4 Numerical Illustrations

In this section, proposed algorithm is illustrated using Bagley-Torvik equation and fractional diffusion problems. Multiquadric RBF (MQ - $\sqrt{1 + \varepsilon^2(x - x_j)^2}$) is used in all these examples. Only 8 quadrature points are used in computation of fractional derivatives. The ill-conditioning issue is tackled by Tikhonov regularization. In trigonometric variable shape parameter, the values are chosen as $\varepsilon_{min} = 2E - 06$ and $\varepsilon_{max} - \varepsilon_{min} \leq 1$. The algorithm discussed by Hansen [15] for finding μ is extended appropriately to regularize fractional order RBF systems.

Example 1 Consider the Bagley-Torvik equation

$$aD^2u(x) + b {}_0^c D_x^{3/2}u(x) + cu(x) = c(1+x), \quad x \in [0, 1] \quad (13)$$

subject to the initial conditions $u(0) = u'(0) = 1$. The exact solution is $u(x) = 1 + x$. Consider $p = 4$, $\alpha = \frac{1}{2}$ and the values of a_l 's are b/a , 0 , 0 , c/a for $l = 0, 1, 2, 3$ respectively in Eqn. (1).

The results obtained are shown in Tables 1 & 2 and Figure 1. Table 1 provides a comparison between triangular function (TF) [5] approach with RBFs and proposed Kansa's method with regularization (KMR) for $a = 1$, $b = c = 0.5$. The solution obtained using KMR with less number of nodes yields same accuracy as that by TF method. In another work by Kazemi and Ghoreishi [7], they have obtained MQ-solutions for $\varepsilon = 1$, but 100-digits precision is used to deal with ill-conditioning. However, this can slow down the algorithm. The proposed solutions are compared with [7] in Table 2. We have obtained solutions for much smaller ε_j 's, whereas $\varepsilon = 1$ for the computations in [7]. MQ-solutions in [7] fail to converge when $n \geq 13$ even for 30-digits precision and in case of $n \geq 18$ even for 40-digits precision. However, with regularization, present Kansa's method provides stable results for large n with default machine precision.

Table 1 Comparison of L_∞ error by TF and KMR for Example 1.

TF [5]			KMR		
h	L_∞ error	CPU time(s)	h	L_∞ error	CPU time(s)
$\frac{1}{100}$	2.86E-13	1.0920	$\frac{1}{50}$	1.31E-13	0.5983

Table 2 Comparison of L_∞ error by MQ RBF [7] with 100-digits precision and KMR with default machine precision for Example 1.

n	L_∞ error	
	MQ RBF [7]	KMR
6	1.97E-04	1.25E-06
8	1.17E-05	1.73E-09
10	1.60E-06	3.52E-11
12	1.48E-07	3.64E-12
20	1.17E-11	9.21E-13
27	7.59E-15	8.21E-13
34	8.83E-19	1.65E-12

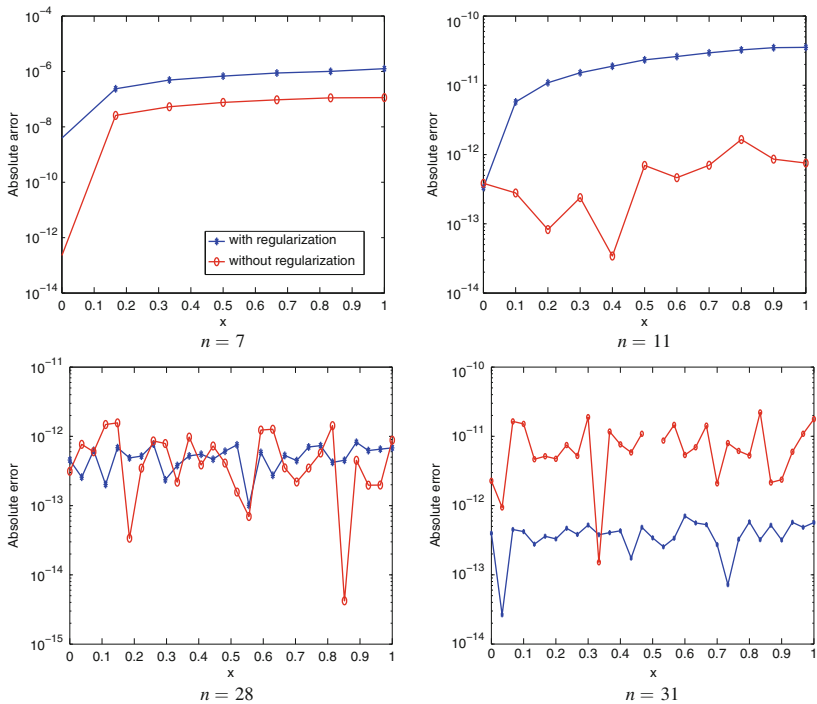


Fig. 1 The graph of absolute errors for various n for Example 1 with regularization (blue) and without regularization (red).

Figure 1 illustrates that for small ε_j 's, Kansa's method produces good accuracy without regularization than the method used along with regularization (i.e., KMR). But as n increases, accuracy using both the methods coincides. On further increment of n , KMR yields good accuracy than the Kansa's method without regularization.

Table 3 Comparison of L_∞ error by KM and KMR at $t = 1$ for Example 2.

$h = \tau$	KM		KMR	
	L_∞ error	Error rate	L_∞ error	Error rate
1/10	3.70E-04		1.15E-04	
1/15	5.65E-05	4.63	5.06E-05	2.02
1/20	1.52E-05	4.57	2.86E-05	1.99
1/25	1.36E-05	0.51	1.83E-05	2.00
1/30	1.07E-05	1.30	1.27E-05	2.01
1/35	1.63E-02	-47.55	9.33E-06	1.99

Traditional diffusion represents the long-time of a random walk, where finite variance jumps occur at regularly spaced intervals. But in many real-world applications, particles follow heavy tails, sharp peaks, etc., which can be expressed as a fractional diffusion equation. Our proposed KMR method can be used successfully in obtaining solution for fractional diffusion equations. Two examples are illustrated below. $p = 3, \alpha = \frac{3}{5}$ and $a_l = 0, l = 0, 1, 2$ have been substituted in Eqn. (1) to get the diffusion equations as in Example 2 and 3. For both the examples, time derivative is approximated by Crank-Nicolson scheme and space derivative is by Kansa’s method.

Example 2 Consider the space fractional diffusion equation

$$\frac{\partial u(x, t)}{\partial t} = \frac{\Gamma(2.2)}{6} x^{2.8} {}_0^c D_x^{1.8} u(x, t) - (1 + x)x^3 e^{(-t)}, \quad x \in (0, 1) \text{ and } t > 0 \quad (14)$$

with initial condition $u(x, 0) = x^3$ and boundary conditions $u(0, t) = 0$ and $u(1, t) = e^{(-t)}$. The exact solution is $u(x, t) = e^{(-t)}x^3$.

Results are displayed in Table 3 and Figure 2. Table 3 shows that the L_∞ error and rate of convergence by Kansa’s method without regularization(KM) and with regularization (KMR) at $t = 1$ for uniform nodal distribution. KMR could produce the displayed accuracy with 8 quadrature points and very small ϵ_j ’s whereas KM used 10 quadrature points with $\epsilon = n^{0.5}/2$. It can be observed that the rate of convergence is decreasing as h (steplength for x_i ’s) or τ (steplength for t_i ’s) is decreasing in KM but it remains same in KMR method, i.e., with regularization, Kansa’s method becomes more stable and achieves better solution.

Figure 2 illustrates the similar relationship between KMR method and Kansa’s method without regularization as explained in Example 1.

Example 3 Consider the space fractional diffusion problem

$$\frac{\partial u(x, t)}{\partial t} = \Gamma(1.2)x^{1.8} {}_0^c D_x^{1.8} u(x, t) + 3x^2(2x - 1)e^{(-t)}, \quad x \in (0, 1) \text{ and } t > 0 \quad (15)$$

with initial condition $u(x, 0) = x^2(1 - x), u(0, t) = u(1, t) = e^{(-t)}$. The exact solution is $u(x, t) = x^2(1 - x)e^{(-t)}$.

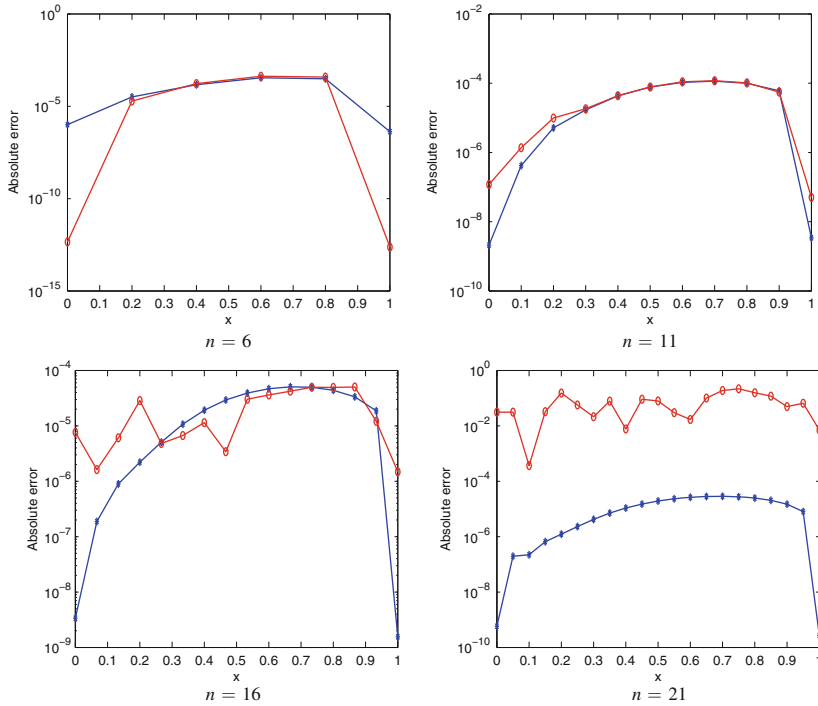


Fig. 2 Absolute error of Example 2 with regularization (blue) and without regularization (red) at $t = 1$.

Table 4 Comparison of absolute error by shifted Chebyshev polynomials (SCP) [28] and KMR method at $t = 1$ and 2 for Example 3.

x	t = 1		t = 2	
	SCP [28]	KMR	SCP [28]	KMR
0.1	5.46E-06	3.08E-07	3.33E-06	2.88E-07
0.2	8.51E-06	6.44E-07	5.65E-06	4.20E-07
0.3	9.60E-06	8.46E-07	7.05E-06	4.56E-07
0.4	9.18E-06	9.14E-07	7.64E-06	4.36E-07
0.5	7.69E-06	8.79E-07	7.52E-06	3.81E-07
0.6	5.60E-06	7.69E-07	6.80E-06	3.08E-07
0.7	3.33E-06	6.04E-07	5.59E-06	2.26E-07
0.8	1.34E-06	4.08E-07	3.98E-06	1.43E-07
0.9	8.39E-08	2.01E-07	2.08E-06	6.54E-08

Tables 4 and 5 show that the accuracy of the KMR method is better than the shifted Chebyshev polynomials (SCP) method [28] for different values of t . The absolute error by KMR method at each point is less than the SCP method. Table 6 displays how error behaves using Kansa’s method with and without regularization

Table 5 Comparison of absolute error by shifted Chebyshev polynomials (SCP) [28] and KMR at $t = 10$ for Example 3.

x	SCP [28]	KMR
0.2	2.34E-08	2.52E-09
0.4	4.78E-09	1.51E-09
0.6	7.39E-09	6.81E-10
0.8	2.84E-08	1.24E-10

Table 6 Comparison of L_∞ error for Kansa’s method: 1) with regularization (KMR) and 2) without regularization (KM) at $t = 1$, $t = 2$, and $t = 10$ for Example 3.

n	t = 1		t = 2		t = 10	
	KMR	KM	KMR	KM	KMR	KM
6	2.30E-04	8.93E-05	2.37E-04	1.62E-04	9.61E-04	5.22E-04
9	3.48E-05	3.52E-05	6.71E-05	6.84E-05	1.09E-04	1.07E-04
11	2.27E-05	1.00E-03	4.46E-05	1.04E-04	3.05E-05	2.66E-05
16	1.01E-05	6.41E-04	1.99E-05	3.55E-02	3.62E-06	2.32E+02
21	5.69E-06	3.98E+07	1.13E-05	3.28E+05	7.16E-07	9.67E+24
26	3.65E-06	5.32E+37	7.22E-06	2.06E+11	1.73E-07	1.31E+06

for various values of t . It can be seen that as n is increased, Kansa’s method can be very unstable and produce unreliable results (see $n = 21, 26$). The solution becomes accurate using appropriate regularization.

5 Conclusions

The aim of the present work is to establish the effect of regularization on RBF collocation system obtained using Kansa’s method. All results show that the Kansa’s method with regularization (KMR) yields stable and accurate results for different types of fractional differential equations. Since Kansa’s method produces ill-conditioned asymmetric collocation matrix, Tikhonov regularization was used to tackle the ill-conditioned issue in which GCV method is used to obtain the regularization parameter.

References

1. M. Arghand and M. Amirfakhrian: A meshless method based on the fundamental solution and radial basis function for solving an inverse heat conduction problem. *Adv. Math. Phys.*, pages Art. ID 256726, **8**, (2015).
2. A. Ben-Israel and T. N. Greville: *Generalized Inverses: Theory and Applications*, volume 15. Springer Science & Business Media, (2003).
3. J. P. Boyd and K. W. Gildersleeve: Numerical experiments on the condition number of the interpolation matrices for radial basis functions. *Appl. Numer. Math.*, **61(4)**, 443–459, (2011).

4. M. D. Buhmann: *Radial basis functions: theory and implementations*. Cambridge monographs on applied and computational mathematics. Cambridge University Press, Cambridge, New York, (2003).
5. S. K. Damarla and M. Kundu: Numerical solution of multi-order fractional differential equations using generalized triangular function operational matrices. *Appl. Math. Comput.*, **263**, 189–203, (2015).
6. M. Di Paola, F. P. Pinnola, and M. Zingales: Fractional differential equations and related exact mechanical models. *Comput. Math. Appl.*, **66**(5), 608–620, (2013).
7. B. Fakhr Kazemi and F. Ghoreishi: Error estimate in fractional differential equations using multiquadratic radial basis functions. *J. Comput. Appl. Math.*, **245**, 133–147, (2013).
8. G. E. Fasshauer and M. J. McCourt: Stable evaluation of Gaussian radial basis function interpolants. *SIAM J. Sci. Comput.*, **34**(2), A737–A762, (2012).
9. B. Fornberg, E. Larsson, and N. Flyer: Stable computations with Gaussian radial basis functions. *SIAM J. Sci. Comput.*, **33**(2), 869–892, (2011).
10. B. Fornberg and C. Piret: A stable algorithm for flat radial basis functions on a sphere. *SIAM J. Sci. Comput.*, **30**(1), 60–80, (2007/08).
11. B. Fornberg and G. Wright: Stable computation of multiquadric interpolants for all values of the shape parameter. *Comput. Math. Appl.*, **48**(5–6), 853–867, (2004).
12. A. Golbabai, E. Mohebianfar, and H. Rabiei: On the new variable shape parameter strategies for radial basis functions. *Comput. Appl. Math.*, **34**(2), 691–704, (2015).
13. J. F. Gómez-Aguilar, H. Yépez-Martí nez, R. F. Escobar-Jiménez, C. M. Astorga-Zaragoza, and J. Reyes-Reyes: Analytical and numerical solutions of electrical circuits described by fractional derivatives. *Appl. Math. Model.*, **40**(21–22), 9079–9094, (2016).
14. P. Gonzalez-Rodriguez, V. Bayona, M. Moscoso, and M. Kindelan: Laurent series based RBF-FD method to avoid ill-conditioning. *Eng. Anal. Bound. Elem.*, **52**, 24–31, (2015).
15. P. C. Hansen: Regularization Tools version 4.0 for Matlab 7.3. *Numer. Algorithms*, **46**(2), 189–194, (2007).
16. J.-H. He: Application of homotopy perturbation method to nonlinear wave equations. *Chaos. Soliton. Fract.*, **26**(3), 695–700, (2005).
17. X. Hei, W. Chen, G. Pang, R. Xiao, and C. Zhang: A new visco–elasto-plastic model via time–space fractional derivative. *Mech. Time-Depend. Mater.*, pages 1–13, (2017).
18. E. J. Kansa: Multiquadrics—a scattered data approximation scheme with applications to computational fluid-dynamics—II. Solutions to parabolic, hyperbolic and elliptic partial differential equations. *Comput. Math. Appl.*, **19**(8), 147–161, (1990).
19. E. Larsson, E. Lehto, A. Heryudono, and B. Fornberg: Stable computation of differentiation matrices and scattered node stencils based on Gaussian radial basis functions. *SIAM J. Sci. Comput.*, **35**(4), A2096–A2119, (2013).
20. J. Lin, W. Chen, and K. Sze: A new radial basis function for Helmholtz problems. *Eng. Anal. Bound. Elem.*, **36**(12), 1923–1930, (2012).
21. J. Lin, W. Chen, and F. Wang: A new investigation into regularization techniques for the method of fundamental solutions. *Math. Comput. Simulation*, **81**(6), 1144–1152, (2011).
22. S. Momani and Z. Odibat: Numerical approach to differential equations of fractional order. *J. Comput. Appl. Math.*, **207**(1), 96–110, (2007).
23. M. D. Paola and M. Zingales: Exact mechanical models of fractional hereditary materials. *J. Rheol.*, **56**(5), 983–1004, (2012).
24. I. Podlubny: *Fractional differential equations: An introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications*, volume 198. Academic Press, (1998).
25. F. A. Rihan: Numerical modeling of fractional-order biological systems. *Abstr. Appl. Anal.*, pages Art. ID 816803, **11**, (2013).
26. S. A. Sarra and D. Sturgill: A random variable shape parameter strategy for radial basis function approximation methods. *Eng. Anal. Bound. Elem.*, **33**(11), 1239–1245, (2009).

27. R. Schaback: Comparison of radial basis function interpolants. In *Multivariate approximation: from CAGD to wavelets (Santiago, 1992)*, volume 3 of *Ser. Approx. Decompos.*, pages 293–305. World Sci. Publ., River Edge, NJ, (1993).
28. N. H. Sweilam, A. M. Nagy, and A. A. El-Sayed: Second kind shifted Chebyshev polynomials for solving space fractional order diffusion equation. *Chaos Solitons Fractals*, **73**, 141–147, (2015).
29. V. Turut and N. Güzel: On solving partial differential equations of fractional order by using the variational iteration method and multivariate padé approximations. *Eur. J. Pure Appl. Math.*, **6(2)**, 147–171, (2013).
30. Y.-F. Zhang and C.-J. Li: A Gaussian RBFs method with regularization for the numerical solution of inverse heat conduction problems. *Inverse Probl. Sci. Eng.*, **24(9)**, 1606–1646, (2016).

The Effect of Toxin and Human Impact on Marine Ecosystem



S. Chakraborty and S. Pal

1 Introduction

Microscopic plankton are a various group of organisms which inhabit the surface waters of oceans, rivers, lakes, ponds, marsh, etc. Aquatic food chain is based on the plankton population. One noticeable characteristic of phytoplankton populations is the rapid rate of increase or bloom in a water system. Usually, nitrogen and phosphorus are two most important nutrients for which phytoplankton grows [1]. Iron, zinc, and manganese are essential nutrients too. In the absence of any of the necessary nutrients, the growth rate can be limited or nonexistent for phytoplankton but the set of favorable conditions can promote the growth rate and reproductive ability (sexual/asexual) of phytoplankton population. Since zooplankton are largely depended on phytoplankton population, hence it can set out large rise in the population numbers simultaneously as the phytoplankton population [2–5]. All blooms are not harmful but harmful blooms can be unfavorable for marine organisms.

Many mathematical models of nutrient-phytoplankton-zooplankton(N-P-Z) interaction with several complexity have been explored and analyzed by researchers [6, 7, 9–11, 14–16, 19, 21, 25, 28, 29].

The ecological system is often deeply perturbed by the exploiting activities of humankind. Some phytoplankton such as nori, kelp, and eucheuma and some zooplankton such as jellyfish, krill, and acetes are harvested for food [20, 22, 26, 27]. Pollution emissions in air and water are produced as a result of human activities which damage the plankton drifting or floating in the ocean. The problem of our fossil fuel age is that huge percentage of extra carbon dioxide (CO_2) has been accumulated in the Earth's atmosphere in the last century. Leakage from storage

S. Chakraborty · S. Pal (✉)

Department of Mathematics, University of Kalyani, Kalyani-741235, India

e-mail: chakrabortyuman07@gmail.com; samaresp@yahoo.co.in

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High*

Performance Computing, Advances in Mechanics and Mathematics 41,

https://doi.org/10.1007/978-3-030-02487-1_6

tanks and pipelines and seepage from waste drums, and dumped nonrecyclable trash and chemicals are causing pollution in sea surface. Turbidity, or the presence of suspended particles in the water, and the thick duvets of CO_2 are nudging the planet to an ocean warming. Generally, phytoplankton production will increase with the temperature, up to an optimum temperature range. But due to excessive heat, this optimum temperature can be exceeded for which growth rate of phytoplankton may in turn be reduced and phytoplankton may drop in population numbers. Relatively, the bigger issue than global warming is that CO_2 nurtures grasses on land, and those grasses interfere with the dust in the wind. Grass loses its water through respiration and breathing. Grasses have to place its wet membranes open to air in order to take up CO_2 and give up oxygen(O_2). Due to high concentration of CO_2 , grasses do not need frequent breathing to obtain the CO_2 they need, hence their wet membranes evaporate less water and keep them green and growing [17]. Green grass is good ground cover, so less dust blows in the wind. The dust in the wind from land carries iron and other vital mineral nutrients into the ocean that result in phytoplankton blooms. Due to low dust deposition flux into the ocean, the growth of phytoplankton becomes severely affected. An immense effect of water pollution is organic and inorganic wastes in sewage water which cause changes in the physicochemical parameters of the water. Fertilizer run-off from agricultural lands causes the abundance of nitrogen and phosphorous into water system. The presence of sewage-derived inorganic nutrients works as additional food for phytoplankton which results in phytoplankton blooms [8, 13, 30]. The discharge of sewage in the water is detrimental to zooplankton growth. Higher levels of calcium (Ca), magnesium (Mg), chlorine gas(Cl_2) due to the continuous influx of contaminated domestic sewage have adverse effects on zooplankton population. The above discussion clearly indicates that exploitation activities and pollution emissions have a great impact upon nutrient-phytoplankton-zooplankton dynamics and to be considered in N-P-Z models.

Toxic substances produced by toxin-producing phytoplankton (TPP) play an important role in plankton dynamics and cannot be underestimated. Some phytoplankton species have toxin-producing ability which is unfavorable for the usual growth of algae and zooplankton [12, 18, 23, 24]. TPP reduces the growth rate and catalyzes the essential mortality of zooplankton by releasing toxic chemicals. Different experiments, field observations, and mathematical models have established that optimal requirements of environmental conditions, environmental stress factors, nutrient-limited conditions, etc. are the main reasons or causes for toxin liberation and TPP acts as a monitoring performer to accomplish plankton bloom.

2 The Mathematical Model 1

We consider a nutrient-phytoplankton-zooplankton model in which nutrient recycling, harvesting of phytoplankton and zooplankton, and negative effects of toxic chemicals upon zooplankton population are taken under consideration. We assume that seven basic processes govern the ecosystem: (1) the nutrient uptake of

phytoplankton, (2) the zooplankton grazing on phytoplankton, (3) the death of phytoplankton, (4) the death of zooplankton, (5) harvesting of both plankton, (6) nutrient recycling, and (7) effect of toxic chemicals on zooplankton. We have considered that nutrient concentration at time t is $s(t)$ and M is the initial concentration of nutrient. Phytoplankton and zooplankton concentrations are $p(t)$ and $z(t)$, respectively, at time t . It is assumed that phytoplankton consume nutrient and zooplankton prey on phytoplankton. Phytoplankton death rate is considered as natural death rate but zooplankton death rate includes its natural death rate and death rate by predation.

We make the following assumptions in formulating the mathematical model:

- (i) Nutrient returns to the dissolved nutrient field after decomposition when a plankton dies. Here, nutrient growth rate is $M - \alpha s + \eta p(t) + \delta z(t)$, when nutrient consumption does not take place.
- (ii) Phytoplankton growth rate is proportional to their population size and to a Holling-type II function of available nutrient $f_1(s) = \frac{m_1 s}{a_1 + s}$.
- (iii) Zooplankton growth rate is proportional to their population size and to a Holling-type II function of available phytoplankton $f_2(p) = \frac{m_2 p}{a_2 + p}$.
- (iv) Harvest rates are proportional to plankton population, defined as: $c_1 h_1 p$ and $c_2 h_2 z$, consecutively for phytoplankton and zooplankton.
- (v) Death rates of plankton are proportional to their population sizes.
- (vi) The rate of toxin release is proportional to zooplankton population and to a Holling-type II function of phytoplankton population $f_3(p) = \frac{p}{a_3 + p}$.

The formulated model by means of differential equation (*Model 1*) is

$$\begin{aligned}
 \frac{ds}{dt} &= M - \alpha s + \gamma_1 p + \gamma_2 z - \frac{m_1 s p}{a_1 + s} \equiv F_1(s, p, z) \\
 \frac{dp}{dt} &= \frac{m_1 s p}{a_1 + s} - \frac{m_2 z p}{a_2 + p} - \beta_1 p - c_1 h_1 p \equiv F_2(s, p, z) \\
 \frac{dz}{dt} &= \frac{m_2 z p}{a_2 + p} - \beta_2 z - \frac{\mu p z}{a_3 + p} - c_2 h_2 z \equiv F_3(s, p, z).
 \end{aligned} \tag{1}$$

System(1) is to be analyzed with initial conditions $s(0) \geq 0, p(0) = p_0 \geq 0, z(0) = z_0 \geq 0$. We consider that all parameters are nonnegative constant. Parameters are interpreted as follows:

- M —constant input nutrient concentration.
- α —rate of nutrient loss.
- β_1 —phytoplankton mortality rate.
- β_2 —zooplankton mortality rate.
- γ_1 —nutrient recycling rate after death of phytoplankton ($\gamma_1 \leq \beta_1$).
- γ_2 —nutrient recycling rate after death of zooplankton ($\gamma_2 \leq \beta_2$).
- m_1 —maximal nutrient uptake rate for the phytoplankton.
- m_2 —maximal zooplankton ingestion rate.
- a_1 —half-saturation constant or Michaelis-Menten constant for Nutrient.

- a_2 —half-saturation constant or Michaelis-Menten constant for phytoplankton.
 h_1 —harvest rate of phytoplankton.
 h_2 —harvest rate of zooplankton.
 μ —rate of toxin production.
 c_1 —catchability coefficient of phytoplankton.
 c_2 —catchability coefficient of zooplankton.
 a_3 —half-saturation constant or Michaelis-Menten constant for zooplankton.

3 Some Preliminary Results

3.1 Boundedness of the System

Theorem 1 *All the solutions of system 1 are ultimately bounded.*

Proof We define a function:

$$w = s + p + z$$

$$\frac{dw}{dt} = \frac{ds}{dt} + \frac{dp}{dt} + \frac{dz}{dt}$$

$$= M - [\alpha s + (\beta_1 - \gamma_1 + c_1 h_1)p + (\beta_2 - \gamma_2 + c_2 h_2)z + \frac{\mu p z}{a_3 + p}]$$

$$\leq M - [\alpha s + (\beta_1 - \gamma_1 + c_1 h_1)p + (\beta_2 - \gamma_2 + c_2 h_2)z]$$

(because $\frac{\mu p z}{a_3 + p}$ is a positive term.)

$$\leq M - \sigma(s + p + z)$$

where $\sigma = \min(\alpha, (\beta_1 - \gamma_1 + c_1 h_1), (\beta_2 - \gamma_2 + c_2 h_2))$

$$= M - \sigma w$$

Hence, $\frac{dw}{dt} + \sigma w \leq M$

Using the variation of transformation formula, the above inequality is transformed into:

$$w(s(t), p(t), z(t)) \leq \frac{M}{\sigma}(1 - e^{-\sigma t}) + w(s(0), p(0), z(0))e^{-\sigma t}$$

For large values of t , we get that $\limsup_{t \rightarrow \infty} [s(t) + p(t) + z(t)] \leq \frac{M}{\sigma}$.

Hence, we can conclude that all solutions of the system are bounded.

3.2 Equilibria

The system 1 possesses the following three equilibria:

The plankton-free equilibrium $E_0 = (\frac{M}{\alpha}, 0, 0)$, and the zooplankton-free equilibrium $E_1(s_1, p_1, 0)$ where $s_1 = \frac{a_1(\beta_1 + c_1 h_1)}{m_1 - (\beta_1 + c_1 h_1)}$ and $p_1 = \frac{M - \alpha s_1}{\beta_1 - \gamma_1 + c_1 h_1}$. It is clear that $m_1 > \beta_1 + c_1 h_1$ and $\beta_1 + c_1 h_1 > \gamma_1$, otherwise E_1 does not exist. If $M = \alpha s_1$, then E_1 switches to E_0 . There is no steady state of the form $E_2 = (s_2, 0, z_2)$. Hence,

we investigate the interior steady state $E^* = (s^*, p^*, z^*)$. From the equilibrium equation, we can see that p^* must satisfy

$$(m_2 - \beta_5 - \mu)p^2 + (m_2a_3 - a_2\beta_5 - \beta_5a_3 - a_2\mu)p - a_2c_3\beta_5 = 0.$$

$$\text{or, } (\beta_5 + \mu - m_2)p^2 + (a_2\beta_5 + \beta_5a_3 + a_2\mu - m_2a_3)p + a_2a_3\beta_5 = 0.$$

(where $\beta_5 = \beta_2 + c_2h_2$).

This equation has no positive solution if $\beta_5 - m_2 > 0$ and if $\beta_5 + \mu - m_2 < 0$, then we get a unique positive value of p^* . However, if $\beta_5 + \mu - m_2 > 0$ and $a_2\beta_5 + \beta_5a_3 + a_2\mu - m_2a_3 < 0$, then either the equation has two positive solutions or no positive solutions.

Here, z can be written in terms of s as:

$$\left(\frac{m_1s}{a_1+s} - \beta_1 - c_1h_1\right)\left(\frac{a_2+p}{m_2}\right) = f_1(s)$$

where p is a positive root of the above equation and s component of the interior steady state, s^* satisfies

$$M - \alpha s + \gamma_1 p + \gamma_2 f_1(s) - \frac{m_1 s p}{a_1 + s} = 0.$$

$$E^* \text{ does not exist if } \frac{m_1 s}{a_1 + s} < \beta_1 + c_1 h_1.$$

3.3 Eigenvalue Analysis

In this section, local stability analysis of the system around the biological feasible equilibria is performed. The central aim of the present analysis is to find out suitable mechanism to explain the planktonic blooms and a possible solution to control it.

Lemma 1 *If $\gamma_1 - \frac{m_1 s}{a_1 + s} < 0$, then the plankton-free steady state E_0 of the system 1 is locally asymptotically stable.*

Lemma 2 *The planar equilibrium E_1 of the system 1 is locally asymptotically stable if $\frac{m_2 p}{a_2 + p} - \frac{\mu p}{a_3 + p} - \beta_2 - c_2 h_2 < 0$.*

Next, we study the equilibrium of the unique positive steady state E^* .

Lemma 3 *E^* is locally asymptotically stable if:*

- (a) $A_1 > 0$
- (b) $A_3 > 0$
- (c) $A_1 A_2 - A_3 > 0$.

where

$$A_1 = \alpha + \frac{m_1 a_1 p^*}{(a_1 + s^*)^2} + \frac{m_2 a_2 z^*}{(a_2 + p^*)^2} - \frac{m_2 z^*}{a_2 + p^*}$$

$$A_2 = \frac{m_2^2 a_2 p^* z^*}{(a_2 + p^*)^3} + \frac{m_1^2 a_1 s^* p^*}{(a_1 + s^*)^3} - \frac{m_2 a_3 \mu p^* z^*}{(a_2 + p^*)(a_3 + p^*)^2} - \frac{m_2 \alpha p^* z^*}{(a_2 + p^*)^2}$$

$$- \frac{m_1 m_2 a_1 p^2 z^*}{(a_1 + s^*)^2 (a_2 + p^*)^2} - \frac{\gamma_1 m_1 a_1 p^*}{(a_1 + s^*)^2},$$

$$A_3 = \frac{m_2^2 \alpha a_2 p^* z^*}{(a_2 + p^*)^3} + \frac{m_1 m_2^2 a_1 a_2 p^2 z^*}{(a_1 + s^*)^2 (a_2 + p^*)^3} - \frac{m_1 m_2 a_1 a_2 \gamma_2 p^* z^*}{(a_1 + s^*)^2 (a_2 + p^*)^2} - \frac{m_2 \alpha \mu a_3 p^* z^*}{(a_2 + p^*) (a_3 + p^*)^2} - \frac{m_1 m_2 a_1 p^2 \mu a_3 z^*}{(a_1 + s^*)^2 (a_2 + p^*) (a_3 + p^*)^2} + \frac{\gamma_2 m_1 a_1 \mu a_3 p^* z^*}{(a_3 + p^*)^2 (a_1 + s^*)^2}.$$

Proof See Appendix.

3.4 Biological Interpretation

3.4.1 Extinction of Both Plankton

Plankton will be extinct if the maximal growth rate of phytoplankton population $\frac{m_1 s^1}{a_1 + s^1}$ is less than β_1 where β_1 is the mortality rate of phytoplankton.

3.4.2 Extinction of Zooplankton

It will happen if:

- (i) $\frac{m_2 p}{a_2 + p} < \beta_2 + c_2 h_2 + \frac{\mu p}{a_3 + p}$ where $\frac{m_2 p}{a_2 + p}$ is the maximal growth rate of zooplankton population and $\beta_2 + c_2 h_2 + \frac{\mu p}{a_3 + p}$ is the total mortality rate of zooplankton.
- (ii) The natural mortality rate of phytoplankton population (β_1) is less than its maximal nutrient uptake rate (m_1).
- (iii) Constant input nutrient concentration must be greater than some threshold value.

3.4.3 Coexistence of Both Plankton

The coexistence of both of the plankton population in an interior steady state happens if:

- (i) The maximal growth rate of phytoplankton population ($\frac{m_1 s}{a_1 + s}$) is greater than its natural mortality rate and harvest rate ($\beta_1 + c_1 h_1$).
- (ii) The growth rate of zooplankton population due to phytoplankton ingestion (m_2) is greater than total mortality rate of zooplankton ($\beta_2 + c_2 h_2 + \mu$).
- (iii) The unique positive interior state E^* has to be locally asymptotically stable.

The most important thing is that the stability of E_0 strictly denies the existence of E^* .

4 Numerical Simulations

We have numerically simulated *system 1* for a range of parameter values around it to study the dynamics of the system around the positive interior state. We have used the following set of parameter values— $M = 0.01, a_1 = 0.2, a_2 = 0.2, a_3 = 0.2, \alpha = 0.01, \beta_1 = 0.21, \beta_2 = 0.1, \gamma_1 = 0.1, \gamma_2 = 0.06, m_1 = 0.6, m_2 = 0.6, \mu = 0.03, c_1 = 0.01, c_2 = 0.01, h_1 = 0.01,$ and $h_2 = 0.01$.

We find the existence of interior steady states of system (1) for $a_1 = 0.2, a_2 = 0.1, a_3 = 0.25,$ and $\mu = 0.6$ where the remaining parameter values are the same as previous. We find the interior steady state (0.1822, 0.0501, 0.0190). Calculating the eigen values, we get that it has one real (-0.0147) and a pair of complex conjugate eigen values (-0.0055 ± 0.0998i) and all eigen values have negative real part which means that it is a stable focus node.

Toxic chemicals into the water act as a biological control which terminate planktonic bloom. The study is focused on how planktonic bloom occurs and then terminates for small changes of values of μ . If toxin production by TPP is very low ($\mu = 0.025$), periodic solutions exist which depicts annual bloom (Figure 1). As the rate of toxin production rises, the bloom height gradually decreases but when the toxin production rate exceeds some critical value ($\mu = 0.06$), the periodic solution disappears and the solution becomes stable (Figure 2). Major increase in toxin production rate ($\mu = 0.25$) results in the extinction of zooplankton population (Figure 3).

Fig. 1 The figures depicts oscillatory behavior of nutrients, phytoplankton, and zooplankton population of system (1) for $\mu = 0.025, \alpha = 0.01$.

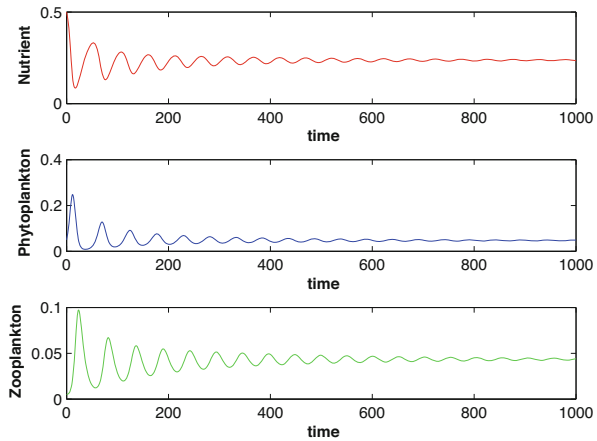


Fig. 2 The stable dynamics of system (1) for $\alpha = 0.01, \mu = 0.06$ with other parameters same as given in Figure 1.

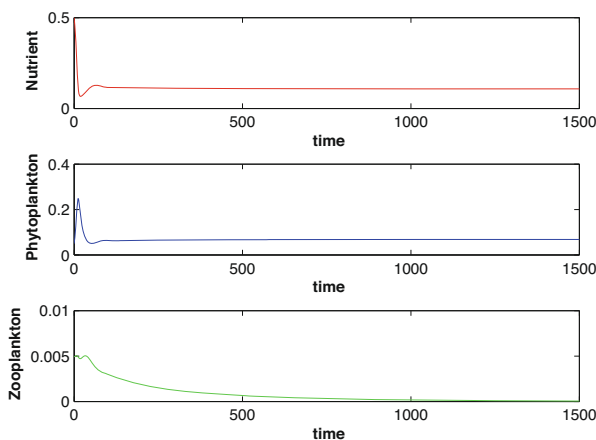
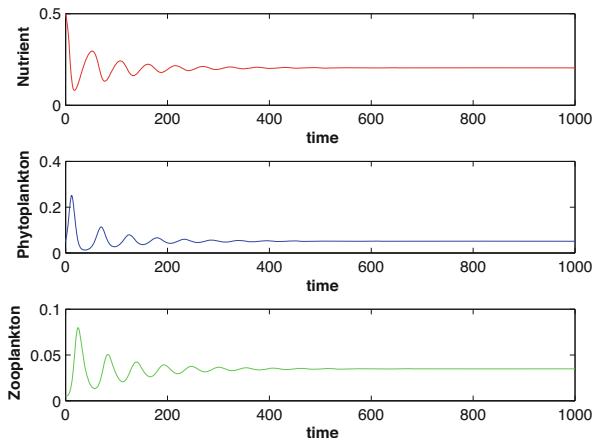


Fig. 3 Extinction of zooplankton population at $\alpha = 0.01, \mu = 0.25$ of system 1, other parametric values remained unchanged

5 Human Impact on Marine Ecosystem

5.1 Mathematical Model 2

Generally, phytoplankton needs both sunlight and nutrients to grow at a high rate. We considered the growth based on nutrient. Since, light is readily available upon water surface and accumulated nutrients distribute additional food for which phytoplankton grows in an exponential way. Favorable conditions allow zooplankton to grow faster. True blooms of zooplankton are typically the result of high food concentration and some other factors. Hence, we modify our model by including the effect of light and additional food for phytoplankton and zooplankton.

Light intensity varies with depth. Suppose, I_h is the light intensity at depth h . Then, $I_h = I_0 e^{-\kappa h}$ where:

I_0 —initial light intensity.

h —depth

κ —attenuation coefficient.

Taking an average, we find:

$$I = \frac{I_0 \kappa (1 - e^{-\kappa h})}{\kappa h}$$

I = light intensity.

$$\frac{ds}{dt} = M_1 - \alpha s + \gamma_1 p + \gamma_2 z - \frac{m_1 s p}{a_1 + s} \equiv G_1(s, p, z)$$

$$\frac{dp}{dt} = r_1 p + \frac{I_0 \kappa (1 - e^{-\kappa h})}{\kappa h} \frac{m_1 s p}{a_1 + s} - \frac{m_2 z p}{a_2 + p} - \beta_1 p - c_1 h_1 p \equiv G_2(s, p, z)$$

$$\frac{dz}{dt} = r_2 z + \frac{m_2 z p}{a_2 + p} - \beta_2 z - \frac{\mu p z}{a_3 + p} - c_2 h_2 z \equiv G_3(s, p, z). \tag{2}$$

Where:

r_1 —intrinsic growth rate of phytoplankton.

r_2 —intrinsic growth rate of zooplankton.

System 2 has three equilibrium points $E_0^{(1)}$, E_1^1 , and $E^{*(1)}$. Where:

$$E_0^{(1)} = (M_1/\alpha, 0, 0), E_1^{(1)} = \left(\frac{a_1(\beta_1 + c_1 h_1 - r_1)}{\phi m_1 - (\beta_1 + c_1 h_1 - r_1)}, \frac{\phi(M_1 - \alpha s_1)}{(\beta_1 + c_1 h_1 - r_1 - \gamma_1)}, 0 \right).$$

Here, $\phi m_1 > \beta_1 + c_1 h_1 - r_1$ and $\beta_1 + c_1 h_1 - r_1 > \gamma_1$, otherwise E_1^* does not exist. If $M_1 = \alpha s_1$, then E_1^* switches to E_0^* .

$$\phi = \frac{I_0 \kappa (1 - e^{-\kappa h})}{\kappa h}.$$

Investigating the interior steady states $E^{(1)*} = (s^{*(1)}, p^{*(1)}, z^{*(1)})$, we can say $p^{*(1)}$ must satisfy the equation:

$$(\beta_5 + \mu - m_2 - r_2) p^2 + (a_2 \beta_5 + \beta_5 a_3 + a_2 \mu - m_2 a_3 - r_2 a_2 - r_2 a_3) p + a_2 a_3 (\beta_5 - r_2) = 0.$$

(where $\beta_5 = \beta_2 + c_2 h_2$)

This equation has no positive solution if $\beta_5 + \mu - m_2 - r_2 > 0$. If $\beta_5 + \mu - m_2 - r_2 < 0$, then we get a unique positive value of $p^{*(1)}$. However, if $\beta_5 + \mu - m_2 - r_2 > 0$ and

$a_2\beta_5 + a_3\beta_5 + a_2\mu - m_2a_3 - r_2a_2 - r_2a_3 < 0$, then either the equation has two positive solutions or no positive solutions.

z can be written in terms of s as:

$$(r_1 + \phi \frac{m_1s}{a_1 + s} - \beta_1 - c_1h_1)(\frac{a_2 + p}{m_2}) = f_2(s)$$

where p is a positive root of the above equation and s component of the interior steady state, $s^{*(1)}$ satisfies

$$M_1 - \alpha s + \gamma_1 p + \gamma_2 f_2(s) - \frac{m_1sp}{a_1+s} = 0.$$

$E^{*(1)}$ does not exist if:

$$\phi \frac{m_1s}{a_1 + s} < \beta_1 + c_1h_1 - r_1.$$

Due to adding new variables, the equations of system (1) are slightly changed. But, the local stability analysis process of system (2) around the biological feasible equilibria is quite the same as the system (1).

5.2 Mathematical Model 3

In the introduction section, we discussed how CO_2 and other heat-consuming substances in the atmosphere are affecting the plankton population. The thick blanket of CO_2 can enhance the process of photosynthesis and the living zone of plankton population but also destroy the plankton population by restricting the nutrients supply and generating overtemperatures. Higher concentration of harmful chemicals (calcium, chlorine, and magnesium) from contaminated sewage also enhances the mortality rate of plankton population.

$$\begin{aligned} \frac{ds}{dt} &= M_2 - \alpha s + \gamma_1 p + \gamma_2 z - \frac{m_1sp}{a_1 + s} \equiv G_1(s, p, z) \\ \frac{dp}{dt} &= r_1 p + \phi \frac{m_1sp}{a_1 + s} - \frac{m_2zp}{a_2 + p} - \beta_3 p - c_1 h_1 p \equiv G_2(s, p, z) \\ \frac{dz}{dt} &= r_2 z + \frac{m_2zp}{a_2 + p} - \beta_4 z - \frac{\mu pz}{a_3 + p} - c_2 h_2 z \equiv G_3(s, p, z). \end{aligned} \quad (3)$$

where:

M_2 —initial nutrient concentration.

β_3 —death rate of phytoplankton ($\beta_3 > \beta_1$).

β_4 —death rate of zooplankton ($\beta_4 > \beta_2$).

System (3) has three equilibrium points $E_0^{(2)}, E_1^2, E^{*(2)}$. Where:

$$E_0^{(2)} = (M_2/\alpha, 0, 0), E_1^{(2)} = \left(\frac{a_1(\beta_3 + c_1h_1 - r_1)}{\phi m_1 - (\beta_3 + c_1h_1 - r_1)}, \frac{\phi(M_2 - \alpha s_1)}{(\beta_3 + c_1h_1 - r_1 - \gamma_1)}, 0 \right)$$

Here, $\phi m_1 > \beta_3 + c_1h_1 - r_1$ and $\beta_3 + c_1h_1 - r_1 > \gamma_1$, otherwise E_1^* does not exist. If $M_2 = \alpha s_1$, then E_1^* switches to E_0^* . Investigating the interior steady states $E^{(2)*} = (s^{*(2)}, p^{*(2)}, z^{*(2)})$, we can say $p^{*(2)}$ must satisfy the equation:

$$(\beta_6 + \mu - m_2 - r_2)p^2 + (a_2\beta_6 + a_3\beta_6 + a_2\mu - m_2a_3 - r_2a_2 - r_2a_3)p + a_2a_3(\beta_6 - r_2) = 0.$$

(where $\beta_6 = \beta_4 + c_2h_2$).

This equation has no positive solution if $\beta_6 + \mu - m_2 - r_2 > 0$. If $\beta_6 + \mu - m_2 - r_2 < 0$, then we get a unique positive value of $p^{*(1)}$. However, if $\beta_6 + \mu - m_2 - r_2 > 0$ and $a_2\beta_6 + a_3\beta_6 + a_2\mu - m_2a_3 - r_2a_2 - r_2a_3 < 0$, then either the equation has two positive solutions or no positive solutions.

z can be written in terms of s as:

$$\left(r_1 + \phi \frac{m_1 s}{a_1 + s} - \beta_3 - c_1 h_1 \right) \left(\frac{a_2 + p}{m_2} \right) = f_2(s)$$

Where p is a positive root of the above equation and s component of the interior steady state, $s^{*(1)}$ satisfies

$$M_2 - \alpha s + \gamma_1 p + \gamma_2 f_2(s) - \frac{m_1 s p}{a_1 + s} = 0.$$

$E^{*(1)}$ does not exist if:

$$\phi \frac{m_1 s}{a_1 + s} < \beta_3 + c_1 h_1 - r_1.$$

Due to addition of new variables, the equations of system (1) are changed. But, the local stability analysis process of system (2) around the biological feasible equilibria is quite the same as the system (1). Conditions for local stability are thoroughly discussed in the Appendix section.

We see how the dynamics change due to change of some variables in the system:

- (i) When $r_1 = 0, r_2 = 0$, the system is stable at the assumed parametric values (Figure 2) (see Table 1) but when $r_1 = 0.1$ (Figure 4), we observe massive increase in bloom heights.
- (ii) When $r_1 = 0, r_2 = 0.1$, we see blooms for both plankton (Figure 5). System stability $\alpha = 0.01, \mu = 0.06$ ($r_1 = 0, r_2 = 0$) is disturbed. Here, the number of bloom decreases than before ($r_1 = 0.1, r_2 = 0$).

Table 1 A hypothetical set of parameter values.

Parameter	Definition	Default value	Unit
M	Constant input of nutrient	0.01	$mgml^{-1}$
α	Dilution rate of nutrient	0.01	day^{-1}
m_1	Nutrient uptake rate for the phytoplankton	0.6	$mlmg^{-1}day^{-1}$
m_2	Maximal zooplankton ingestion rate for the growth of phytoplankton	0.6	$mlmg^{-1}day^{-1}$
β_1	Mortality rate of phytoplankton	0.21	day^{-1}
β_2	Mortality rate of zooplankton	0.1	day^{-1}
γ_1	Nutrient recycle rate due to the death of phytoplankton	0.1	$mgday^{-1}$
γ_2	Nutrient recycle rate due to the death of zooplankton	0.07	$mgday^{-1}$
a_1	Half-saturation constant for nutrient	0.2	$mlday^{-1}$
a_2	Half-saturation constant for phytoplankton	0.2	$mlday^{-1}$
a_3	Half-saturation constant for zooplankton	0.2	$mlday^{-1}$
c_1	Carrying capacity	0.1	$mlday^{-1}$
c_2	Carrying capacity	0.1	$mlday^{-1}$
h_1	Harvest rate of phytoplankton	0.1	$mlday^{-1}$
h_2	Harvest rate of zooplankton	0.1	$mlday^{-1}$
μ	Rate of toxin production by phytoplankton species	0.06	$mlday^{-1}$

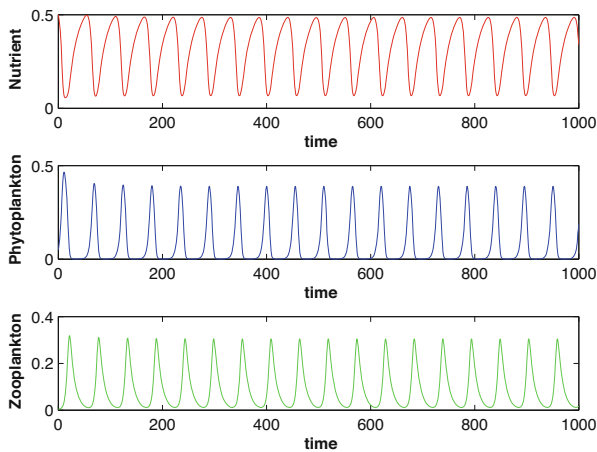


Fig. 4 Stability disturbed, huge blooms for $r_1 = 0.1$; other parametric values are same as *Figure 2*.

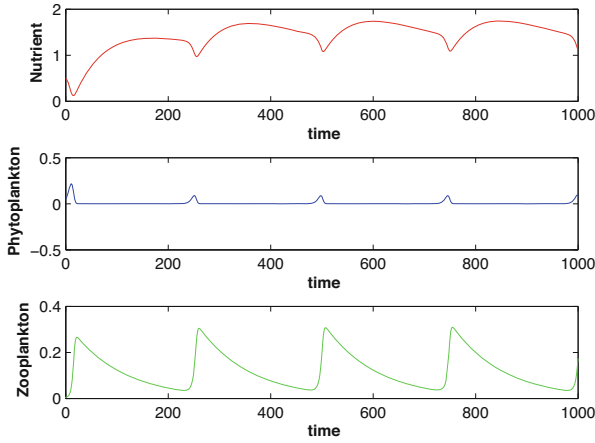


Fig. 5 Stability disturbed, huge blooms for $r_2 = 0.1$

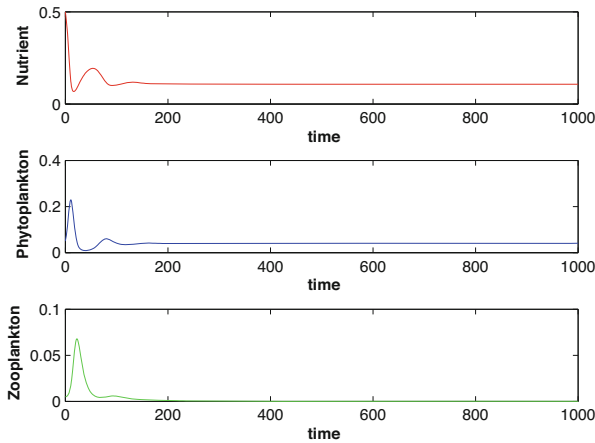


Fig. 6 Extinction of zooplankton for $M_2 = 0.007, \mu = 0.07, \alpha = 0.025, r_1 = 0, r_2 = 0$ and other parametric values are unchanged.

(iii) Now, we observe an interesting thing, when concentration of initial nutrient is low ($M_2 = 0.007$) and $m_1 = 0.63, \mu = 0.07$, zooplankton population goes to extinction (Figure 6) but the presence of additional food for phytoplankton ($r_1 = 0.1, r_2 = 0$) helps them to survive (Figure 7). Similarly, additional food for zooplankton ($r_1 = 0, r_2 = 0.1$) helps zooplankton population to survive (Figure 8).

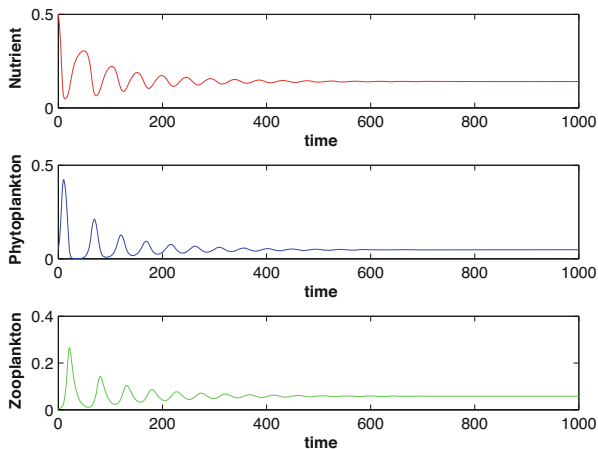


Fig. 7 Zooplankton survives again from extinction for $M_2 = 0.007$, $\mu = 0.07$, $\alpha = 0.025$, $r_1 = 0.1$

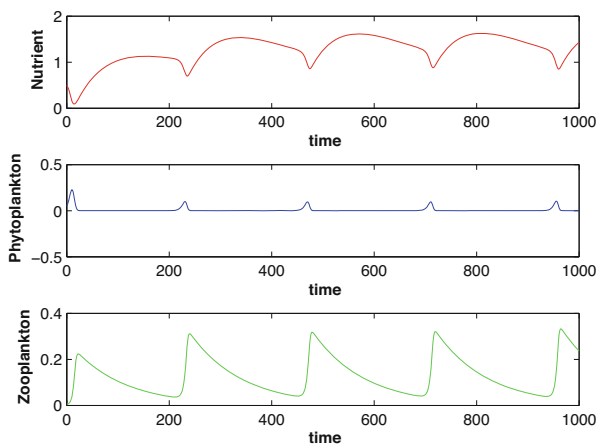


Fig. 8 Zooplankton survives again from extinction for $M_2 = 0.007$, $\mu = 0.07$, $\alpha = 0.025$, $r_2 = 0.1$

6 Discussion

In this paper, we have constructed and analyzed models of nutrient-phytoplankton-zooplankton populations. In our first model, an N-P-Z model with general nutrient uptake functions, instantaneous nutrient recycling, and harvesting on plankton populations is considered. In an addition, we also have discussed the effect of human impact upon the N-P-Z model. The effect of toxin produced by TPP on zooplankton population is also taken under our consideration. The main objective of the present

study is to analyze the dynamics of N-P-Z models and achieve fruitful results in the presence of natural resources, toxic chemicals, and human impact on marine system.

We notice that three primary parameters, namely toxic substances, positive atmospheric situations, and human-made pollution, have monitored the system. We know that TPP is an important key factor to control planktonic bloom by reducing the grazing pressure of zooplankton. System (1) gets oscillation for low toxin production rate but finds the stability at a fitting rate of toxin production. An increase in toxic level may cause the extinction of zooplankton population. Interesting factors arise when we include accumulated nutrients due to human impact and natural sources (like sunlight, and inevitable minerals) into the system. Sunlight and additional food for phytoplankton definitely enhance its growth rate and also transmitted to zooplankton population. Due to additional food, we see the rise in bloom heights where nutrient dilution rate (α) and toxin production rate (μ) remained unchanged. Even, additional nutrient availability favors the zooplankton population to survive from extinction under high toxin production rate and low initial nutrient concentration. According to our study, CO_2 can enhance photosynthesis rate but is responsible for nutrient unavailability and higher mortality rate of plankton population. But, small amount of additional food can change the entire dynamics. Extinction can be avoided and plankton can retrieve from extinction.

Appendix

Stability Analysis of the System

We construct the 3×3 Jacobian matrix:

$$\begin{pmatrix} -\alpha - \frac{m_1 a_1 p}{(a_1 + s)^2} & \gamma_1 - \frac{m_1 s}{a_1 + s} & \gamma_2 \\ \frac{m_1 a_1 p}{(a_1 + s)^2} & \frac{m_1 s}{a_1 + s} - \frac{m_2 a_2 z}{(a_2 + p)^2} - \beta_1 - h_1 c_1 & -\frac{m_2 p}{a_2 + p} \\ 0 & \frac{m_2 a_2 z}{(a_2 + p)^2} - \frac{\mu a_3 z}{(a_3 + p)^2} & \frac{m_2 p}{a_2 + p} - \beta_2 - h_2 c_2 - \frac{\mu p}{a_3 + p} \end{pmatrix}$$

We check the eigen values at every equilibrium point E_0, E_1, E^* .

The Jacobian at the equilibrium point $E_0 = (\frac{M}{\alpha}, 0, 0)$ has the eigen values $-\alpha, \frac{m_1 s}{a_1 + s} - \beta_1 - c_1 h_1$, and $-(\beta_2 + c_2 h_2)$. So, we can say that E_0 is locally asymptotically stable if $\frac{m_1 s}{a_1 + s} - \beta_1 - c_1 h_1 < 0$.

The variational matrix of the system around the positive equilibrium $E^* = (s^*, p^*, z^*)$ is

$$\begin{pmatrix} n_{11} & n_{12} & n_{13} \\ n_{21} & n_{22} & n_{23} \\ 0 & n_{32} & 0 \end{pmatrix}$$

where $n_{11} = -\alpha - \frac{m_1 a_1 p^*}{(a_1 + s^*)^2} < 0$, $n_{12} = \gamma_1 - \frac{m_1 s^*}{a_1 + s^*} < 0$, $n_{13} = \gamma_2 > 0$, $n_{21} = \frac{m_1 a_1 p^*}{(a_1 + s^*)^2} > 0$, $n_{22} = \frac{m_1 s^*}{a_1 + s^*} - \frac{m_2 a_2 z^*}{(a_2 + p^*)^2} - \beta_1 - h_1 c_1 = \frac{m_2 z}{a_2 + p} - \frac{m_2 a_2 z^*}{(a_2 + p^*)^2} > 0$, $n_{23} = -\frac{m_2 p^*}{a_2 + p^*} < 0$, and $n_{32} = \frac{m_2 a_2 z^*}{(a_2 + p^*)^2} - \frac{\mu a_3 z^*}{(a_3 + p^*)^2} > 0$

The characteristic equation is of the form $\lambda^3 + A_1 \lambda^2 + A_2 \lambda + A_3$,

where $A_1 = -(n_{11} + n_{22})$, $A_2 = n_{11} n_{22} - n_{12} n_{21} - n_{23} n_{32}$, and $A_3 = n_{11} n_{23} n_{32} - n_{13} n_{21} n_{32}$.

By the Routh-Hurwitz criteria, all roots of above equation have negative real parts if and only if $A_1 > 0$, $A_3 > 0$, and $A_1 A_2 - A_3 > 0$.

$A_1 = -(n_{11} + n_{22}) = -(-\alpha - \frac{m_1 a_1 p^*}{(a_1 + s^*)^2} + \frac{m_2 z^*}{a_2 + p^*} - \frac{m_2 a_2 z^*}{(a_2 + p^*)^2}) > 0$

if $\alpha + \frac{m_1 a_1 p^*}{(a_1 + s^*)^2} + \frac{m_2 a_2 z^*}{(a_2 + p^*)^2} > \frac{m_2 z^*}{a_2 + p^*}$.

$A_2 = n_{11} n_{22} - n_{12} n_{21} - n_{23} n_{32} > 0$

if $n_{11} n_{22} > n_{12} n_{21} + n_{23} n_{32}$

$$\begin{aligned} A_3 &= n_{11} n_{23} n_{32} - n_{13} n_{21} n_{32} \\ &= (-\alpha - \frac{m_1 a_1 p^*}{(a_1 + s^*)^2}) (-\frac{m_2 p^*}{a_2 + p^*}) (\frac{m_2 a_2 z^*}{(a_2 + p^*)^2} - \frac{\mu a_3 z^*}{(a_3 + p^*)^2}) \\ &\quad - \gamma_2 \frac{m_1 a_1 p^*}{(a_1 + s^*)^2} (\frac{m_2 a_2 z^*}{(a_2 + p^*)^2} - \frac{\mu a_3 z^*}{(a_3 + p^*)^2}) \\ &= [\alpha + \frac{m_1 a_1 p^*}{(a_1 + s^*)^2}] (\frac{m_2 p^*}{a_2 + p^*}) - \gamma_2 \frac{m_1 a_1 p^*}{(a_1 + s^*)^2} (\frac{m_2 a_2 z^*}{(a_2 + p^*)^2} - \frac{\mu a_3 z^*}{(a_3 + p^*)^2}) \\ &= [\alpha (\frac{m_2 p^*}{a_2 + p^*}) + (\frac{m_1 a_1 p^*}{(a_1 + s^*)^2}) (\frac{m_2 p^*}{a_2 + p^*}) - \gamma_2 \frac{m_1 a_1 p^*}{(a_1 + s^*)^2}] (\frac{m_2 a_2 z^*}{(a_2 + p^*)^2} \\ &\quad - \frac{\mu a_3 z^*}{(a_3 + p^*)^2}) = [\alpha (\frac{m_2 p^*}{a_2 + p^*}) + (\frac{m_1 a_1 p^*}{(a_1 + s^*)^2}) \\ &\quad (\frac{m_2 p^*}{a_2 + p^*} - \gamma_2)] (\frac{m_2 a_2 z^*}{(a_2 + p^*)^2} - \frac{\mu a_3 z^*}{(a_3 + p^*)^2}) > 0 \end{aligned}$$

Since, $\frac{m_2 p^*}{a_2 + p^*} > \gamma_2$ and $\frac{m_2 a_2 z^*}{(a_2 + p^*)^2} > \frac{\mu a_3 z^*}{(a_3 + p^*)^2}$

Finally, $A_1 A_2 - A_3 = -(n_{11} + n_{22})(n_{11} n_{22} - n_{12} n_{21} - n_{23} n_{32}) - (n_{11} n_{22} - n_{12} n_{21} - n_{23} n_{32}) > 0$ since, $-(n_{11} + n_{22})(n_{11} n_{22} - n_{12} n_{21} - n_{23} n_{32}) > n_{11} n_{22} - n_{12} n_{21} - n_{23} n_{32}$.

Acknowledgement The research is partially supported by the University Grants Commission, New Delhi [grant number MRP-MAJ-MATH-2013-609].

References

1. M. Droop, Some thoughts on nutrient limitation in algae, *J. Phycol.*, **9** (1973) 264–272.
2. K. DEIMLING, Nonlinear Functional Analysis, *Springer-Verlag*, Berlin, 2008.
3. Y. Du, S.B. Hsu, Concentration phenomena in a nonlocal quasilinear problem modelling phytoplankton I: existence, *SIAM J. Math. Anal.* **40** (2008) 1419–1440.
4. Y. Du, S.B. Hsu, Concentration phenomena in a nonlocal quasilinear problem modelling phytoplankton II: limiting profile, *SIAM J. Math. Anal.* **40** (2008) 1441–1470.
5. Y. Du, S.B. Hsu, On a nonlocal reaction-diffusion problem arising from the modelling of the phytoplankton growth: *SIAM J. Math. Anal.* **42** (2010) 1305–1333.
6. Evans, G.T. and Parslow, J.S. : A model of annual plankton cycles. *Biol. Oceanogr.* **3**, (1985), 327–427.
7. Dancer EN (1984), On positive solutions of some pairs of differential equation, *Trans Amer Math Soc* **284** 729–743.
8. U. Ebert, M. Arrayas, N. Temme, B. Sommeijer, J. Huisman, Critical condition for phytoplankton blooms, *Bull. Math. Biol.* **63** (2001) 1095–1124.
9. H.L. Smith, P.E. Waltman, The Theory of the Chemostat, *Cambridge University Press*. (2008).
10. Courant R, Hilbert D (1953) Methods of Mathematical Physics, Vol. I *Wiley Interscience*, New York.
11. X.-Q. ZHAO, Dynamical Systems in Population Biology, *Springer*, New York, 2003.
12. S. Chakraborty, S. Roy, J. Chattopadhyay, Nutrient-limited toxin production and the dynamics of two phytoplankton in culture media: a mathematical model, *Ecol. Model.* **213**, (2008), 191–201.
13. Brewer PG, Goldman JC (1976), Alkalinity changes generated by phytoplankton growth, *Limnol Oceanogr* **21** 108–117.
14. Busenberg, S., Kishore, K.S., Austin, P. and Wake, G. : The dynamics of a model of a plankton - nutrient interaction. *J. Math. Biol.* **52**, (1990), 677–696.
15. D. Tilman, Resource Competition and community structure. *Princeton University Press* New Jersey, 1982.
16. Greer AT, Cowen RK, Guiland CM, McManus MA, Sevadjian JC, Timmerman AHV (2013), Relationships between phytoplankton thin layers and the fine-scale vertical distributions of two trophic levels of zooplankton. *Journal of Plankton Research* **35**, 939–956.
17. Badger MR, Price GD, Long BM, Woodger FJ (2006) The environmental plasticity and ecological genomics of cyanobacterial CO_2 concentrating mechanisms, *J. Exp Bot* **57** 249–265.
18. Ruan, S. : Persistence and coexistence in zooplankton-phytoplankton-nutrient models with instantaneous nutrient recycling. *J. Math. Biol.* **31**, (1993), 633–654.
19. Pardo, O. : Global stability for a phytoplankton-nutrient system. *J. Biological Systems* **8**, (2000), 195–209.
20. Edwards, A.M. and Brindley, J. : Zooplankton mortality and the dynamical behaviour of plankton population models. *Bull. Math. Biol.*, **61**, (1999), 303–339.
21. Cushing DH (1975), Marine ecology and fisheries, *Cambridge University Press*, London.
22. Edwards, A.M. and Yool, A. : The role of higher predation in plankton population models. *J. Plankton Res.*, **22**, (2000), 1085–1112.
23. Ruan, S. : Oscillations in Plankton Models with Nutrient Recycling *J. Theor. Biol.* **208**, (2001), 15–26.
24. Bairagi, N., Pal, S., Chatterjee, S., Chattopadhyay, J.: Nutrient, non-toxic phytoplankton, toxic phytoplankton and zooplankton interaction in the open marine system. In: Hosking, R.J., Venturino, E. (Eds), Aspects of Mathematical Modelling. Mathematics and Biosciences in Interaction. Birkhauser Verlag Basel, Switzerland, (2008), 41–63.
25. Oscar Angulo, J.C. Lopez-Marcos, M.A. Lopez-Marcos, Numerical Analysis of a Size-Structured Population Model with a Dynamical Resource, *Biomath* **3** (2014), 1403241, <http://dx.doi.org/10.11145/j.biomath.2014.03.241>.

26. Mitra, A and Flynn, K.J. : Promotion of harmful algal blooms by zooplankton predatory activity. *Biol. Lett.*, **2** (2),(2006),194–197.
27. Morozov, A. and Arashkevich, E. : Patterns of Zooplankton Functional Response in Communities with Vertical Heterogeneity : a Model Study. *Math Model. Nat. Phenom.* **3** (3), (2008), 131–149.
28. Poggiale, J.-C., Gauduchon, M., Auger, P. : Enrichment Paradox Induced by Spatial Heterogeneity in a Phytoplankton-Zooplankton System. *Math Model. Nat. Phenom.* **3**, (2008), 87–102.
29. Rene Alt, Jean-Luc Lamotte, Stochastic Arithmetic as a Tool to Study the Stability of Biological Models, *Biomath* 2 (2013), 1312291,<http://dx.doi.org/10.11145/j.biomath.2013.12.291>.
30. Condon RH, Duarte CM, Pitt KA, Robinson KL, Lucas CH, Sutherland KR, Mianzan HW, Bogeberg M, Purcell JE, Decker MB, and others (2013) Recurrent jellyfish blooms are a consequence of global oscillations. *Proc Natl Acad Sci U S A* 110: 1000–1005

A Computational Study of Reduction Techniques for the Minimum Connectivity Inference Problem



Muhammad Abid Dar, Andreas Fischer, John Martinovic,
and Guntram Scheithauer

1 Introduction

The minimum connectivity inference (MCI) problem is an NP-hard discrete optimization problem [6, 11] which has been independently discussed in various areas of research under different names. Most probably, its first appearance dates back to a Chinese journal [10] from 1976, where this problem was considered for the design of vacuum systems. A reference to this paper can be found on Ding-Zhu Du's homepage [9] and in [8].

Moreover, the problem under consideration was studied in [11] and in [15] under the term Subset Interconnection Design problem. Besides this, the name Minimum Topic-Connected Overlay problem (see [4–6, 13]) has been established in the context of scalable overlay networks. Furthermore, the MCI problem was investigated for the design of reconfigurable interconnection networks and called Interconnection Graph Problem [12]. With reference to underlying social networks, the MCI problem was also dealt with as Network Inference problem [3]. Sometimes, related problems were studied as well, in particular the case with nonuniform edge weights is dealt with. This weighted MCI problem is out of the scope of our contribution.

The name Minimum Connectivity Inference problem itself appeared first in recent publications dealing with applications in structural biology to discover connections within macromolecular assemblies [1, 2]. Here, this most current term

M. A. Dar · A. Fischer (✉) · J. Martinovic · G. Scheithauer
Institute of Numerical Mathematics, Technische Universität Dresden, 01062 Dresden, Germany
e-mail: Muhammad_Abid.Dar@tu-dresden.de; Andreas.Fischer@tu-dresden.de;
John.Martinovic@tu-dresden.de; Guntram.Scheithauer@tu-dresden.de

© Springer Nature Switzerland AG 2019
V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_7

will be used to address the problem under consideration, especially since (among all the existing names) the latter seems to show best the actual objective behind the given optimization problem.

An abstract definition of the MCI problem is as follows:

Definition 1 Consider a simple, undirected, and complete graph $G = (V, E)$, where $V = \{1, \dots, m\}$ and $E = \{e = \{i, j\} \mid i, j \in V, i \neq j\}$ are the sets of vertices and edges, respectively. Moreover, let a finite collection $\mathcal{C} = \{V_i \mid V_i \subseteq V, i \in I\}$ of subsets of V , called *clusters*, be given. Then, the MCI problem is to find an edge set $E^* \subseteq E$ of minimal cardinality so that the subgraphs $G^*[V_i]$ induced by V_i in $G^* = (V, E^*)$ are connected for all $i \in I$.

Note that for a graph $G' = (V, E')$ with an edge set $E' \subseteq E$, the subgraph $G'[V_i]$ of G' induced by V_i in G' is the graph with vertex set V_i and all those edges of E' that connect two vertices of V_i . Moreover, recall that an edge $\{i, j\}$ has no orientation and is identical to $\{j, i\}$. The pair (V, \mathcal{C}) is termed an *instance* of the MCI problem. Moreover, a set $E' \subseteq E$ is called *feasible*, if the subgraphs $G'[V_i]$ are connected for all $i \in I$. This means that any two distinct vertices of V_i are connected by a path which contains vertices of V_i only. If a feasible edge set E' also has minimal cardinality, then it is a *solution* of the MCI instance (V, \mathcal{C}) . Without loss of generality, we may assume that $|V_i| \geq 2$, and $V_i \neq V_j$ hold for all $i, j \in I$ with $i \neq j$.

The MCI problem has been studied with respect to heuristic solution techniques, complexity issues, reduction methods, and application-based modeling. As far as we are aware, the only effort to determine an exact solution of general MCI instances is due to [1] by proposing an MILP model. Based on this MILP formulation, only MCI instances of small size can be solved in a reasonable amount of time. Recently, the authors of the current paper presented an improved MILP formulation [7], which, to some extent, allows to successfully solve larger instances of the MCI problem. In [7], some new instance reduction techniques were developed. Further reduction techniques from the literature were discussed as well. The solution of reduced (in some sense smaller) instances is expected to require less computation time and, therefore, can be meaningful for tackling even larger instances.

The reduction techniques presented in [7] are *exact*, i.e., it can be shown that a solution of a reduced instance provides a solution of the original MCI instance by simply adding some of the removed edges. However, for a promising reduction technique, suggested in [1], it was shown in [4] that the property to be exact is not fulfilled, in general. This reduction technique will always lead to a feasible edge set, but not necessarily to an optimal one. Therefore, we call it *heuristic* reduction technique.

In this paper, we aim to demonstrate benefits and limits of applying instance reduction techniques. Therefore, we present the results obtained without any reduction of instances and compare them with those resulting from the use of several reduction techniques. Moreover, besides the application of only exact reduction

techniques we also analyze effects of additionally using the heuristic reduction rule from [1]. All the computations and comparisons are based on the improved MILP formulation [7]. Moreover, they are done with respect to different classes of MCI instances proposed in [7].

The paper is organized as follows. After a short description of the improved MILP model for the MCI problem in Section 2, we describe several exact instance reduction techniques as well as the heuristic reduction rule in Section 3. The computational study is given in Section 4. There, in particular, the test environment including the choice of test instances, the computational results, and their discussion with regard to several aspects are given. Finally, some concluding remarks follow in Section 5.

2 An MILP Model of the MCI Problem

In this section, we briefly describe the improved MILP formulation. The numerical experiments in [7] showed a significantly better performance for the improved formulation than for the first MILP formulation from [1]. The description, based on a given instance (V, \mathcal{C}) of the MCI problem, requires some notations. Firstly, in addition to the undirected graph $G = (V, E)$, we introduce a directed graph (V, A) with *arc set*:

$$A = \{(i, j) \mid i, j \in V, i \neq j\}.$$

Moreover, for a subset $U \subset V$ of vertices, the corresponding sets of arcs and edges induced by U are defined as:

$$A(U) = \{(i, j) \mid i, j \in U, i \neq j\}, \quad E(U) = \{(i, j) \mid i, j \in U, i \neq j\}.$$

Furthermore, for any cluster $i \in I$ and each vertex $j \in V_i$, the sets of incoming and outgoing arcs are given by:

$$A_i^+(j) = \{(k, j) \mid k \in V_i \setminus \{j\}\} \quad \text{and} \quad A_i^-(j) = \{(j, k) \mid k \in V_i \setminus \{j\}\}.$$

Finally, for any $i \in I$, let $r_i \in V_i$ denote an arbitrary but fixed vertex. Then, as shown in [7], the MCI problem can be modeled by means of two types of variables. Decision variables $x_e \in \{0, 1\}$, $e \in E$, are used to indicate whether an edge $e \in E$ belongs to a feasible edge set ($x_e = 1$) or not ($x_e = 0$). By means of the flow variables f_a^i , $a \in A(V_i)$, $i \in I$, the connectivity constraint for each cluster is described. The resulting MILP formulation can be stated as follows:

$$\text{minimize } \sum_{e \in E} x_e \quad (1)$$

$$\text{s.t. } \sum_{e \in E(V_i)} x_e \geq |V_i| - 1, \quad i \in I, \quad (2)$$

$$\sum_{a \in A_i^-(j)} f_a^i - \sum_{a \in A_i^+(j)} f_a^i = -1, \quad j \in V_i \setminus \{r_i\}, i \in I, \quad (3)$$

$$f_{(j,k)}^i + f_{(k,j)}^i \leq (|V_i| - 1) \cdot x_e, \quad i \in I, e = \{j, k\} \in E(V_i), \quad (4)$$

$$f_a^i \geq 0, \quad i \in I, a \in A(V_i), \quad (5)$$

$$x_e \in \{0, 1\}, \quad e \in E. \quad (6)$$

Note that constraints in (2) are, in fact, not necessary to formulate a correct model of the MCI problem. However, their presence essentially strengthens the corresponding LP relaxation and, therefore, they are important for the reduction of the overall effort to solve an instance, for details see [7]. Moreover, we note that restricting the MILP model (1)–(6) to the case of a single cluster $V_1 = V$ leads to a single commodity flow model for the minimum spanning tree problem. It is well known that there are further models for the latter problem [14]. Their possible use for establishing further MILP models of the MCI problem is a topic of future research, in particular with respect to the achievement of an even better numerical performance.

3 Instance Reduction Techniques

The first subsection below describes the exact reduction rules we use for our computational study in Section 4. Moreover, the heuristic reduction rule from [1] is presented in Subsection 3.2. In what follows, $\mathcal{H} = (V, \mathcal{C})$ denotes some instance of the MCI problem that is used as input for a reduction rule. For any vertex $u \in V$, let $\mathcal{C}(u)$ denote the set of all those clusters which contain vertex u , i.e., $\mathcal{C}(u) = \{V_i \mid u \in V_i, i \in I\}$. Furthermore, \mathcal{H}_u or \mathcal{H}_U , respectively, denotes an instance of the MCI problem, where the vertex u or, respectively, all vertices in U are removed from V and from the clusters in \mathcal{C} .

Sometimes, the reduction rules presented in Subsections 3.1 and 3.2 lead to an instance $\tilde{\mathcal{H}} = (\tilde{V}, \tilde{\mathcal{C}})$, where $\tilde{\mathcal{C}}$ contains a cluster with less than two elements. Then, according to [1], such clusters are simply removed before any further reduction or before solving the reduced instance. Further note that some reduction rules may provide two or more identical clusters. Since the collection $\tilde{\mathcal{C}}$ of clusters is a set, only one of those identical clusters remains.

3.1 Exact Reduction Rules

Rules 1 and 2 below are taken from [5]. Other rules provided therein are either made for special cases (very small number of clusters compared to the number of vertices) or are not applicable since the knowledge of an optimal edge set of the instance is required. The Rules 3–5 were suggested and analyzed in [7].

Rule 1 Assume that there are $q + 1$ vertices $u, v_1, \dots, v_q \in V$ with q being a positive integer such that

- $\mathcal{C}(u) \subseteq \mathcal{C}(v_i)$ holds for $i = 1, \dots, q$ and
- $|V_j| \leq q + 3$ is satisfied for all $V_j \in \mathcal{C}(u)$.

Then, any solution E_u of \mathcal{H}_u can be converted into a solution of \mathcal{H} by adding a single edge $\{u, v\}$ to the edge set E_u , where v is an arbitrarily chosen vertex from the set $\{v_1, \dots, v_q\}$.

Rule 2 Assume that there is a cluster $V_i \in \mathcal{C}$ with $V_i = \{u, u_1, \dots, u_l\}$ such that $\mathcal{C}(u_j) \subseteq \mathcal{C}(u)$ holds for $j = 1, \dots, l$. Then, any solution $E_{\{u_1, \dots, u_l\}}$ of $\mathcal{H}_{\{u_1, \dots, u_l\}}$ can be converted into a solution of \mathcal{H} by adding the edges in $\{\{u, u_j\} \mid j = 1, \dots, l\}$ to the edge set $E_{\{u_1, \dots, u_l\}}$.

To describe the reduction rules proposed in [7], we need to introduce some more notation. First, let the graph $\mathcal{G} = (I, \mathcal{E})$ be assigned to (V, \mathcal{C}) , where the edge set \mathcal{E} is defined as:

$$\mathcal{E} = \{\{j, k\} \mid V_j \cap V_k \neq \emptyset, j, k \in I, j \neq k\}.$$

For any $J \subset I$, let $\mathcal{G}[J] = (J, \mathcal{E}(J))$ with

$$\mathcal{E}(J) = \{\{i, j\} \mid V_i \cap V_j \neq \emptyset, i, j \in J, i \neq j\}$$

denote the corresponding induced subgraph of \mathcal{G} . Moreover, we define the set:

$$V(J) = \bigcup_{j \in J} V_j.$$

For a particular cluster $V_i \in \mathcal{C}$, let $J_i = \{j \in I \mid V_j \subset V_i, j \neq i\}$ collect the clusters completely contained in V_i , and let γ_i denote the number of *connected components* of $\mathcal{G}[J_i]$.

If, for a certain cluster $V_i \in \mathcal{C}$, the graph $\mathcal{G}[J_i]$ has at least two connected components, i.e., if $\gamma_i > 1$, then we represent each connected component $k \in \{1, \dots, \gamma_i\}$ by the set $J_{i,k} \subset J_i$ of those indices of clusters which are involved in the component. Obviously:

$$\bigcup_{k=1}^{\gamma_i} J_{i,k} = J_i \quad \text{and} \quad J_{i,k} \cap J_{i,l} = \emptyset \quad \text{for } k, l \in \{1, \dots, \gamma_i\} \text{ with } k \neq l$$

are satisfied.

Rule 3 Assume that there is a cluster $V_i \in \mathcal{C}$ with $V(J_i) = V_i$ and $\gamma_i = 1$, then any solution of $\mathcal{H}' = (V, \mathcal{C} \setminus \{V_i\})$ is also a solution of the instance \mathcal{H} .

Rule 4 Assume that there is a cluster $V_i \in \mathcal{C}$ with $V(J_i) = V_i$ and $\gamma_i > 1$. If there is no cluster V_p with $p \in I \setminus \{i\}$, $V_p \cap V(J_{i,k}) \neq \emptyset$, and $V_p \cap V(J_{i,l}) \neq \emptyset$ for any pair (k, l) with $k, l \in \{1, \dots, \gamma_i\}$ and $k \neq l$, then any solution E' of the instance $\mathcal{H}' = (V, \mathcal{C} \setminus \{V_i\})$ can be converted into a solution of \mathcal{H} by adding $\gamma_i - 1$ edges $\{u, v_k\}$ to the edge set E' , where $u \in V(J_{i,1})$ and $v_k \in V(J_{i,k})$ for $k = 2, \dots, \gamma_i$ are arbitrarily chosen vertices.

Rule 5 Assume that $V_i \in \mathcal{C}$ is a cluster satisfying $V(J_i) \neq V_i$ and $\gamma_i = 1$. If it holds that

$$V_i \cap V_k = \{v\} \quad \text{for all } V_k \in \mathcal{C}(v) \setminus \{V_i\} \text{ and all } v \in V_i \setminus V(J_i),$$

then any solution E' of the instance $\mathcal{H}' = (V, \mathcal{C} \setminus \{V_i\})$ can be converted into a solution of the instance \mathcal{H} by adding the edges in $T \cup \{\{u, v\}\}$ to the edge set E' , where T is the edge set of an arbitrarily chosen spanning tree for the vertex set $V_i \setminus V(J_i)$, and $u \in V(J_i)$ and $v \in V_i \setminus V(J_i)$ are arbitrarily chosen vertices.

3.2 A Heuristic Reduction Rule

For the instance reduction rule proposed in [1], it has to be noticed that, in contrast to the exact Rules 1–5, the following Rule 6 is only a heuristic one. This was proven in [5]. Nevertheless, as subsequent computational results will show, this rule works well in some cases.

Rule 6 Assume that there are vertices $u, v \in V$ with $u \neq v$ so that $\mathcal{C}(u) \subseteq \mathcal{C}(v)$. Then, any solution E_u of \mathcal{H}_u can be converted into a feasible solution of \mathcal{H} by adding the edge $\{u, v\}$ to the edge set E_u .

4 Computational Study

This section is divided into three parts. In Subsection 4.1, we explain the test environment. In particular, we provide the principles of generating random test instances for the MCI problem. Several tables with our computational results including the description of their entries follow in Subsection 4.2. Finally, observations and conclusions are given in Subsection 4.3.

4.1 Test Environment

To generate a certain variety of MCI instances, we follow the lines suggested in [7]. More precisely, for a certain pair $(|V|, |\mathcal{C}|)$ of cardinalities, we define four *instance types* for the range the cardinalities of the clusters V_i in \mathcal{C} must belong to, namely:

$$\begin{aligned}
 \text{Type 1: } & |V_i| \in \{2, \dots, |V|\}, \\
 \text{Type 2: } & |V_i| \in \{2, \dots, \lfloor |V|/2 \rfloor\}, \\
 \text{Type 3: } & |V_i| \in \{\lceil |V|/4 \rceil, \dots, |V|\}, \\
 \text{Type 4: } & |V_i| \in \{\lceil |V|/4 \rceil, \dots, \lfloor |V|/2 \rfloor\}.
 \end{aligned} \tag{7}$$

Then, for given $|V|$, $|\mathcal{C}|$, and for a fixed instance type, the following procedure generates 50 random instances. The cardinality of each of the clusters $V_1, \dots, V_{|\mathcal{C}|}$ is drawn randomly from the set of integers in (7) according to the uniform distribution. Thereafter, the vertex set of each of these clusters is drawn from $\{1, \dots, |V|\}$ with uniform distribution. If, during the generation of a cluster, a vertex is drawn a second time, then it is not used. The same is done if identical clusters occur. For our computations, $|V|$ is chosen greater than or equal to 10, whereas $|\mathcal{C}| \in \{|V|, \dots, 5|V|\}$ is used, details are shown in the tables in Subsection 4.2.

The solution of any MCI instance is based on the solution of the MILP model in Section 2 by means of CPLEX⁶ Version 12.6.3 on a PC with Intel⁶ Xeon⁶ processor X5670 at 2.93 GHz using 96 GB of memory. The preprocessing including the application of reduction rules and the call of the CPLEX⁶ routine `plexmilkp` is done in MATLAB⁶ Release R2016a. After 600 seconds, the solution of any instance is stopped.

4.2 Computational Results

The results are presented in five tables below. Most of their entries were obtained by averaging the measurements for the 50 random instances obtained for fixed values of $|V|$, $|\mathcal{C}|$ and of the instance types (7). The structure of Tables 1, 2, 3 is similar. They provide results for the solution of instances with different numbers $|V|$ of vertices, where the number $|\mathcal{C}|$ of clusters is equal to $|V|$ for Table 1, to $2|V|$ for Table 2, and to $3|V|$ and $5|V|$ for Table 3, subject to the use of different combinations of reduction rules. These tables include also the percentage of clusters and vertices that could be removed. In contrast to this, Tables 4 and 5 present only results on such percentages of reductions but for significantly higher values of $|V|$ and $|\mathcal{C}|$.

To understand the tables, let us explain the notions used therein:

- The column **Type** shows the instance type of a row, according to (7).
- In each row, the column $|E^*|_{\emptyset}$ provides the averaged optimal values of the instances that were solved within the time limit of 600 seconds (out of 50 random instances).

Table 1 Comparison of exact and heuristic reduction rules for instances with $|V|$ clusters

$ V $	Type	$ E^* _{\emptyset}$	Clusters removed (%)			Vertices removed (%) (#)				Runtime (sec)			
			R1–R5	+R6	R6	R1–R5	+R6	R6	+R6+	R1–R5	R1–R6	R6	NoR
10	1	10.8	25.0	6.8	24.8	15.6	34.4	48.8	45	0.11	0.06	0.06	0.14
	2	13.0	11.6	2.2	7.2	19.2	11.0	28.2	32	0.08	0.06	0.06	0.07
	3	10.7	23.6	10.4	27.6	12.2	37.4	49.0	47	0.11	0.06	0.07	0.15
	4	12.9	4.6	2.0	4.4	14.0	12.6	26.4	33	0.08	0.06	0.06	0.08
14	1	16.6	14.0	2.9	8.6	2.4	31.3	33.3	49	0.37	0.20	0.21	0.52
	2	19.8	1.9	0.7	2.1	5.0	15.9	20.3	48	0.16	0.13	0.13	0.15
	3	16.4	13.7	4.6	10.0	1.0	32.6	33.6	50	0.37	0.20	0.25	0.59
	4	19.76	0.1	0.1	0.3	1.4	14.4	15.9	46	0.19	0.17	0.18	0.18
18	1	22.8	10.2	0.2	2.4	1.3	20.9	22.1	48	1.91	0.98	1.38	2.37
	2	27.5	0.6	0.0	0.6	1.6	12.9	14.4	50	0.36	0.37	0.36	0.33
	3	22.1	11.7	1.3	3.9	0.0	21.7	21.7	48	4.27	1.80	2.45	5.44
	4	27.0	0.0	0.0	0.0	0.2	6.6	6.8	34	0.89	0.85	0.83	0.90
22	1	29.4	8.5	0.3	1.6	0.2	15.0	15.2	44	10.4	6.12	8.84	17.0
	2	35.4	0.1	0.0	0.1	0.9	6.9	7.8	41	1.47	1.35	1.34	1.46
	3	28.4	10.8	0.0	0.9	0.0	11.4	11.4	45	18.4	10.4	18.0	26.9
	4	35.1	0.0	0.0	0.0	0.2	2.9	3.1	24	4.11	3.84	3.81	4.02
26	1	35.8	6.7	0.0	0.4	0.1	10.5	10.6	46	61.5	32.7	62.5	95.5
										[48]	[49]		[49]
	2	44.2	0.1	0.0	0.1	0.2	4.4	4.6	36	3.21	3.08	3.0	3.18
	3	34.3	10.1	0.1	0.5	0.0	7.5	7.5	39	143	164	175	202
										[37]	[38]	[34]	[31]
	4	42.4	0.0	0.0	0.0	0.0	1.9	1.9	20	38.9	37.2	37.2	50.7
										[46]	[47]	[47]	[47]

- The multiple columns **Clusters removed** and **Vertices removed** show percentages of how many clusters and vertices, respectively, could be removed by means of reduction rules.
- The columns **R1–R5**, **R1–R6**, or **R6** refer to results obtained when Rules 1–5, Rules 1–6, or only Rule 6 are/is employed before solving the instances. Note that Rules 1–5 are exact reduction rules, whereas Rule 6 is a heuristic rule, see Subsection 3.2.
- The two columns **+R6** show the percentage of reduction (of clusters or vertices) we obtained by applying Rule 6 to instances that were already reduced by Rules 1–5.
- There is a single column **+R6+** that contains the number (#) of the 50 instances for which the application of Rule 6 (to instances that were already reduced by Rules 1–5) yields an additional reduction in terms of the number of clusters or vertices.

Table 2 Comparison of exact and heuristic reduction rules for instances with $2|V|$ clusters

V	Type	E* _∅	Clusters removed (%)			Vertices removed (%) (#)				Runtime (sec)			
			R1-R5	+R6	R6	R1-R5	+R6	R6	+R6+	R1-R5	R1-R6	R6	NoR
10	1	13.9	23.8	0.8	3.7	0.8	6.0	6.8	18	0.19	0.18	0.23	0.25
	2	17.4	6.0	0.1	0.8	1.6	2.0	3.2	9	0.12	0.12	0.11	0.11
	3	13.6	26.6	0.5	2.5	0.8	5.2	6.0	18	0.19	0.18	0.25	0.28
	4	16.8	2.8	0.1	0.0	0.0	2.2	2.0	10	0.15	0.14	0.14	0.14
14	1	21.3	18.6	0.1	0.4	0.0	2.4	2.4	13	0.91	0.81	1.27	1.23
	2	26.8	1.0	0.0	0.0	0.0	1.3	1.3	8	0.29	0.29	0.27	0.27
	3	19.9	20.2	0.0	0.3	0.4	1.7	2.1	11	1.50	1.42	2.13	2.19
	4	25.3	0.4	0.0	0.0	0.0	0.1	0.1	1	0.67	0.67	0.67	0.67
18	1	29.4	14.5	0.0	0.1	0.0	0.7	0.7	5	5.09	5.00	8.99	9.12
	2	36.7	0.3	0.0	0.0	0.0	0.2	0.2	2	1.66	1.65	1.62	1.63
	3	26.9	16.7	0.0	0.0	0.0	0.3	0.3	2	15.8	15.8	30.7	30.7
	4	34.5	0.0	0.0	0.0	0.0	0.2	0.2	2	5.38	5.35	5.24	5.26
22	1	37.4	12.6	0.0	0.0	0.0	0.0	0.0	0	50.2	50.2	93.7	93.7
										[48]	[48]	[48]	[48]
	2	47.1	0.1	0.0	0.0	0.0	0.1	0.1	1	6.89	6.98	6.91	6.82
	3	34.4	14.9	0.0	0.1	0.0	0.2	0.2	1	189	190	259	269
										[42]	[42]	[32]	[32]
	4	43.9	0.0	0.0	0.0	0.0	0.1	0.1	1	241	241	241	241
										[32]	[32]	[32]	[32]
26	1	46.7	11.3	0.0	0.0	0.0	0.0	0.0	0	154	154	248	248
										[32]	[32]	[22]	[22]
	2	58.6	0.1	0.0	0.0	0.0	0.0	0.0	0	65.6	65.6	65.6	65.6
										[49]	[49]	[49]	[49]
	3	—	13.0	0.0	0.0	0.0	0.1	0.1	1	—	—	—	—
	4	—	0.0	0.0	0.0	0.0	0.0	0.0	0	—	—	—	—

- The multiple column **Runtime** shows the averaged runtimes for solving the 50 instances depending on the reduction rules employed. The column **NoR** provides the averaged runtime if no reduction rule is used at all. In some of the entries of the column **Runtime**, a **bracketed number [x]** can be found below the average runtime. This means that $x < 50$ instances are solved within the time limit of 600 seconds and the average is taken only over these instances.

The measurements (average number of edges in $|E^*|$, average percentages, and average runtimes) given in the tables are rounded.

In Table 2, for $|V| = 26$ and Type 3 or 4, the time limit of 600 seconds was exceeded for all 50 instances. The time for only applying the reduction rules in Tables 4 and 5 (i.e., without any solution of the MILP formulation) required less than 0.3 seconds.

Table 3 Comparison of exact and heuristic reduction rules for instances with $3|V|$ and $5|V|$ clusters

$ V $	Type	$ \mathcal{C} $	Clusters removed (%)			Vertices removed (%) (#)				Runtime (sec)				
			$ E^* _\emptyset$	R1-R5	+R6	R6	R1-R5	+R6	R6	+R6+	R1-R5	R1-R6	R6	NoR
10	1	30	16.1	30.0	0.1	0.5	0.4	0.6	1.0	3	0.21	0.21	0.35	0.36
		50	19.4	42.1	0.0	0.0	0.0	0.0	0.0	0	0.24	0.24	0.60	0.60
	2	30	20.4	10.8	0.0	0.0	0.0	0.2	0.2	1	0.16	0.16	0.15	0.15
		50	24.8	26.6	0.0	0.0	0.0	0.0	0.0	0	0.24	0.2	0.24	0.24
	3	30	15.5	29.6	0.0	0.1	0.0	0.4	0.4	2	0.25	0.25	0.42	0.42
		50	18.3	41.2	0.0	0.0	0.0	0.0	0.0	0	0.32	0.32	0.72	0.72
4	30	19.6	6.5	0.0	0.0	0.0	0.0	0.0	0	0.22	0.22	0.22	0.22	
	50	22.4	15.0	0.0	0.0	0.0	0.0	0.0	0	0.31	0.31	0.34	0.34	
14	1	42	24.7	23.9	0.0	0.0	0.0	0.0	0.0	0	1.27	1.27	1.91	1.91
		70	30.1	30.8	0.0	0.0	0.0	0.0	0.0	0	1.73	1.73	3.51	3.51
	2	42	31.3	3.9	0.0	0.0	0.0	0.0	0.0	0	0.46	0.46	0.50	0.50
		70	38.1	10.9	0.0	0.0	0.0	0.0	0.0	0	0.81	0.81	0.98	0.98
	3	42	22.7	24.1	0.0	0.0	0.0	0.1	0.1	1	2.69	2.69	4.81	4.82
		70	26.5	31.6	0.0	0.0	0.0	0.0	0.0	0	4.72	4.72	11.20	11.20
4	42	29.1	0.7	0.0	0.0	0.0	0.0	0.0	0	1.87	1.87	1.90	1.90	
	70	33.4	2.5	0.0	0.0	0.0	0.0	0.0	0	4.64	4.64	4.62	4.62	
18	1	54	33.8	20.0	0.0	0.0	0.0	0.0	0.0	0	8.82	8.82	15.0	15.0
		90	41.0	26.4	0.0	0.0	0.0	0.0	0.0	0	13.4	13.4	40.1	40.1
	2	54	42.8	0.9	0.0	0.0	0.0	0.0	0.0	0	2.97	2.97	2.99	2.99
		90	52.3	3.8	0.0	0.0	0.0	0.0	0.0	0	4.67	4.67	5.04	5.04
	3	54	30.4	21.3	0.0	0.0	0.0	0.0	0.0	0	41.5	41.5	95.0	95.0
											[46]	[46]	[38]	[38]
		90	35.7	25.3	0.0	0.0	0.0	0.0	0	135	135	285	285	
										[41]	[41]	[28]	[28]	
4	54	39.1	0.1	0.0	0.0	0.0	0.0	0.0	0	68.4	68.4	71.3	71.3	
										[49]	[49]	[49]	[49]	
		90	44.4	0.1	0.0	0.0	0.0	0.0	0	356	356	289	289	
										[5]	[5]	[4]	[4]	

4.3 Observations and Conclusions

Based on the computational results obtained, we now provide several observations and derive conclusions.

1. *Complexity of Instances.* There are several factors that influence the time needed to solve the MILP formulation of an instance. Clearly, increasing values of $|V|$ and $|\mathcal{C}|$ will, in average, lead to higher runtimes. However, in addition to this, we see from Tables 1, 2, 3 that instances of Type 2 (compared to instances of Types 1, 3, and 4) need significantly less runtime (and do not exceed our time limit). At

Table 4 Comparison of exact and heuristic reduction rules for larger numbers of vertices with $|\mathcal{C}| = |V|$

V	Type	Clusters removed (%)			Vertices removed (%)			#
		R1-R5	+R6	R6	R1-R5	+R6	R6	
28	1	7.8	0.0	0.2	0.2	7.6	7.9	42
	2	0.0	0.0	0.0	0.1	3.9	4.0	33
	3	8.8	0.0	0.3	0.0	4.6	4.6	32
	4	0.0	0.0	0.0	0.0	0.9	0.9	12
32	1	6.9	0.0	0.1	0.0	4.4	4.4	35
	2	0.0	0.0	0.0	0.1	2.3	2.3	25
	3	8.4	0.0	0.1	0.0	2.8	2.8	28
	4	0.0	0.0	0.0	0.0	0.3	0.3	5
36	1	6.6	0.0	0.2	0.1	2.7	2.8	26
	2	0.0	0.0	0.0	0.0	1.5	1.5	21
	3	7.8	0.0	0.1	0.0	2.1	2.1	20
	4	0.0	0.0	0.0	0.0	0.1	0.1	2
40	1	5.4	0.0	0.1	0.0	1.3	1.3	16
	2	0.0	0.0	0.0	0.1	1.0	1.0	15
	3	8.6	0.0	0.0	0.0	1.0	1.0	15
	4	0.0	0.0	0.0	0.0	0.1	0.1	1

Table 5 Effect of exact reduction rules (Rules 1-5) for large numbers $|\mathcal{C}|$ of clusters

V	\mathcal{C}	Clusters removed (%)	
		Type 1	Type 3
32	32	6.9	8.4
	64	10.2	11.0
	96	13.3	13.7
	128	15.4	15.4
	160	18.1	17.3
36	36	6.6	7.8
	72	9.8	10.6
	108	12.1	13.0
	144	14.3	15.0
	180	16.5	14.8
40	40	5.3	8.6
	80	9.2	10.4
	120	11.8	11.6
	160	14.1	13.6
	200	15.5	14.5

least for higher numbers of clusters, instances of Type 3 or 4 seem particularly difficult to solve. In general, if some (noticeable) reduction occurs, then this leads to a decrease in the computation time.

2. *Removal of Clusters.* It can be observed from all tables that the exact reduction by Rules 1–5 leads to a remarkable decrease in the number of clusters for Types 1 and 3. It is shown by Tables 4 and 5 that this remains true for larger numbers of vertices and, particularly, quite large number of clusters. In addition to this, Tables 1 and 2 show that an additional heuristic reduction by Rule 6 is not helpful for removing clusters when dealing with instances having more than about 20 vertices.
3. *Removal of Vertices.* Advantages can be observed for problem instances with a small number of clusters, to see this one may compare Tables 1 with Tables 2 and 3. In addition to this, as Table 4 shows, the reduction percentage decreases for an increasing number of vertices. This is mainly caused by the exponential growth of the number of possible clusters for a linearly increasing number of vertices. The application of the heuristic Rule 6 in addition to the exact rules is not useful for instances with a high number of clusters (compared to the number of vertices), see Table 3.
4. *Heuristic Reduction Rule.* Interestingly, for all the generated instances for Tables 1, 2, 3, reductions caused by Rule 6 always led to an exact reduction. Thus, in certain cases, using this rule may be helpful in obtaining at least a good approximate solution. Moreover, for the case that the number of clusters is similar to the number of vertices, an additional use of Rule 6 after the exact Rules 1–5 were applied can lead to further reductions, see Table 1.
5. *Reduction versus No Reduction.* The results recommend to always use the reduction rules (at least the exact Rules 1–5) since, on the one hand, the time needed to apply these rules is negligible (less than 0.3 seconds for all instances considered in all tables). On the other hand, the reduction may lead (in certain cases) to significant savings of solution time, for example see Table 2 with $|V| = 18$, Type 3, and $|\mathcal{C}| = 90$.

An obvious but important fact we would finally like to underline is that any reduction of clusters or vertices leads to a smaller number of variables and constraints in the MILP model.

5 Future Research

In this contribution, we have studied the importance of reduction methods. Although it depends on the particular type of MCI instances and, naturally, on their size, it seems useful to apply all reduction rules. If a solution of an MCI instance is needed, then the exact Rules 1–5 can be used. The development of additional exact reduction rules is an interesting direction of research to further enlarge the range of exactly solvable MCI instances. But also a heuristic reduction (as Rule 6) might be valuable. Then, appropriate sufficient conditions have to be found to verify the optimality of the obtained feasible edge set. This can also be a part of future research. Thirdly, since the MCI problem belongs to the class of NP-hard problems, heuristic approaches are of high interest. As results in [7] show, the heuristic suggested in [1]

computes approximate solutions whose function value can be too far away from the optimal number of edges. Therefore, attempts to develop further heuristic methods might become an important field of research. Finally, to the best of our knowledge, no work has been done so far to develop ideas which would allow to split the MCI instance into several smaller problems, to solve them in parallel and, based on this, to derive a solution or an approximate solution for the original instance. Of course, parallelization can be used at lower levels, in particular for the solution of the MILP model.

Acknowledgements This work is supported in parts by a scholarship of the Governmental Scholarship Programme Pakistan – DAAD/HEC Overseas and by the German Research Foundation (DFG) in the Collaborative Research Center 912 “Highly Adaptive Energy-Efficient Computing (HAEC).”

References

1. Agarwal, D., Araujo, J.-C.S., Caillouet, C., Cazals, F., Coudert, D., Pérennes, S.: Connectivity inference in mass spectrometry based structure determination. In: Bodlaender, H.L., Italiano, G.F. (eds.) European Symposium on Algorithms, Lecture Notes in Computer Science vol. 8125, pp. 289–300. Springer, Berlin (2013)
2. Agarwal, D., Caillouet, C., Coudert, D., Cazals, F.: Unveiling contacts within macromolecular assemblies by solving minimum weight connectivity inference (MWC) problems. *Molecular & Cellular Proteomics* **14**(8), 2274–2284 (2015)
3. Angluin, D., Aspnes, J., Reyzin, L.: Inferring social networks from outbreaks. In: Hutter, M., Stephan, F., Vovk, V., Zeugmann, T. (eds.) Algorithmic Learning Theory. Lecture Notes in Computer Science vol. 6331, pp. 104–118. Springer, Berlin (2010)
4. Chen, C., Jacobsen, H.-A., Vitenberg, R.: Algorithms based on divide and conquer for topic-based publish/subscribe overlay design. *IEEE/ACM Transactions on Networking* **24**(1), 422–436 (2016)
5. Chen, J., Komusiewicz, C., Niedermeier, R., Sorge, M., Suchý, O., Weller, M.: Polynomial-time data reduction for the subset interconnection design problem. *SIAM Journal on Discrete Mathematics* **29**, 1–25 (2015)
6. Chockler, G., Melamed, R., Tock, Y., Vitenberg, R.: Constructing scalable overlays for pub-sub with many topics. In: Gupta, I. (ed.) Proceedings of the Twenty-Sixth Annual ACM Symposium on Principles of Distributed Computing, pp.109–118. ACM, New York (2007)
7. Dar, A., Fischer, A., Martinovic, J., Scheithauer, G.: An improved flow-based formulation and reduction principles for the Minimum Connectivity Inference problem. *Optimization* (2018). <https://doi.org/10.1080/02331934.2018.1465944>
8. Du, D.-Z.: An optimization problem on graphs. *Discrete Applied Mathematics* **14**(1), 101–104 (1986)
9. Du, D.-Z.: Curriculum Vitae of Ding-Zuh Du. <http://www.utdallas.edu/~dx056000/>. Accessed 25 July 2017
10. Du, D.-Z., Chen, Y.-M.: Placement of valves in vacuum systems. *Communication on Electric Light Source Technology* **4**, 22–28 (in Chinese, 1976)
11. Du, D.-Z., Miller, Z.: Matroids and subset interconnection design. *SIAM Journal on Discrete Mathematics* **1**(4), 416–424 (1988)

12. Fan, H., Hundt, C., Wu, Y.-L., Ernst, J.: Algorithms and implementation for interconnection graph problem. In: Yang, B., Du, D.-Z., Wang, C. A. (eds.) *Combinatorial Optimization and Applications*. Lecture Notes in Computer Science vol. 5165, pp. 201–210. Springer, Berlin (2008)
13. Hosoda, J., Hromkovič, J., Izumi, T., Ono, H., Steinová, M., Wada, K.: On the approximability and hardness of minimum topic connected overlay and its special instances. *Theoretical Computer Science* **429**, 144–154 (2012)
14. Magnanti, T.L., Wolsey, L.A.: *Optimal Trees*. In: Ball, M.O., Magnanti, T.L., Monma, B.L., Nemhauser, G. (eds.) *Network Routing*, Handbooks in Operations Research and Management Science vol. 7, pp. 503–615, Elsevier, Amsterdam (1995)
15. Prisner, E.: Two algorithms for the subset interconnection design problem. *Networks* **22**(4), 385–395 (1992)

Approximate Controllability of Nonlocal Impulsive Stochastic Differential Equations with Delay



Surendra Kumar

1 Introduction

The concept of controllability is introduced by Kalman [1]. Controllability is a qualitative property of a dynamical control system and plays a crucial role in modern control theory. In infinite-dimensional case, two basic concepts of controllability can be distinguished, namely, exact and approximate controllability. Exact controllability means that it is possible to steer the system from an arbitrary initial state to an arbitrary final state. To show exact controllability of a dynamical system, the main approach is to convert the controllability issue into a fixed point problem which requires that the controllability operator must have an induced inverse [2, 3]. However, Triggiani [4] proved that if the semigroup associated with the corresponding linear system is compact, then the controllability operator is also compact and hence the induced inverse does not exist. Therefore, the concept of exact controllability is too strong and approximate controllability is more appropriate for infinite-dimensional systems. Approximate controllability means that the system can be steered to arbitrary small neighborhood of the arbitrary final state.

It is well known that many real-world problems in science and engineering are modeled as stochastic differential equations [5]. As a result of its widespread use, the controllability problems related to stochastic differential equations attract many researchers. In 2001, Mahmudov [6] discussed the issue of controllability related to infinite-dimensional linear stochastic system in Hilbert space setting. Then, the results were extended for semilinear stochastic equations by Mahmudov [7, 8],

S. Kumar (✉)

Department of Mathematics, University of Delhi, Delhi 110 007, India
e-mail: mathdma@gmail.com

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_8

149

Dauer and Mahmudov [9], Balachandran, Karthikeyan, and Kim [10], Matar [11], and Karthikeyan, Balachandran, and Sathya [12].

On the other hand, Sakthivel [13] discussed sufficient conditions for the approximate controllability of impulsive stochastic system via contraction mapping principle. Using resolvent operators and fixed point theory, Sakthivel et al. [14] studied the approximate controllability of impulsive deterministic and stochastic nonlinear systems with unbounded delay. Shen and Sun [15] obtained a set of sufficient conditions for the relative controllability and relative approximate controllability in both finite and infinite-dimensional spaces with delay in control. Further, Shen and Sun [16] produced results related to the approximate controllability of stochastic impulsive systems with multiple time-varying delays using the Nussbaum fixed point theorem. The approximate controllability of the stochastic impulsive system with control acting on the nonlinear terms was examined by Shen and Wu [17]. In 2015, Ning and Qing [18] studied the approximate controllability of stochastic nonlinear system with infinite delay. Shukla et al. [19] obtained sufficient conditions for approximate controllability of retarded semilinear stochastic control system with nonlocal conditions and finite delay. Using the Banach fixed point theorem, Arora and Sukavanam [20] examined the approximate controllability of impulsive semilinear stochastic system with finite delay in state in Hilbert space setting. Recently, Mokkedem and Fu [21] formulated and proved a set of sufficient conditions for the approximate controllability of an infinite-dimensional delayed stochastic system in Banach space. However, to the best of my knowledge, there is no result on the approximate controllability of impulsive stochastic system with nonlocal conditions and considered delay in state. Motivated by this fact, in this paper, I will examine the approximate controllability of impulsive stochastic system with time varying delay and nonlocal conditions in Hilbert space setting via semigroup theory, stochastic analysis techniques, and the Banach contraction principle.

The paper is organized as follows: In Section 2, we present some basic definitions and lemmas as preliminaries. The approximate controllability of the system (1) is studied in Section 3. In Section 4, an example is given to illustrate the developed theory.

2 Preliminaries

This section concerns with some basic definitions, lemmas, and notations, which are used in the article. Throughout this paper, unless otherwise specified, let $(\Omega, \mathcal{F}, \mathbf{P})$ be a complete probability space furnished with complete family of right continuous increasing σ -algebras $\{\mathcal{F}_t : t \in J = [0, \tau]\}$ satisfying $\mathcal{F}_t \subset \mathcal{F}$ for $t \geq 0$. Let H , U , and K be real separable Hilbert spaces. For convenience, we denote the inner products and norms in all spaces by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$. Let $\omega = \{\omega(t) : t \geq 0\}$ be a Q -Wiener process defined on the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ with the covariance operator Q such that $Tr(Q) < \infty$. Suppose that $\{e_n\}_{n=1}^{\infty}$ be a complete orthonormal system in K , and $\{\alpha_n\}_{n=1}^{\infty}$ a sequence of independent Brownian motions such that

$$\omega(t) = \sum_{n=1}^{\infty} \sqrt{\delta_n} e_n \alpha_n(t), \quad t \in J,$$

where $\{\delta_n\}_{n=1}^{\infty}$ is a bounded sequence of nonnegative real numbers such that $Qe_n = \delta_n e_n, n = 1, 2, \dots$.

Let $L_2^0 = L_2(Q^{1/2}K, H)$ be the space of all Hilbert-Schmidt operators from $Q^{1/2}K$ into H with the inner product $\langle \chi_1, \chi_2 \rangle_{L_2^0} = Tr[\chi_1 Q \chi_2^*]$. Let $h > 0$ and $C([-h, 0], H)$ denote the Banach space of all continuous functions from $[-h, 0]$ into H with the norm $\|\zeta\| = \sup_{-h \leq t \leq 0} \|\zeta(t)\|$. The space of all \mathcal{F}_t -measurable, square integrable random variables with values in H , denoted by $L_2^{\mathcal{F}}(\Omega, H)$, is a Hilbert space. Let $L_2^{\mathcal{F}_\tau}(\Omega, H)$ be the Hilbert space of all \mathcal{F}_τ -measurable, square integrable random variables with values in H .

Define $PC(J_1, L_2^{\mathcal{F}}(\Omega, H))$ as the Banach space of all piecewise continuous functions $y(t)$ from $J_1 = [-h, \tau]$ into $L_2^{\mathcal{F}}(\Omega, H)$ with the norm $\|y\| = (\sup_{t \in J_1} \mathbb{E}\|y(t)\|^2)^{1/2}$, where $\mathbb{E}(\cdot)$ is the expectation with respect to the measure \mathbf{P} . Let H_2 be the closed subspace of $PC(J_1, L_2^{\mathcal{F}}(\Omega, H))$ consisting of all \mathcal{F}_t -measurable H -valued processes $y(\cdot) \in PC(J_1, L_2^{\mathcal{F}}(\Omega, H))$ endowed with the norm:

$$\|y\|_{H_2} = \left(\sup_{t \in J_1} \mathbb{E}\|y(t)\|^2 \right)^{1/2}.$$

The purpose of this paper is to examine the approximate controllability for a class of nonlocal impulsive stochastic differential equation with time varying delay given by:

$$\begin{aligned} dy(t) &= [Ay(t) + Bu(t) + f(t, y(t - \mu(t)))]dt \\ &\quad + \sigma(t, y(t - \mu(t)))d\omega(t), \quad t \neq t_k, \quad t \in (0, \tau] \\ y(t) &= \varphi(t) + g(y)(t), \quad t \in [-h, 0] \\ \Delta y|_{t=t_k} &= I_k(y(t_k)), \quad k = 1, 2, \dots, m, \end{aligned} \tag{1}$$

where the state $y(t)$ is H -valued stochastic processes; A is the infinitesimal generator of a compact C_0 -semigroup $\{S(t) : t \geq 0\}$; the control function $u(\cdot)$ takes its values in $L_2^{\mathcal{F}}(J, U)$, the Hilbert space of admissible control functions; B is a bounded linear operator from U into H ; $f : [0, \tau] \times H \rightarrow H$ and $\sigma : [0, \tau] \times H \rightarrow L_2^0$ are the appropriate functions defined later; $g : PC((0, \tau], H) \rightarrow PC([-h, 0], H)$ is a continuous nonlinear function; $\varphi(\cdot) \in C_{\mathcal{F}_0}([-h, 0], H)$, the space of all \mathcal{F}_0 -measurable random variables independent of the Wiener process ω satisfying $\sup_{-h \leq t \leq 0} \mathbb{E}\|\varphi(t)\|^2 < \infty$; $I_k : H \rightarrow H, k = 1, 2, \dots, m$, are continuous functions and μ is a continuous function from \mathbb{R}^+ to $(0, h]$; Furthermore, let $0 = t_0 < t_1 < \dots < t_m < t_{m+1} = \tau$ be prefixed points,

and $\Delta y(t_k) = y(t_k^+) - y(t_k^-)$ represents the jump of the function y at t_k with I_k determining the size of the jump, where $y(t_k^+)$ and $y(t_k^-)$ are the right and left limits of $y(t)$ at $t = t_k$, respectively.

Definition 2.1 A continuous \mathcal{F}_t -adapted stochastic process $y : [-h, \tau] \rightarrow H$ is called a mild solution of the system (1) if for each $u(\cdot) \in L_2^{\mathcal{F}}(J, U)$, $y(t)$ is measurable with almost surely that $\int_{-h}^{\tau} \|y(s)\|^2 ds < \infty$ and the following stochastic integral equation is satisfied:

$$y(t) = \begin{cases} S(t)[\varphi(0) + g(y)(0)] + \int_0^t S(t-s)[Bu(s) + f(s, y(s - \mu(s)))]ds \\ + \int_0^t S(t-s)\sigma(s, y(s - \mu(s)))d\omega(s) \\ + \sum_{0 < t_k < t} S(t-t_k)I_k(y(t_k)), \quad t \in J, \\ \varphi(t) + g(y)(t), \quad t \in [-h, 0]. \end{cases}$$

Let $y(t; \varphi, u)$ denote the state value of the system (1) at time t corresponding to the control $u(\cdot) \in L_2^{\mathcal{F}}(J, U)$ and the initial value φ .

Introduce the set:

$$\mathfrak{R}(\tau; \varphi, u) = \{y(\tau; \varphi, u) : u(\cdot) \in L_2^{\mathcal{F}}(J, U)\}$$

which is called the reachable set of the system (1) at terminal time τ . Suppose that its closure in $L_2^{\mathcal{F}\tau}(\Omega, H)$ is denoted by $\overline{\mathfrak{R}(\tau; \varphi, u)}$.

Definition 2.2 The system (1) is said to be approximately controllable on J if and only if $\overline{\mathfrak{R}(\tau; \varphi, u)} = L_2^{\mathcal{F}\tau}(\Omega, H)$.

To discuss the approximate controllability of the system (1), define the operator $\mathcal{L}_\tau : L_2^{\mathcal{F}}(J, U) \rightarrow L_2^{\mathcal{F}\tau}(\Omega, H)$ by:

$$\mathcal{L}_\tau u = \int_0^\tau S(\tau-s)Bu(s)ds.$$

Clearly, the adjoint operator $\mathcal{L}_\tau^* : L_2^{\mathcal{F}\tau}(\Omega, H) \rightarrow L_2^{\mathcal{F}}(J, U)$ is given by:

$$\mathcal{L}_\tau^* z = B^* S^*(\tau-s)\mathbb{E}\{z|\mathcal{F}_t\}$$

where B^* and S^* denote the adjoint operators of B and S , respectively.

The linear controllability operator $\Pi_0^\tau : L_2^{\mathcal{F}\tau}(\Omega, H) \rightarrow L_2^{\mathcal{F}\tau}(\Omega, H)$ associated with the linear part corresponding to the system (1) is defined by:

$$\Pi_0^\tau\{\cdot\} = \mathcal{L}_\tau \mathcal{L}_\tau^*\{\cdot\} = \int_0^\tau S(\tau-t)BB^*S^*(\tau-t)\mathbb{E}\{\cdot|\mathcal{F}_t\}dt$$

and controllability operator corresponding to the linear deterministic control system is given by:

$$\Psi_t^\tau = \int_t^\tau S(\tau - s)BB^*S^*(\tau - s)ds.$$

Lemma 2.3 ([5]) *Let $\Phi : J \times \Omega \rightarrow L_2^0$ be a measurable mapping. Then for every $p \geq 2$, there exists $c_p > 0$ such that for every $t \geq 0$:*

$$\mathbb{E} \sup_{s \in [0,t]} \left\| \int_0^t \Phi(s)dw(s) \right\|^p \leq c_p \left[\int_0^t \left(\mathbb{E} \|\Phi(s)\|_{L_2^0}^p \right)^{2/p} ds \right]^{p/2}$$

where $c_p = (p(p - 1)/2)^{p/2}$. For $p = 2$, $c_p = 1$ and hence we have

$$\mathbb{E} \sup_{s \in [0,t]} \left\| \int_0^t \Phi(s)dw(s) \right\|^2 \leq \int_0^t \mathbb{E} \|\Phi(s)\|_{L_2^0}^2 ds. \tag{2}$$

We impose the following conditions to the system parameters:

- (H1) The operator A is the infinitesimal generator of a compact C_0 -semigroup $\{S(t) : t \geq 0\}$ in the Hilbert space H . Suppose $M \geq 1$ is such that $\|S(t)\|^2 \leq M$.
- (H2) The nonlinear functions $f : J \times H \rightarrow H$ and $\sigma : J \times H \rightarrow L_2^0$ satisfy linear growth and Lipschitz conditions. Moreover, there exist positive constants M_f , M_σ , N_f , and N_σ such that

$$\begin{aligned} \|f(t, x) - f(t, y)\|^2 &\leq M_f \|x - y\|^2, \quad \|f(t, y)\|^2 \leq N_f (1 + \|y\|^2), \\ \|\sigma(t, x) - \sigma(t, y)\|_{L_2^0}^2 &\leq M_\sigma \|x - y\|^2, \quad \|\sigma(t, y)\|_{L_2^0}^2 \leq N_\sigma (1 + \|y\|^2). \end{aligned}$$

- (H3) The function $g : PC((0, \tau], H) \rightarrow PC([-h, 0], H)$ satisfies linear growth and Lipschitz conditions. That is, there are positive constants M_g and N_g such that

$$\|g(x) - g(y)\|^2 \leq M_g \|x - y\|_{H_2}^2, \quad \|g(x)\|^2 \leq N_g (1 + \|y\|_{H_2}^2).$$

- (H4) The functions $I_k : H \rightarrow H$ are continuous and there exist positive constants p_k and \tilde{p}_k such that

$$\|I_k(x) - I_k(y)\|^2 \leq p_k \|x - y\|^2, \quad \|I_k(x)\|^2 \leq \tilde{p}_k (1 + \|x\|^2)$$

- (H5) For each $0 \leq t \leq \tau$, $\beta(\beta I + \Psi_t^\tau)^{-1}$ tends to zero in the strong operator topology as $\beta \rightarrow 0^+$. Observe that the linear deterministic system corresponding to the system (1) is approximately controllable on $[t, \tau]$ if and only if the operator $\beta(\beta I + \Psi_t^\tau)^{-1}$ tends to zero strongly as $\beta \rightarrow 0^+$ [6, Theorem 4.1].

Also without any loss of generality, we can assume that $\|(\beta I + \Psi_t^\tau)^{-1}\| \leq \frac{1}{\beta}$ for any $t \in J$.

3 Controllability Results

To define the control function, we need the following lemma. For more details, one can see [7, Lemma 3].

Lemma 3.1 *For any $\tilde{y}_\tau \in L_2^{\mathcal{F}_\tau}(\Omega, H)$, there exists a process $\tilde{\phi} \in L_2^{\mathcal{F}}(J, L_2^0)$ such that*

$$\tilde{y}_\tau = \mathbb{E}\tilde{y}_\tau + \int_0^\tau \tilde{\phi}(s)d\omega(s).$$

Now, for any $\beta > 0$ and $\tilde{y}_\tau \in L_2^{\mathcal{F}_\tau}(\Omega, H)$, define the control function:

$$\begin{aligned} u^\beta(t, y) = & B^*S^*(\tau - t)(\beta I + \Psi_0^\tau)^{-1} \left[\mathbb{E}\tilde{y}_\tau - S(\tau)[\varphi(0) + g(x)(0)] \right. \\ & \left. + \int_0^\tau (\beta I + \Psi_s^\tau)^{-1} \tilde{\phi}(s)d\omega(s) \right] \\ & - B^*S^*(\tau - t) \int_0^\tau (\beta I + \Psi_s^\tau)^{-1} S(\tau - s)f(s, y(s - \mu(s)))ds \\ & - B^*S^*(\tau - t) \int_0^\tau (\beta I + \Psi_s^\tau)^{-1} S(\tau - s)\sigma(s, y(s - \mu(s)))d\omega(s) \\ & - B^*S^*(\tau - t)(\beta I + \Psi_0^\tau)^{-1} \sum_{0 < t_k < \tau} S(\tau - t_k)I_k(y(t_k)). \end{aligned} \quad (3)$$

Lemma 3.2 *There exists a positive constant \hat{M} such that for all $x, y \in H_2$, we have*

$$\begin{aligned} \mathbb{E}\|u^\beta(t, x) - u^\beta(t, y)\|^2 & \leq \frac{\hat{M}}{\beta^2} \|x - y\|_{H_2}^2, \\ \mathbb{E}\|u^\beta(t, x)\|^2 & \leq \frac{\hat{M}}{\beta^2} (1 + \|x\|_{H_2}^2). \end{aligned}$$

Proof Let $x, y \in H_2$ be arbitrary. Then, we have

$$\begin{aligned} & \mathbb{E}\|u^\beta(t, x) - u^\beta(t, y)\|^2 \\ & \leq 4\mathbb{E} \left\| B^*S^*(\tau - t)(\beta I + \Psi_0^\tau)^{-1} S(\tau)[g(x)(0) - g(y)(0)] \right\|^2 \end{aligned}$$

$$\begin{aligned}
 &+ 4\mathbb{E}\|B^*S^*(\tau - t) \int_0^\tau (\beta I + \Psi_s^\tau)^{-1}S(\tau - s)[f(s, x(s - \mu(s))) \\
 &- f(s, y(s - \mu(s)))]ds\|^2 \\
 &+ 4\mathbb{E}\|B^*S^*(\tau - t) \int_0^\tau (\beta I + \Psi_s^\tau)^{-1}S(\tau - s)[\sigma(s, x(s - \mu(s))) \\
 &- \sigma(s, y(s - \mu(s)))]d\omega(s)\|^2 \\
 &+ 4\mathbb{E}\|B^*S^*(\tau - t)(\beta I + \Psi_0^\tau)^{-1}\sum_{0 < t_k < t}S(\tau - t_k)[I_k(x(t_k)) - I_k(y(t_k))]\|^2 \\
 &= 4(I_1 + I_2 + I_3 + I_4). \tag{4}
 \end{aligned}$$

Using assumptions (H1) and (H3), we get

$$I_1 \leq \frac{1}{\beta^2} \|B\|^2 M^2 M_g \|x - y\|_{H_2}^2.$$

The Cauchy-Schwarz inequality and assumptions (H1) and (H2) yield that

$$\begin{aligned}
 I_2 &\leq \|B\|^2 M \tau \int_0^\tau \mathbb{E}\|(\beta I + \Psi_s^\tau)^{-1}S(\tau - s)[f(s, x(s - \mu(s))) \\
 &- f(s, y(s - \mu(s)))]\|^2 ds \\
 &\leq \frac{1}{\beta^2} \|B\|^2 M^2 \tau \int_0^\tau M_f \mathbb{E}\|x(s - \mu(s)) - y(s - \mu(s))\|^2 ds \\
 &\leq \frac{1}{\beta^2} \|B\|^2 M^2 \tau \int_0^\tau M_f \sup_{s \in J} \mathbb{E}\|x(s - \mu(s)) - y(s - \mu(s))\|^2 ds \\
 &\leq \frac{1}{\beta^2} \|B\|^2 M^2 \tau^2 M_f \|x - y\|_{H_2}^2.
 \end{aligned}$$

Now using inequality (2) and assumptions (H1) and (H2), we obtain

$$\begin{aligned}
 I_3 &\leq \|B\|^2 M \mathbb{E} \sup_{s \in J} \left\| \int_0^\tau (\beta I + \Psi_s^\tau)^{-1}S(\tau - s)[\sigma(s, x(s - \mu(s))) \right. \\
 &- \sigma(s, y(s - \mu(s)))]d\omega(s) \|^2 \\
 &\leq \frac{1}{\beta^2} \|B\|^2 M^2 \int_0^\tau \mathbb{E}\|\sigma(s, x(s - \mu(s))) - \sigma(s, y(s - \mu(s)))\|_{L_2^2}^2 ds \\
 &\leq \frac{1}{\beta^2} \|B\|^2 M^2 \tau M_\sigma \|x - y\|_{H_2}^2.
 \end{aligned}$$

The assumptions (H1) and (H4) imply that

$$I_4 \leq \frac{1}{\beta^2} \|B\|^2 M^2 m \left(\sum_{k=1}^m p_k \right) \|x - y\|_{H_2}^2.$$

Therefore, (4) becomes

$$\begin{aligned} \mathbb{E} \|u^\beta(t, x) - u^\beta(t, y)\|^2 &\leq \frac{4}{\beta^2} \|B\|^2 M^2 \left[M_g + \tau^2 M_f + \tau M_\sigma + m \left(\sum_{k=1}^m p_k \right) \right] \|x - y\|_{H_2}^2 \\ &\leq \hat{M} \|x - y\|_{H_2}^2, \end{aligned}$$

where \hat{M} is a suitable positive constant. The second inequality can be established in the similar way. This completes the proof. \square

Consider the operator $\Phi_\beta : H_2 \rightarrow H_2$ defined by:

$$(\Phi_\beta y)(t) = \begin{cases} S(t)[\varphi(0) + g(y)(0)] + \int_0^t S(t-s)[Bu^\beta(s, y) + f(s, y(s - \mu(s)))]ds \\ + \int_0^t S(t-s)\sigma(s, y(s - \mu(s)))d\omega(s) \\ + \sum_{0 < t_k < t} S(t - t_k)I_k(y(t_k)), \quad t \in J \\ \varphi(t) + g(y)(t), \quad t \in [-h, 0]. \end{cases} \quad (5)$$

Now, the system (1) has a solution on J_1 if for all $\beta > 0$ there exists a fixed point of the operator Φ_β . For this purpose, the Banach contraction mapping principle is used.

Theorem 3.3 *If hypotheses (H1)–(H4) hold, then the system (1) has a unique mild solution on J_1 .*

Proof For every $y \in H_2$, the Cauchy-Schwarz inequality, hypothesis (H2)–(H4), and inequality (2) imply that

$$\begin{aligned} \mathbb{E} \|\Phi_\beta y\|_{H_2}^2 &\leq \sup_{t \in J_1} \mathbb{E} \|(\Phi_\beta y)(t)\|^2 \\ &\leq 6M[\mathbb{E}\|\varphi(0)\|^2 + N_g(1 + \|y\|_{H_2}^2)] \\ &\quad + \frac{6}{\beta^2} M\tau^2 \|B\|^2 \hat{M}(1 + \|y\|_{H_2}^2) + 6M^2\tau^2 N_f(1 + \|y\|_{H_2}^2) \\ &\quad + 6M\tau N_\sigma(1 + \|y\|_{H_2}^2) + 6mM^2 \left(\sum_{k=1}^m \tilde{p}_k \right) (1 + \|y\|_{H_2}^2) \\ &\leq \lambda_1 + \lambda_2 \|y\|_{H_2}^2, \end{aligned}$$

where λ_1 and λ_2 are suitable positive constants. Therefore, Φ_β maps H_2 into itself.

We now show that Φ_β is a contraction map on H_2 . For $x, y \in H_2$, hypothesis (H1)–(H4), inequality (2), and the Cauchy-Schwarz inequality yield that

$$\begin{aligned} \mathbb{E}\|(\Phi_\beta x)(t) - (\Phi_\beta y)(t)\|^2 &\leq 5[\mathbb{E}\|S(t)[g(x)(0) - g(y)(0)]\|^2 \\ &\quad + \mathbb{E}\|\int_0^t S(t-s)B[u^\beta(s, x) - u^\beta(s, y)]ds\|^2 \\ &\quad + \mathbb{E}\|\int_0^t S(t-s)[f(s, x(s - \mu(s))) - f(s, y(s - \mu(s)))]ds\|^2 \\ &\quad + \mathbb{E}\|\int_0^t S(t-s)[\sigma(s, x(s - \mu(s))) - \sigma(s, y(s - \mu(s)))]d\omega(s)\|^2 \\ &\quad + \mathbb{E}\|\sum_{0 < t_k < t} S(t - t_k)[I_k(x(t_k)) - I_k(y(t_k))]\|^2 \\ &\leq \gamma(\beta) \int_0^t \mathbb{E}\|x(s) - y(s)\|^2 ds, \end{aligned}$$

where $\gamma(\beta) = 5M \left[M_g + \frac{\|B\|^2 \tau^2 \hat{M}}{\beta^2} + M_f \tau^2 + M\sigma\tau + m \sum_{k=1}^m p_k \right]$. For any natural number n , mathematical induction yields that

$$\begin{aligned} \mathbb{E}\|\Phi_\beta^n x - \Phi_\beta^n y\|_{H_2}^2 &= \sup_{t \in J_1} \mathbb{E}\|(\Phi_\beta^n x)(t) - (\Phi_\beta^n y)(t)\|^2 \\ &\leq \frac{(\tau\gamma(\beta))^n}{n!} \|x - y\|_{H_2}^2 \end{aligned}$$

For any $\beta > 0$, there exists n such that $\frac{(\tau\gamma(\beta))^n}{n!} < 1$. It follows that Φ_β^n is a contraction mapping for sufficiently large n . Hence by the contraction mapping principle, Φ_β has a unique fixed point, say, y_β in H_2 , which is a solution to the system (1). □

Theorem 3.4 *Assume that the conditions in Theorem 3.3 and the assumption (H5) are satisfied. If f and σ are uniformly bounded, then the system (1) is approximately controllable on J_1 .*

Proof By Theorem 3.3, y_β is the fixed point of the operator Φ_β in H_2 . By substituting (3) into (5) and using the stochastic Fubini theorem [5], it can be easily seen that

$$\begin{aligned} y_\beta(\tau) &= y_\tau - \beta(\beta I + \Psi_0^\tau)^{-1} [\mathbb{E}\tilde{y}_\tau - S(\tau)(\varphi(0) + g(y_\beta)(0)) \\ &\quad - \sum_{0 < t_k < \tau} S(\tau - t_k)I_k(y_\beta(t_k))] \\ &\quad + \int_0^\tau \beta(\beta I + \Psi_s^\tau)^{-1} S(\tau - s)f(s, y_\beta(s - \mu(s)))ds \end{aligned}$$

$$+ \int_0^\tau \beta(\beta I + \Psi_s^\tau)^{-1} [S(\tau - s)\sigma(s, y_\beta(s - \mu(s))) - \tilde{\phi}(s)] d\omega(s). \quad (6)$$

Since f and σ are uniformly bounded, there is a subsequence still denoted by $\{f(s, y_\beta(s - \mu(s))), \sigma(s, y_\beta(s - \mu(s)))\}$ converging to, say, $\{f(s), \sigma(s)\}$ weakly in $H \times L_2^0$. Thus from (6), we obtain

$$\begin{aligned} \mathbb{E}\|y_\beta(\tau) - y_\tau\|^2 &\leq 6\|\beta(\beta I + \Psi_0^\tau)^{-1}\|[\mathbb{E}\tilde{y}_\tau - S(\tau)(\varphi(0) + g(y_\beta(0)) \\ &\quad - \sum_{0 < t_k < \tau} S(\tau - t_k)I_k(y_\beta(t_k)))]^2 \\ &\quad + 6\mathbb{E}\left(\int_0^\tau \|\beta(\beta I + \Psi_s^\tau)^{-1} S(\tau - s)[f(s, y_\beta(s - \mu(s))) - f(s)]\| ds\right)^2 \\ &\quad + 6\mathbb{E}\left(\int_0^\tau \|\beta(\beta I + \Psi_s^\tau)^{-1} S(\tau - s)f(s)\| ds\right)^2 \\ &\quad + 6\mathbb{E}\left(\int_0^\tau \|\beta(\beta I + \Psi_s^\tau)^{-1} S(\tau - s)[\sigma(s, y_\beta(s - \mu(s))) - \sigma(s)]\|_{L_2^0}^2 ds\right) \\ &\quad + 6\mathbb{E}\left(\int_0^\tau \|\beta(\beta I + \Psi_s^\tau)^{-1} S(\tau - s)\sigma(s)\|_{L_2^0}^2 ds\right) \\ &\quad + 6\mathbb{E}\left(\int_0^\tau \|\beta(\beta I + \Psi_s^\tau)^{-1} \tilde{\phi}(s)\|_{L_2^0}^2 ds\right) \end{aligned}$$

On the other hand, by (H5) for all $0 \leq s \leq \tau$, the operator $\beta(\beta I + \Psi_s^\tau)^{-1} \rightarrow 0$ strongly as $\beta \rightarrow 0^+$, and moreover $\|\beta(\beta I + \Psi_s^\tau)^{-1}\| < 1$. Thus, by the Lebesgue's dominated convergence theorem and the compactness of $S(\cdot)$, it follows that

$$\mathbb{E}\|y_\beta(\tau) - y_\tau\|^2 \rightarrow 0 \text{ as } \beta \rightarrow 0^+.$$

Therefore, $y_\beta(\tau) \rightarrow y_\tau$ holds in H and consequently we obtain the approximate controllability of the system (1). \square

4 Example

In this section, we give an example to show the applications of the obtained results. Consider the following stochastic control system of the form:

$$dz(t, y) = \left[\frac{\partial^2}{\partial y^2} z(t, y) + B\vartheta(t, y) + \frac{z(t - \sin t, y)}{1 + z(t - \sin t, y)} \right] dt$$

$$\begin{aligned}
 & + \frac{z(t - \sin t, y)}{1 + z(t - \sin t, y)} d\omega(t); \quad t_k \neq t \in (0, \tau], \quad 0 \leq y \leq \pi, \\
 & z(t, 0) = z(t, \pi) = 0, \quad t \in J, \\
 & z(\theta, y) = \varphi(\theta, y) + \int_0^\tau \ell(s, \theta) \cos(z(s, y)) ds, \quad -h \leq \theta \leq 0 \\
 & \Delta z(t_k, y) = I_k(z(t_k^-, y)), \quad k = 1, 2, \dots, m.
 \end{aligned} \tag{7}$$

Let $H = U = K = L_2[0, \pi]$ and $A : D(A) \subset H \rightarrow H$ be defined by $Aw = w''$ with domain $D(A) = \{z(\cdot) \in H : z, z' \text{ are absolutely continuous, } z'' \in H, z(0) = z(\pi) = 0\}$. Furthermore, A has discrete spectrum, the eigenvalues of A are $-n^2, n = 1, 2, \dots$, with the corresponding normalized eigenvectors $e_n(y) = \sqrt{2/\pi} \sin(ny)$. Then:

$$Az = - \sum_{n=1}^\infty n^2 \langle z, e_n \rangle e_n, \quad z \in D(A).$$

It is well known that A is the infinitesimal generator of a C_0 -semigroup $\{S(t) : t \geq 0\}$ on H and is given by:

$$S(t)z = \sum_{n=1}^\infty \exp(-n^2 t) \langle z, e_n \rangle e_n, \quad z \in H.$$

Now, define a continuous linear map B from

$$U = \left\{ u \mid u = \sum_{n=2}^\infty u_n e_n, \text{ with } \sum_{n=2}^\infty u_n^2 < \infty \right\}.$$

to H by:

$$Bu = 2u_2 e_1 + \sum_{n=2}^\infty u_n e_n \text{ for } u = \sum_{n=2}^\infty u_n e_n \in U.$$

Define $z(t)(y) = z(t, y), u(t)(y) = \vartheta(t, y)$ where $\vartheta(t, y) : J \times [0, \pi] \rightarrow [0, \pi]$ is continuous, $\mu(t) = \sin t, f(t, z(t))(y) = \frac{z(t,y)}{1+z(t,y)}, \sigma(t, z(t))(y) = \frac{z(t,y)}{1+z(t,y)}, g(z)(\theta)(y) = \int_0^\tau \ell(s, \theta) \sin(z(s, y)) ds$, and $\ell \in C(J \times [-h, 0], \mathbb{R})$. Thus, the stochastic control system (7) can be represented in the abstract form (1).

Now if $\|B^* S^*(t)z\| = 0, t \in J$, then it follows that

$$\|2z_1 e^{-t} + z_2 e^{-4t}\|^2 + \sum_{n=3}^\infty \|z_n e^{-n^2 t}\|^2 = 0$$

which yield that $z_n = 0$, $n = 1, 2, \dots$. Thus, we get $z = 0$. Therefore, the deterministic linear control system corresponding to the system (7) is approximately controllable on J (see [22, Theorem 4.1.7]).

Clearly, f and σ are uniformly bounded and satisfy assumption (H2). Also, the function g satisfies assumption (H3). If I_k , $k = 1, 2, \dots, m$ satisfy assumption (H4), then all the hypotheses of Theorem 3.4 are satisfied. Thus by Theorem 3.4, the system (7) is approximately controllable on J .

References

1. Kalman, R.E., Ho, Y.C., Narendra, K.S.: Controllability of linear dynamical systems, *Contrib. Differ. Equ.* 1, 189–213 (1963).
2. Quinn, M.D., Carmichael, N.: An approach to nonlinear control problem using fixed point methods, degree theory and pseudo-inverses, *Numer. Funct. Anal. Optim.* 7, 197–219 (1984).
3. Kumar, S., Sukavanam, N.: Controllability of second-order systems with nonlocal conditions in Banach spaces, *Numer. Funct. Anal. Optim.* 35, 423–431 (2014).
4. Triggiani, R.: A note on the lack of exact controllability for mild solutions in Banach spaces, *SIAM J. Control Optim.* 15, 407–411 (1977).
5. Da Prato, G., Zabczyk, J.: *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.
6. Mahmudov, N.I.: Controllability of linear stochastic systems in Hilbert spaces, *J. Math. Anal. Appl.* 259, 64–82 (2001).
7. Mahmudov, N.I.: Controllability of semilinear stochastic systems in Hilbert spaces, *J. Math. Anal. Appl.* 288, 197–211 (2003).
8. Mahmudov, N.I.: Approximate controllability of semilinear deterministic and stochastic evolution equations in abstract spaces, *SIAM J. Control Optim.* 42, 1604–1622 (2003).
9. Dauer, J.P., Mahmudov, N.I.: Controllability of stochastic semilinear functional differential equations in Hilbert spaces, *J. Math. Anal. Appl.* 290, 373–394 (2004).
10. Balachandran, K., Karthikeyan, S., Kim, J.H.: Controllability of semilinear stochastic integrodifferential systems, *Kybernetika*, 43, 31–44 (2007).
11. Matar, M.M.: On controllability of some stochastic semilinear integrodifferential systems in Hilbert spaces, *Int. Math. Forum*, 6, 1225–1239 (2011).
12. Karthikeyan, S., Balachandran, K., Sathya, M.: Controllability of nonlinear stochastic systems with multiple time-varying delays in control, *Int. J. Appl. Math. Comput. Sci.* 25, 207–215 (2015).
13. Sakthivel, R.: Approximate controllability of impulsive stochastic evolution equations, *Funkcial. Ekvac.* 52, 381–393 (2009).
14. Sakthivel, R., Nieto J.J., Mahmudov, N.I.: Approximate controllability of nonlinear deterministic and stochastic systems with unbounded delay, *Taiwanese J. Math.* 14(5), 1777–1797 (2010).
15. Shen, L., Sun, J.: Relative controllability of stochastic nonlinear systems with delay in control, *Nonlinear Anal.: Real World Applications*, 13, 2880–2887 (2012).
16. Shen, L., Sun, J.: Approximate controllability of abstract stochastic impulsive systems with multiple time-varying delays, *Int. J. Robust. Nonlinear Control*, 23, 827–838 (2013).
17. Shen, L., Wu, Q.: Approximate controllability of nonlinear stochastic impulsive systems with control acting on the nonlinear terms, *Int. J. Control*, 87(8), 1672–1680 (2014).
18. Ning, H., Qing, G. Approximate controllability of nonlinear stochastic partial differential systems with infinite delay, *Adv. Difference Equ.* 85, DOI 10.1186/s13662-015-0434-6, (2015).

19. Shukla, A., Arora, U., Sukavanam, N.: Approximate controllability of retarded semilinear stochastic system with non local conditions, *J. Appl. Math. Comput.* 49, 513–527 (2015).
20. Arora, U., Sukavanam, N.: Approximate controllability of impulsive semilinear stochastic system with delay in state *Stoch. Anal. Appl.* 34(6), 1111–1123 (2016).
21. Mokkedem, F.Z., Fu, X.: Approximate controllability for a semilinear stochastic evolution system with infinite delay in L_p space, *Appl. Math. Optim.* 75, 253–283 (2017).
22. Curtain, R.F., Zwart, H.: *An Introduction to Infinite-Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1995.

Convergence of an Operator Splitting Scheme for Abstract Stochastic Evolution Equations



Joshua L. Padgett and Qin Sheng

1 Introduction

Geometric integration techniques have received much attention in the study of differential equations [1, 3, 8, 16]. In particular, operator splitting methods have been shown to be effective and efficient numerical methods, as they may often be constructed to preserve stability while being explicit with desirable convergence rates [9, 10, 19, 20, 23, 24]. While splitting methods have primarily been studied in the deterministic setting, there have been several recent studies regarding their efficacy in application to stochastic problems [2, 17, 18, 21]. In particular, it has been shown that the splitting of deterministic and stochastic counterparts of differential equations can prove effective by increasing convergence rates without the inclusion of derivative terms [2, 5, 17]. Moreover, it is known that operator splitting methods may preserve many desirable geometric properties of the true solution, including the monotonicity and positivity [9, 12, 21].

Due to its wide range of applications in sciences and engineering, this chapter considers the following semi-linear stochastic differential equation problem:

$$du = [Au + f(u)]dt + g(u)dW, \quad 0 \leq t \leq T, \quad (1)$$

$$u(0) = u_0 \in H, \quad (2)$$

J. L. Padgett (✉)

Department of Mathematics and Statistics, Texas Tech University, Lubbock,
TX 79409-1042, USA

e-mail: joshua.padgett@ttu.edu

Q. Sheng

Department of Mathematics and Center for Astrophysics, Space Physics and Engineering
Research, Baylor University, Waco, TX 76798-7328, USA

e-mail: qin_sheng@baylor.edu

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_9

163

where H is a separable Hilbert space. In the above, $A : \text{Dom}(A) \subset H \rightarrow H$ is a linear operator whose domain is dense in H and compactly embedded into H . We will further assume that A generates an analytic semigroup e^{tA} , $t \geq 0$. The operators f and g are assumed to be Lipschitz continuous and possess continuous, uniformly bounded Fréchet derivatives up to order two. These assumptions and the precise analytic framework for (1)–(2) will be further outlined in Section 2. For technical reasons, we assume $u_0 \in H$ to be deterministic.

Without loss of generality, we let $N \in \mathbb{N}$ be fixed, and define $h = 1/N$. We are concerned with developing an approximation to the true solution to (1)–(2) at time $t_n = nh$, denoted u_n , being given by:

$$u_n = S^n(u_0), \quad (3)$$

where $S : H \rightarrow H$ is the nonlinear operator defined as:

$$S := e^{hA} e^{hf} e^{\Delta W(h)g}. \quad (4)$$

The nonlinear operator $v(t) = e^{hf}(v_0)$ is the solution to the differential equation $dv = f(v) dt$ at time h with initial condition $v(0) = v_0$, while $z(t) = e^{\Delta W(h)g}(z_0)$ is the solution to the stochastic differential equation $dz = g(z) dW$ at time h with initial condition $z(0) = z_0$. Such operators are often referred to as the nonlinear semigroup for each problem [14].

The splitting scheme given by (3) and (4) is classically known as the Lie-Trotter splitting scheme and has been well studied in numerous settings [9, 12, 13, 25]. Such methods have been studied in the finite-dimensional stochastic setting for ordinary differential equations via Lie algebraic techniques [2, 17, 18]. There has also been a recent study of such problems for linear equations with additive noise in UMD Banach spaces [5]. In this study, the optimal convergence rate was recovered, while the effects of nonlinearities were not included. However, the inclusion of nonlinear multiplicative noise terms complicates the required analysis and becomes one of the concerns of this current chapter.

This chapter is organized as follows. In Section 2, the abstract setting utilized throughout the chapter is detailed with several necessary results recalled. Section 3 outlines several basic properties regarding stability issues of the proposed operator splitting scheme. Section 4 is concerned with a detailed consistency analysis, while Section 5 demonstrates the desired convergence result.

2 Abstract Stochastic Evolution Problems

Let H be a separable Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \| = \langle \cdot, \cdot \rangle^{1/2}$. For another Hilbert space U equipped with norm $\| \cdot \|_U$, we denote by $L(U, H)$ the set of bounded linear operators from U to H . For the simplicity of notations, we let $L(U, U) = L(U)$. Further, we denote by $L_1(U, H)$ the set of

nuclear operators from U to H and $L_2(U, H)$ the set of Hilbert-Schmidt operators from U to H . Further, if $\{e_i\}_{i \in \mathbb{N}}$ forms an arbitrary orthonormal basis of H , then we have the following norms associated with the aforementioned spaces:

$$\| \Gamma \|_{L_1(U)} := \text{Tr} (\Gamma^* \Gamma)^{1/2} = \sum_{i=1}^{\infty} \langle (\Gamma^* \Gamma)^{1/2} e_i, e_i \rangle, \quad \Gamma \in L_1(U),$$

and

$$\| \Gamma \|_{L_2(U)}^2 := \sum_{i=1}^{\infty} \| \Gamma e_i \|_U^2, \quad \Gamma \in L_2(U),$$

where Γ^* denotes the adjoint of Γ . We further let $\mathbb{E} \| \cdot \|_{L_1(H)}$ and $\mathbb{E} \| \cdot \|_{L_2(H)}$ denote the corresponding expected values of each norm. Moreover, the trace and Hilbert-Schmidt norms are independent of the given basis.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with normal filtration $\{\mathcal{F}(t)\}_{t \geq 0}$, and let $W(t)$ be a standard Wiener process with covariance operator Q , where $Q : H \rightarrow H$ is a positive self-adjoint operator. If $q_i > 0$ are the eigenvalues of Q corresponding to eigenfunctions $e_i, i \in \mathbb{N}$, we then have

$$W(t) := \sum_{i \in \mathbb{N}} \sqrt{q_i} \beta_i(t) e_i, \quad 0 \leq t \leq T,$$

where $\{\beta_i\}_{i \in \mathbb{N}}$ are independent, real-valued Brownian motions on the probability space.

We denote the set of Hilbert-Schmidt operators from $Q^{1/2}(H)$ to H by $L_2^0(H)$ and its norm, for $\Gamma \in L_2^0(H)$, is given by:

$$\| \Gamma \|_{L_2^0(H)} := \| Q^{1/2} \Gamma \|_{L_2(H)} = \left(\sum_{i=1}^{\infty} \| Q^{1/2} (\Gamma^* \Gamma)^{1/2} e_i \|^2 \right)^{1/2}.$$

Now, let $\varphi : [0, T] \times \Omega \rightarrow L_2^0(H)$ be an $L_2^0(H)$ -valued predictable stochastic process with

$$\int_0^t \mathbb{E} \| Q^{1/2} \varphi \|_{L_2(H)}^2 ds < \infty, \quad 0 \leq t \leq T,$$

then Ito's isometry (see, for instance, [6]) gives

$$\mathbb{E} \left\| \int_0^t \varphi dW \right\|^2 = \int_0^t \mathbb{E} \| \varphi \|_{L_2^0(H)}^2 ds = \int_0^t \mathbb{E} \| Q^{1/2} \varphi \|_{L_2(H)}^2 ds, \quad 0 \leq t \leq T.$$

We now recall some basic properties of Hilbert space operators that will be of interest throughout this work.

Proposition 1 *Let $\Gamma, \Gamma_1, \Gamma_2$ be three operators in Hilbert spaces. Then, we have the following results.*

i. *If $\Gamma \in L_1(U)$, then*

$$|\text{Tr}(\Gamma)| \leq \|\Gamma\|_{L_1(U)}.$$

ii. *If $\Gamma_1 \in L(U)$ and $\Gamma_2 \in L_1(U)$, then both $\Gamma_1\Gamma_2$ and $\Gamma_2\Gamma_1$ belong to $L_1(U)$ with*

$$\text{Tr}(\Gamma_1\Gamma_2) = \text{Tr}(\Gamma_2\Gamma_1).$$

iii. *If $\Gamma_1 \in L(U, H)$ and $\Gamma_2 \in L(H, U)$, then $\Gamma_1\Gamma_2 \in L_1(H)$ with*

$$\|\Gamma_1\Gamma_2\|_{L_1(H)} \leq \|\Gamma_1\|_{L_2(U, H)}\|\Gamma_2\|_{L_2(H, U)}.$$

iv. *If $\Gamma \in L_2(U, H)$, then $\Gamma^* \in L_2(H, U)$ with*

$$\|\Gamma^*\|_{L_2(H, U)} = \|\Gamma\|_{L_2(U, H)}.$$

v. *If $\Gamma \in L(U, H)$ and $\Gamma_1, \Gamma_2 \in L_i(U)$, $i = 1, 2$, then $\Gamma\Gamma_1, \Gamma\Gamma_2 \in L_i(U, H)$, $i = 1, 2$, with*

$$\|\Gamma\Gamma_i\|_{L_j(U, H)} \leq \|\Gamma\|_{L(U, H)}\|\Gamma_i\|_{L_j(H)}, \quad i = 1, 2, \quad j = 1, 2.$$

More details on the proposition and the spaces used can be found in [4, 22].

We now outline several assumptions necessary for the existence, uniqueness, and well-posedness of the solution to (1)–(2).

Assumption 1 *The linear operator $A : \text{Dom}(A) \subset H \rightarrow H$ is the generator of a bounded C_0 semigroup e^{tA} , $t \geq 0$.*

Without loss of generality, by Assumption 1, it follows that we may assume that

$$\|e^{tA}\| \leq 1, \quad t \geq 0.$$

We now outline some basic properties of the semigroup generated by A (see, for instance, [11]).

Proposition 2 *Let $\alpha \geq 0$ and $0 \leq \gamma \leq 1$. Then, there exists a constant $C > 0$ such that:*

- i. $\|(-A)^\alpha e^{tA}\|_{L(H)} \leq Ct^{-\alpha}$, for $t > 0$,
- ii. $(-A)^\alpha e^{tA} = e^{tA}(-A)^\alpha$, on $\text{Dom}((-A)^\alpha)$,
- iii. If $\alpha \geq \gamma$, then $\text{Dom}((-A)^\alpha) \subset \text{Dom}((-A)^\gamma)$.

Recall (1). For nonlinear terms f and g , we need following restrictions.

Assumption 2 For the drift term $f : H \rightarrow H$, assume that there exists a positive constant $L_f > 0$ such that f satisfies the following Lipschitz condition:

$$\|f(u) - f(v)\| \leq L_f \|u - v\|, \quad \text{for all } u, v \in H.$$

This yields the following growth condition:

$$\|f(u)\| \leq C(1 + \|u\|), \quad \text{for all } u \in H.$$

We further assume that the derivatives $Df[u] : H \rightarrow H$ and $D^2 f[u] : H \times H \rightarrow H$ are continuous and uniformly bounded for all $u \in H$.

Assumption 3 For the diffusion term $g : H \rightarrow L_2^0(H)$, assume that there exists a positive constant $L_g > 0$ such that g satisfies the following Lipschitz condition:

$$\|g(u) - g(v)\|_{L_2^0(H)} \leq L_g \|u - v\|, \quad \text{for all } u, v \in H.$$

Similarly, the above leads to the growth condition:

$$\|g(u)\|_{L_2(H)} \leq C(1 + \|u\|), \quad \text{for all } u \in H.$$

We further assume that the derivatives $Dg[u] : H \rightarrow L_2^0(H)$ and $D^2 g[u] : H \times H \rightarrow L_2^0(H)$ are continuous and uniformly bounded for all $u \in H$.

In order to guarantee the existence of a well-defined mild solution to (1)–(2), we must also invoke a standard regularity assumption on the covariance operator of the noise W .

Assumption 4 Assume that there exist $\beta \in (0, 1]$ and $C > 0$ such that

$$\left\| (-A)^{(\beta-1)/2} Q^{1/2} \right\|_{L_2(H)} = \left\| Q^{1/2} (-A)^{(\beta-1)/2} \right\|_{L_2(H)} \leq C. \quad (5)$$

In the following analysis, any reference to a parameter β is the same β defined in (5).

If Assumptions 1–4 are satisfied and $u_0 \in H$ is \mathcal{F}_0 -measurable, then it follows that (1)–(2) admit a unique (up to the equivalence of paths) mild solution $u : [0, T] \times \Omega \rightarrow H$ with continuous sample path given by:

$$u(t) = e^{tA} u_0 + \int_0^t e^{(t-s)A} f(u(s)) ds + \int_0^t e^{(t-s)A} g(u(s)) dW(s), \quad \mathbb{P}\text{-a.s.}, \quad (6)$$

with the expectation:

$$\mathbb{E}\|u(t)\|^2 < \infty, \quad 0 \leq t \leq T, \quad (7)$$

(see [6]).

Let the Banach space $\text{Dom}((-A)^{\alpha/2})$, $\alpha \geq 0$, be equipped with the standard norm given by $\|\cdot\|_{\alpha} := \|(-A)^{\alpha/2} \cdot\|$. Then, we have the following regularity result for the solution to (1)–(2) [15].

Theorem 1 *Assume that Assumptions 1–4 hold. Let u be the mild solution to (1)–(2) given by (6). If $u_0 \in L^2(\Omega, \text{Dom}((-A)^{\alpha/2}))$, $\alpha \in [0, 1]$, then for all $0 \leq t \leq T$, $u \in L^2(\Omega, \text{Dom}((-A)^{\alpha/2})$ and*

$$\sup_{0 \leq t \leq T} \left(\mathbb{E} \|u(t)\|_{\alpha}^2 \right)^{1/2} \leq C \left(1 + \left(\mathbb{E} \|u_0\|_{\alpha}^2 \right)^{1/2} \right).$$

In addition, we employ two more assumptions.

Assumption 5 *We have that $\text{Dom}(A) \subset H$ and $\text{Dom}(A^2) \subset H$ are both invariant under f and g , with $\text{Dom}(A)$ also being invariant under Df and Dg , for all $u \in \text{Dom}(A)$.*

Assumption 6 *Let $\beta \in (0, 1]$ be defined as in (5). Then, we assume that there exists a constant $C > 0$ such that*

$$\|(-A)^{(\beta-1)/2} Dg[\xi](u - v)\|_{L^0_2(H)} \leq C \|u - v\|$$

and

$$\|(-A)^{(\beta-1)/2} D^2g[\xi](u - v)^2\|_{L^0_2(H)} \leq C \|u - v\|,$$

for all $\xi, u, v \in H$.

Assumption 6 initially appears to be restrictive. However, since $\beta \in (0, 1]$, the assumption actually allows for the derivatives Dg and D^2g to be slightly less regular.

Throughout this chapter, we will denote function and operator composition by left multiplication. That is, for two operators F_1 and F_2 , we use the standard notation:

$$F_1 F_2(u) = F_1(F_2(u)),$$

whenever the composition in consideration is well defined. Furthermore, throughout this chapter, we let $C > 0$ represent a generic constant independent of n and h . Note that this constant may assume different values throughout arguments.

In order to avoid repetition, it is henceforth assumed that Assumptions 1–6 hold throughout the remainder of the chapter. It is worth noting that Assumption 4 is quite standard and allows for the consideration of both space-time and trace class white noise. Space-time white noise corresponds to $Q = I$ and it is known that (5) is satisfied when $\beta < 1/2$, in the case of one spatial dimension. When considering trace class noise, that is when $\text{Tr}(Q) < \infty$, it follows that (5) is satisfied for $\beta = 1$

[7]. By considering trace class noise, we are able to recover the results presented in [17, 18].

3 Properties of the Splitting Operator

We first define the *least upper bound (lub) Lipschitz constant* and *(lub) logarithmic Lipschitz constant* of a function $F : H \rightarrow H$ by:

$$L[F] := \sup_{u \neq v} \frac{\|F(u) - F(v)\|}{\|u - v\|}$$

and

$$M[F] := \lim_{h \rightarrow 0^+} \frac{L[I + hF] - 1}{h},$$

respectively. For the following lemmas, we will consider the following problems:

$$dv = f(v) dt, \quad v(0) = v_0, \tag{8}$$

and

$$dv = g(v) dW, \quad v(0) = v_0. \tag{9}$$

Lemma 1 *Let $v(t) = e^{hf}(v_0)$ be the solution to (8). It then follows that*

$$L[e^{hf}] \leq e^{hL_f}.$$

Proof Let v and w be two distinct solutions to (8). Let D_t^+ denote the upper-right Dini derivative. Then, due to the assumptions on f and its derivatives, we have

$$\begin{aligned} D_t^+ \|v - w\| &= \limsup_{h \rightarrow 0^+} \frac{\|v(t+h) - w(t+h)\| - \|v(t) - w(t)\|}{h} \\ &\leq \lim_{h \rightarrow 0^+} \frac{\|(I + hf)(v(t) - w(t))\| - \|v(t) - w(t)\|}{h} \\ &\leq \lim_{h \rightarrow 0^+} \frac{L[I + hf]\|v(t) - w(t)\| - \|v(t) - w(t)\|}{h} \\ &\leq M[f]\|v(t) - w(t)\|. \end{aligned}$$

Solving the above inequality yields

$$\|v(t) - w(t)\| \leq e^{hM[f]}\|v_0 - w_0\|.$$

By the fact that f is Lipschitz continuous in H , we have

$$M[f] \leq L_f.$$

This yields the desired result. ■

For the following lemma, we mirror the approach employed in Lemma 1, but we need to consider slightly modified Lipschitz constants. To that end, we define the *lub stochastic Lipschitz constant* and *lub logarithmic stochastic Lipschitz constant* of a function $G : H \rightarrow L_2^0(H)$ by

$$\mathbb{E}[L_2[G]] := \sup_{u \neq v} \frac{\mathbb{E} \|G(u) - G(v)\|_{L_2^0(H)}^2}{\|u - v\|^2}$$

and

$$M_2[G] := \lim_{h \rightarrow 0} \frac{L_2[I + hG] - 1}{h},$$

respectively.

Lemma 2 *Let $v(t) = e^{\Delta W(h)g}(v_0)$ be the solution (9). It then follows that*

$$\mathbb{E} \left[L_2[e^{\Delta W(h)g}] \right] \leq e^{hL_g^2}.$$

Proof We proceed in a fashion similar to that of the previous proof. Let v and w be two distinct solutions to (9) and let D_t^+ denote the upper-right Dini derivative. Hence, we have

$$\begin{aligned} D_t^+ \mathbb{E} \|v - w\|^2 &= \limsup_{h \rightarrow 0^+} \frac{\mathbb{E} [\|v(t+h) - w(t+h)\|^2 - \|v(t) - w(t)\|^2]}{h} \\ &\leq \lim_{h \rightarrow 0^+} \frac{\mathbb{E} \left[\|(I + hg)(v(t) - w(t))\|_{L_2^0(H)}^2 - \|v(t) - w(t)\|^2 \right]}{h} \\ &\leq M_2[g] \mathbb{E} \|v(t) - w(t)\|^2. \end{aligned}$$

Note that deriving the second inequality follows from the fact that the remainder terms from the expansion are bounded. The details of this claim can be found in the proofs of Lemmas 4 and 5. Due to the expectation, the above inequality is deterministic and its solution is given by:

$$\mathbb{E} \|v(t) - w(t)\|^2 \leq e^{hM_2[g]} \mathbb{E} \|v_0 - w_0\|^2.$$

Once again, since g is Lipschitz in H , we have

$$M_2[g] = \lim_{h \rightarrow 0} \frac{1}{h} \left[\sup_{u \neq v} \frac{\mathbb{E} \|(1 + hg)(u - v)\|_{L_2^0(H)}^2}{\|u - v\|^2} - 1 \right] \leq L_g^2.$$

This yields the desired result. ■

Lemma 3 Consider (3) and (4). Then, we have

$$\mathbb{E} \|S^n(u) - S^n(v)\|^2 \leq C \mathbb{E} \|u - v\|^2,$$

and in particular,

$$\mathbb{E} [L_2[S]] \leq e^{hC}.$$

Proof By Lemmas 1 and 2, we readily have the following estimates:

$$\begin{aligned} \mathbb{E} \|S(u) - S(v)\|^2 &\leq \mathbb{E} \left[L[e^{hf}]^2 \|e^{\Delta W(h)g}(u - v)\|_{L_2^0(H)}^2 \right] \\ &\leq e^{2hL_f} \mathbb{E} \left[L_2[e^{\Delta W(h)g}](\|u - v\|)^2 \right] \\ &\leq e^{hC} \mathbb{E} \|u - v\|^2. \end{aligned} \tag{10}$$

Via iterations, it follows immediately that

$$\mathbb{E} \|S^n(u) - S^n(v)\|^2 \leq \prod_{j=0}^n e^{hC} \mathbb{E} \|u - v\|^2 \leq e^{TC} \mathbb{E} \|u - v\|^2,$$

which gives the desired result. ■

4 Approximation Consistency

Similar to discussions in [9] and [13], we define

$$\phi(t) := \frac{1}{t} \int_0^t e^{(t-s)A} f(T(u(s))) ds, \quad \psi(t) := \frac{1}{t} \int_0^t e^{(t-s)A} g(T(u(s))) dW(s),$$

where $T(u)$ is the solution operator for (1)–(2), and

$$T(u) = e^{hA}u + \int_0^h e^{(h-s)A} f(u) ds + \int_0^h e^{(h-s)A} g(u) dW(s), \quad \mathbb{P} - a.s. \tag{11}$$

Note that the operators are well defined and map H into itself for $u \in \text{Dom}(A)$.

Lemma 4 Assume that $u \in \text{Dom}(A)$. Then:

$$\mathbb{E}\|(T - S)(u)\|^2 \leq Ch^{2+\beta},$$

where $\beta \in (0, 1]$ is defined in (5).

Proof By appealing to the stochastic version of Taylor's theorem, we arrive at

$$\begin{aligned} S(u) &= e^{hA} e^{hf} e^{\Delta W(h)g} u \\ &= e^{hA} e^{\Delta W(h)g} u + h e^{hA} f e^{\Delta W(h)g} u + R_1(u) \\ &= e^{hA} u + h e^{hA} f(u + g(u)\Delta W(h) + R_2(u)) + e^{hA} g(u)\Delta W(h) + R_1(u) + R_2(u) \\ &= e^{hA} u + h e^{hA} f(u) + e^{hA} g(u)\Delta W(h) + R_1(u) + R_2(u) + R_3(u), \end{aligned} \quad (12)$$

where

$$\begin{aligned} \Delta W(h) &:= \int_0^h dW(s), \\ R_1(u) &:= h^2 \int_0^1 (1-s) Df[e^{shf} e^{\Delta W(h)g} u] f e^{shf} e^{\Delta W(h)g} u ds, \end{aligned} \quad (13)$$

$$\begin{aligned} R_2(u) &:= \int_0^h \int_0^1 (1-y) e^{hA} D^2 g[u + y(e^{\Delta W(s)g} u - u)] (e^{\Delta W(s)g} u - u)^2 dy dW(s) \\ &\quad + \int_0^h e^{hA} Dg[u](e^{\Delta W(s)g} u - u) dW(s), \end{aligned} \quad (14)$$

and

$$R_3(u) := h \int_0^1 e^{hA} Df[u + y(g(u)\Delta W(h) + R_2(u))] (g(u)\Delta W(h) + R_2(u)) dy. \quad (15)$$

Recall (11). We observe that

$$\begin{aligned} T(u) &= e^{hA} u + \int_0^h e^{(h-s)A} f(u) ds + \int_0^h e^{(h-s)A} g(u) dW(s) \\ &= e^{hA} u + \int_0^h e^{(h-s)A} f(e^{sA} u + s\phi(s) + s\psi(s)) ds \\ &\quad + \int_0^h e^{(h-s)A} g(e^{sA} u + s\phi(s) + s\psi(s)) dW(s) \end{aligned}$$

$$= e^{hA}u + \int_0^h e^{(h-s)A} f(e^{sA}u) ds + \int_0^h e^{(h-s)A} g(e^{sA}u) dW(s) + R_T(u), \quad (16)$$

where

$$R_T(u) := \int_0^h \int_0^1 e^{(h-s)A} Df[\xi(y; s)](s\phi(s) + s\psi(s)) dy ds + \int_0^h \int_0^1 e^{(h-s)A} Dg[\xi(y; s)](s\phi(s) + s\psi(s)) dy dW(s), \quad (17)$$

where $\xi(y; s) := e^{sA}u + y(s\phi(s) + s\psi(s)) \in H$. Combining the above yields

$$(T - S)(u) = \int_0^h [z_1(s) - z_1(0)] ds + \int_0^h [z_2(s) - z_2(0)] dW(s) + R(u), \quad (18)$$

where

$$z_1(s) := e^{(h-s)A} f(e^{sA}u), \quad z_2(s) := e^{(h-s)A} g(e^{sA}u)$$

and

$$R(u) := (R_T - R_1 - R_2 - R_3)(u).$$

Upon further expansion of (18), we obtain

$$(T - S)(u) = h \int_0^h z'_1(\xi_1) ds + h \int_0^h z'_2(\xi_2) dW(s) + R(u), \quad (19)$$

for some $\xi_1, \xi_2 \in [0, h]$.

Utilizing the above equalities and Ito's isometry, we acquire that

$$\begin{aligned} \mathbb{E}\|(T-S)(u)\|^2 &\leq 3\mathbb{E}\left\|h \int_0^h z'_1(\xi_1) ds\right\|^2 + 3\mathbb{E}\left\|h \int_0^h z'_2(\xi_2) dW(s)\right\|^2 + 3\mathbb{E}\|R(u)\|^2 \\ &= 3h^2 \int_0^h \mathbb{E}\|z'_1(\xi_1)\|^2 ds \\ &\quad + 3h^2 \int_0^h \mathbb{E}\|Q^{1/2}z'_2(\xi_2)\|_{L_2(H)}^2 ds + 3\mathbb{E}\|R(u)\|^2. \end{aligned} \quad (20)$$

It now remains to estimate each of the integrals in (20). To this end, we observe that

$$z'_1(\xi_1) = -Ae^{(h-\xi_1)A} f(e^{\xi_1 A}u) + e^{(h-\xi_1)A} Df[e^{\xi_1 A}u]e^{\xi_1 A}Au.$$

When $u \in \text{Dom}(A)$, it follows immediately that $z'_1 \in H$ and thus

$$\mathbb{E} \|z'_1(\xi_1)\|^2 \leq C.$$

By the same token, we have

$$z'_2(\xi_2) = -Ae^{(h-\xi_2)A}g(e^{\xi_2A}u) + e^{(h-\xi_2)A}Dg[e^{\xi_2A}u]e^{\xi_2A}Au \in H.$$

Thus:

$$\begin{aligned} & \mathbb{E} \left\| Q^{1/2} \left[-Ae^{(h-\xi_2)A}g(e^{\xi_2A}u) + e^{(h-\xi_2)A}Dg[e^{\xi_2A}u]e^{\xi_2A}Au \right] \right\|_{L_2(H)}^2 \\ & \leq 2\mathbb{E} \left\| Q^{1/2}(-A)e^{(h-\xi_2)A}g(e^{\xi_2A}u) \right\|_{L_2(H)}^2 \\ & \quad + 2\mathbb{E} \left\| Q^{1/2}e^{(h-\xi_2)A}Dg[e^{\xi_2A}u]e^{\xi_2A}Au \right\|_{L_2(H)}^2. \end{aligned}$$

Considering the first quantity in the above inequality, we find that

$$\begin{aligned} & \left\| Q^{1/2}(-A)e^{(h-\xi_2)A}g(e^{\xi_2A}u) \right\|_{L_2(H)}^2 \\ & \leq \left\| Q^{1/2}(-A)^{(\beta-1)/2} \right\|_{L_2(H)}^2 \left\| (-A)^{-\beta/2}e^{(h-\xi_2)A} \right\|_{L(H)}^2 \left\| g(e^{\xi_2A}u) \right\|_{L_2(H)}^2 \\ & \leq Ch^\beta(1 + \|u\|)^2, \end{aligned}$$

and thus by Theorem 1 and Lemma 3, we have

$$\mathbb{E} \left\| Q^{1/2}(-A)e^{(h-\xi_2)A}g(e^{\xi_2A}u) \right\|_{L_2(H)}^2 \leq Ch^\beta. \quad (21)$$

Considering the remaining quantity yields

$$\begin{aligned} & \left\| Q^{1/2}Ae^{(h-\xi_2)A}Dg[e^{\xi_2A}u]e^{\xi_2A}Au \right\|_{L_2(H)}^2 \\ & \leq \left\| Q^{1/2}(-A)^{(\beta-1)/2} \right\|_{L_2(H)}^2 \left\| (-A)^{-\beta/2}e^{(h-\xi_2)A} \right\|_{L(H)}^2 \left\| Dg[e^{\xi_2A}u]e^{\xi_2A}Au \right\|_{L_2(H)}^2 \\ & \leq Ch^\beta \left\| Dg[e^{\xi_2A}u]e^{\xi_2A}Au \right\|_{L_2(H)}^2 \leq Ch^\beta. \end{aligned} \quad (22)$$

Combining (21) and (22) with (20) yields

$$\mathbb{E}\|(T - S)(u)\|^2 \leq Ch^{2+\beta} + \mathbb{E}\|R(u)\|^2. \tag{23}$$

The desired result follows by applying the bound in Lemma 5 to (23). ■

Lemma 5 *Assume that $u \in \text{Dom}(A)$. Then:*

$$\mathbb{E}\|R(u)\|^2 \leq Ch^{2+\beta},$$

where $\beta \in (0, 1]$ is defined in (5).

Proof We now demonstrate that all terms in $R(u)$ have the expected error bounds. Recalling $R(u)$, we have

$$\mathbb{E}\|R(u)\|^2 \leq 4\mathbb{E}\|R_T(u)\|^2 + 4\mathbb{E}\|R_1(u)\|^2 + 4\mathbb{E}\|R_2(u)\|^2 + 4\mathbb{E}\|R_3(u)\|^2. \tag{24}$$

Let us estimate each of the terms in (24) individually. First, we observe that

$$\begin{aligned} \mathbb{E}\|R_T(u)\|^2 &\leq \mathbb{E} \left\| \int_0^h \int_0^1 e^{(h-s)A} Df[\xi(y; s)](s\phi(s) + s\psi(s)) dy ds \right\|^2 \\ &\quad + \mathbb{E} \left\| \int_0^h \int_0^1 e^{(h-y)A} Dg[\xi(y; s)](s\phi(s) + s\psi(s)) dy dW(s) \right\|_{L^2_2(H)}^2 \\ &\leq \int_0^h \int_0^1 \mathbb{E} \|Df[\xi(y; s)](s\phi(s) + s\psi(s))\|^2 dy ds, \\ &\quad + Ch^{\beta-1} \int_0^h \int_0^1 \mathbb{E} \|Dg[\xi(y; s)](s\phi(s) + s\psi(s))\|_{L^2_2(H)}^2 dy ds. \end{aligned}$$

and by recalling that Df and Dg are uniformly bounded in H , we obtain

$$\begin{aligned} \mathbb{E}\|R_T(u)\|^2 &\leq C \int_0^h \int_0^1 \mathbb{E} \|s\phi(s) + s\psi(s)\|^2 dy ds \\ &\quad + Ch^{\beta-1} \int_0^h \int_0^1 \mathbb{E} \|s\phi(s) + s\psi(s)\|_{L^2(H)}^2 dy ds \\ &\leq Ch^{2+\beta}. \end{aligned} \tag{25}$$

Recall (13) and (15). Due to the fact that Df is uniformly bounded, it is straightforward to show that

$$\mathbb{E}\|R_1(u)\|^2 \leq Ch^4 \quad \text{and} \quad \mathbb{E}\|R_3(u)\|^2 \leq Ch^3. \tag{26}$$

Finally, according to (14), by invoking Assumption 6 we have

$$\begin{aligned} & \mathbb{E} \|R_2(u)\|^2 \\ & \leq 2 \int_0^h \int_0^s \mathbb{E} \left\| e^{hA} D^2 g[u + y(e^{\Delta W^{(s)}g} u - u)] (e^{\Delta W^{(s)}g} u - u)^2 \right\|_{L^0_2(H)}^2 dy ds \\ & \quad + 2 \int_0^h \mathbb{E} \left\| e^{hA} Dg[u] (e^{\Delta W^{(s)}g} u - u) \right\|_{L^0_2(H)}^2 ds. \\ & \leq Ch^{\beta-1} \int_0^h \left[\int_0^s \mathbb{E} \|e^{\Delta W^{(s)}g} u - u\|^2 dy + \mathbb{E} \|e^{\Delta W^{(s)}g} u - u\|^2 \right] ds. \end{aligned}$$

By employing Lemma 6 in the above inequality, we obtain

$$\mathbb{E} \|R_2(u)\|^2 \leq Ch^{2+\beta}. \quad (27)$$

A combination of (25)–(27) yields our anticipated error bound. ■

Continuing, we may state the following estimate.

Lemma 6 *Let $0 < s < T$. Then, for $u \in H$, we have*

$$\mathbb{E} \left\| e^{\Delta W^{(s)}g} u - u \right\|^2 \leq Cs.$$

Proof By recalling (9), we see that

$$e^{\Delta W^{(s)}g} u = u + \int_0^s g(e^{\Delta W^{(y)}g} u) dW(y).$$

Thus, by Lemma 2, we have

$$\begin{aligned} \mathbb{E} \left\| e^{\Delta W^{(s)}g} u - u \right\|^2 &= \mathbb{E} \left\| \int_0^s g(e^{\Delta W^{(y)}g} u) dW(y) \right\|^2 \\ &= \int_0^s \mathbb{E} \left\| g(e^{\Delta W^{(y)}g} u) \right\|_{L^0_2(H)}^2 dy \leq Cs, \end{aligned}$$

which completes our proof. ■

5 Algorithmic Convergence

We now state our main result.

Theorem 2 Let $u_n = S^n(u_0)$, as defined in (3), be an approximation to the solution $u(nh) = T^n(u_0)$ of (1)–(2). If $u_0 \in \text{Dom}(A)$, then for h sufficiently small we have

$$\mathbb{E}\|(S^n - T^n)(u_0)\|^2 \leq Ch^\beta,$$

where $\beta \in (0, 1]$ is given in Assumption 4.

Proof Recall (11). It follows immediately that

$$\begin{aligned} T^n(u_0) &= e^{nhA}u_0 + \int_0^{nh} e^{(nh-s)A} f(u(s)) ds \\ &\quad + \int_0^{nh} e^{(nh-s)A} g(u(s)) dW(s), \quad \mathbb{P} - a.s. \end{aligned}$$

We now have the following representation of the difference:

$$(S^n - T^n)(u_0) = \sum_{j=0}^{n-1} \left(S^{n-j}T^j - S^{n-j-1}T^{j+1} \right) (u_0). \tag{28}$$

By taking the norm and expectation of (28), we observe that

$$\begin{aligned} \mathbb{E}\|(S^n - T^n)(u_0)\|^2 &= \mathbb{E} \left\| \sum_{j=0}^{n-1} \left(S^{n-j}T^j - S^{n-j-1}T^{j+1} \right) (u_0) \right\|^2 \\ &\leq (n-1) \sum_{j=0}^{n-1} \mathbb{E} \left[L[S^{n-j-1}] \right]^2 \mathbb{E}\|(S - T)(T^j(u_0))\|^2. \tag{29} \end{aligned}$$

If $u_0 \in \text{Dom}(A)$, then it follows that $T^j(u_0) \in \text{Dom}(A)$, $0 \leq j \leq n - 1$, due to Assumption 5. Therefore, we have

$$\mathbb{E}\|(S - T)(T^j(u_0))\|^2 \leq Ch^{2+\beta}, \tag{30}$$

for $0 \leq j \leq n - 1$. Recall Lemma 3. We find that

$$\mathbb{E} \left[L[S^{n-j-1}] \right]^2 \leq \mathbb{E} [L[S]]^{2(n-j-1)} \leq e^{2(n-j-1)hC}, \tag{31}$$

where C is independent of h , n , and j . Combining (30) and (31) gives

$$\mathbb{E}\|(S^n - T^n)(u_0)\|^2 \leq (n-1) \sum_{j=0}^{n-1} e^{2(n-j-1)hC} Ch^{2+\beta} \leq Ch^\beta.$$

■

From Theorem 2, we see that the maximal mean square convergence rate is given by $\beta/2$. Since $\beta \in (0, 1]$, it follows that the maximal convergence rate is $1/2$. Such a convergence rate is recovered when (1)–(2) is driven by trace class noise.

References

1. Blanes, S., Casas, F.: A concise introduction to geometrical numerical integration, 1st edn. CRC Press (2016)
2. Burrage, K., Burrage, P.M.: High strong order methods for non-commutative stochastic differential equations systems and the Magnus formula. *Physica D: Nonlinear Phenomena* **133**(1), 34–48 (1999)
3. Casas, F., Iserles, A.: Explicit Magnus expansions for nonlinear equations. *Journal of Physics A: Mathematical and General* **39**(19), 5445 (2006). URL <http://stacks.iop.org/0305-4470/39/i=19/a=S07>
4. Chow, P.L.: Stochastic partial differential equations. CRC Press (2014)
5. Cox, S., Van Neerven, J.: Convergence rates of the splitting scheme for parabolic linear stochastic Cauchy problems. *SIAM Journal on Numerical Analysis* **48**(2), 428–451 (2010)
6. Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions, 2 edn. *Encyclopedia of Mathematics and its Applications*. Cambridge University Press (2014). DOI 10.1017/CBO9781107295513
7. Debussche, A.: Weak approximation of stochastic partial differential equations: the nonlinear case. *Mathematics of Computation* **80**(273), 89–117 (2011)
8. Hairer, E., Lubich, C., Wanner, G.: Geometric numerical integration: structure-preserving algorithms for ordinary differential equations, vol. 31. Springer Science & Business Media (2006)
9. Hansen, E., Kramer, F., Ostermann, A.: A second-order positivity preserving scheme for semilinear parabolic problems. *Applied Numerical Mathematics* **62**(10), 1428–1435 (2012)
10. Hansen, E., Ostermann, A.: Dimension splitting for evolution equations. *Numerische Mathematik* **108**(4), 557–570 (2008). DOI 10.1007/s00211-007-0129-3. URL <https://doi.org/10.1007/s00211-007-0129-3>
11. Henry, D.: Geometric theory of semilinear parabolic equations, vol. 840. Springer (2006)
12. Iserles, A.: A First Course in the Numerical Analysis of Differential Equations, 2 edn. Cambridge Texts in Applied Mathematics. Cambridge University Press (2008). DOI 10.1017/CBO9780511995569
13. Jahnke, T., Lubich, C.: Error bounds for exponential operator splittings. *BIT Numerical Mathematics* **40**(4), 735–744 (2000)
14. Kato, T.: Nonlinear semigroups and evolution equations. *Journal of the Mathematical Society of Japan* **19**(4), 508–520 (1967)
15. Lord, G.J., Tambue, A.: Stochastic exponential integrators for the finite element discretization of SPDEs for multiplicative and additive noise. *IMA Journal of Numerical Analysis* **33**(2), 515–543 (2012)
16. Malham, S.J.A., Wiese, A.: Stochastic Lie group integrators. *SIAM J. Sci. Comput.* **30**(2), 597–617 (2008)
17. Misawa, T.: Numerical integration of stochastic differential equations by composition methods. *Sūrikaiseikikenkyūsho Kōkyūroku* **1180**, 166–190 (2000) [Dynamical systems and differential geometry (Japanese) (Kyoto, 2000)]

18. Misawa, T.: A Lie algebraic approach to numerical integration of stochastic differential equations. *SIAM J. Sci. Comput.* **23**(3), 866–890 (2006)
19. Padgett, J.L., Sheng, Q.: Modern Mathematical Methods and High Performance Computing in Science and Technology, chap. On the Stability of a Variable Step Exponential Splitting Method for Solving Multidimensional Quenching-Combustion Equations, pp. 155–167. Springer (2016)
20. Padgett, J.L., Sheng, Q.: On the positivity, monotonicity, and stability of a semi-adaptive LOD method for solving three-dimensional degenerate Kawarada equations. *J. Math. Anal. Appl.* **439**, 465–480 (2016)
21. Padgett, J.L., Sheng, Q.: Numerical solution of degenerate stochastic Kawarada equations via a semi-discretized approach. *Applied Mathematics and Computation* **325**, 210–226 (2018). DOI <https://doi.org/10.1016/j.amc.2017.12.034>. URL <https://www.sciencedirect.com/science/article/pii/S0096300317308937>
22. Prévôt, C., Röckner, M.: A concise course on stochastic partial differential equations, vol. 1905. Springer (2007)
23. Sheng, Q.: Solving linear partial differential equations by exponential splitting. *IMA Journal of numerical analysis* **9**(2), 199–212 (1989)
24. Sheng, Q.: Global error estimates for exponential splitting. *IMA Journal of Numerical Analysis* **14**(1), 27–56 (1994)
25. Trotter, H.: On the product of semi-groups of operators. *Proceedings of the American Mathematical Society* **10**(4), 545–551 (1959)

Modified Post-Widder Operators Preserving Exponential Functions



Vijay Gupta and Vinai K. Singh

1 Introduction

Post-Widder operators [10] are defined for $f \in C[0, \infty)$ as:

$$W_n(f, x) := \frac{1}{n!} \left(\frac{n}{x}\right)^{n+1} \int_0^{\infty} t^n e^{-\frac{nt}{x}} f(t) dt.$$

Approximation properties of these operators were recently studied in [4]. In [8], May considered slightly modified form of the Post-Widder operators while he discussed saturation and inverse results for exponential-type operators. The following form of Post-Widder operators was studied in [8] as:

$$P_n(f, x) := \frac{1}{(n-1)!} \left(\frac{n}{x}\right)^n \int_0^{\infty} t^{n-1} e^{-\frac{nt}{x}} f(t) dt. \quad (1)$$

These operators preserve constant and linear functions. Rempulska and Skorupka in [9] considered the modified form of the above Post-Widder operators P_n , which

V. Gupta (✉)

Department of Mathematics, Netaji Subhas University of Technology (Formerly Netaji Subhas Institute of Technology), Sector 3 Dwarka, New Delhi-110078, India
e-mail: vijaygupta2001@hotmail.com

V. K. Singh

Department of Applied Science and Humanities, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh, India
e-mail: drvinaiksingh@gmail.com

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_10

181

preserve the test function $e_2(x) = x^2$. It was observed in [9] that the modified form provides better approximation results over the original operators P_n . Also, some other approximation properties of Post-Widder operators have been discussed in the recent book by Gupta and Tachev [5]. The main motivation here is to study the modified form of the operators (1). For $a_n(x), b_n(x) > 0$, we start with the following modified form:

$$\tilde{P}_n(f, x) = \frac{1}{(n-1)!} \left(\frac{n}{a_n(x)} \right)^n \int_0^\infty t^{n-1} e^{-\frac{nt}{b_n(x)}} f(t) dt.$$

For $f(t) = e^{\theta t}$, $\theta \in R$ and $n > \theta b_n(x)$, we have

$$\tilde{P}_n(e^{\theta t}, x) = \frac{1}{(n-1)!} \left(\frac{n}{a_n(x)} \right)^n \int_0^\infty t^{n-1} e^{-\left(\frac{n}{b_n(x)} - \theta\right)t} dt.$$

Substituting $\left(\frac{n}{b_n(x)} - \theta\right)t = u$, we can write

$$\begin{aligned} \tilde{P}_n(e^{\theta t}, x) &= \frac{1}{(n-1)!} \left(\frac{n}{a_n(x)} \right)^n \frac{1}{\left(\frac{n}{b_n(x)} - \theta\right)^n} \int_0^\infty u^{n-1} e^{-u} du \\ &= \left(\frac{n}{a_n(x)} \right)^n \left(\frac{n}{b_n(x)} - \theta \right)^{-n}. \end{aligned} \quad (2)$$

Suppose that the modified Post-Widder operators preserve the test functions e^{ax} and e^{bx} , with the condition that $a_n(x) = b_n(x)$, if $a = b$, then using (2), we can write

$$\tilde{P}_n(e^{at}, x) = e^{ax} = \left(\frac{n}{a_n(x)} \right)^n \left(\frac{n}{b_n(x)} - a \right)^{-n},$$

and

$$\tilde{P}_n(e^{bt}, x) = e^{bx} = \left(\frac{n}{a_n(x)} \right)^n \left(\frac{n}{b_n(x)} - b \right)^{-n},$$

implying

$$a_n(x) = \frac{n(e^{(a-b)x/n} - 1)}{(a-b)e^{ax/n}}$$

and

$$b_n(x) = \frac{n(e^{(a-b)x/n} - 1)}{(-b + ae^{(a-b)x/n})}.$$

Obviously, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} a_n(x) &= x \\ \lim_{n \rightarrow \infty} b_n(x) &= x. \end{aligned}$$

Thus, our modified operators \tilde{P}_n take the following form:

$$\tilde{P}_n(f, x) := \frac{1}{(n-1)!} \frac{(a-b)^n e^{ax}}{(e^{(a-b)x/n} - 1)^n} \int_0^\infty t^{n-1} e^{-\frac{(-b+ae^{(a-b)x/n})t}{(e^{(a-b)x/n}-1)}} f(t) dt, \quad (3)$$

which preserves the test functions e^{ax} and e^{bx} and does not preserve constant if both a and b are nonzero. The preservation of constant function is possible if one of the values of a or b is zero. From the definition of P_n and \tilde{P}_n , it is immediate that two sequences are connected by the identity:

$$\tilde{P}_n(f, x) = \frac{(a-b)^n e^{ax}}{(-b + ae^{(a-b)x/n})^n} P_n\left(f, \frac{n(e^{(a-b)x/n} - 1)}{(-b + ae^{(a-b)x/n})}\right).$$

The most recent work on linear positive operators, which preserve exponential functions can be found in [1, 6] and [3]. Here in the present paper, we consider the modification of Post-Widder operators preserving exponential functions. We estimate direct results including quantitative asymptotic formula.

2 Auxiliary Results

We first present some lemmas required to prove main results of this section.

Lemma 1 *We have for $\theta > 0$ that*

$$\tilde{P}_n(e^{\theta t}, x) = (a-b)^n e^{ax} [(-b + ae^{(a-b)x/n}) - \theta(e^{(a-b)x/n} - 1)]^{-n}.$$

Lemma 2 *The r -th order moment $\mu_r^{\tilde{P}_n}(x) = \tilde{P}_n(e_r, x)$, where $e_r(t) = t^r$, $r \in \mathbb{N} \cup \{0\}$ are given by:*

$$\mu_r^{\tilde{P}_n}(x) = \frac{(n)_r \cdot (a-b)^n e^{ax} (e^{(a-b)x/n} - 1)^r}{(-b + ae^{(a-b)x/n})^{n+r}},$$

where the rising factorial $(n)_r = n(n + 1)(n + 2) \cdots (n + r - 1)$.

First few moments are given by:

$$\begin{aligned} \mu_0^{\tilde{P}_n}(x) &= \frac{(a - b)^n e^{ax}}{(-b + ae^{(a-b)x/n})^n} \\ \mu_1^{\tilde{P}_n}(x) &= \frac{n \cdot (a - b)^n e^{ax} (e^{(a-b)x/n} - 1)}{(-b + ae^{(a-b)x/n})^{n+1}} \\ \mu_2^{\tilde{P}_n}(x) &= \frac{n(n + 1) \cdot (a - b)^n e^{ax} (e^{(a-b)x/n} - 1)^2}{(-b + ae^{(a-b)x/n})^{n+2}} \\ \mu_3^{\tilde{P}_n}(x) &= \frac{n(n + 1)(n + 2) \cdot (a - b)^n e^{ax} (e^{(a-b)x/n} - 1)^3}{(-b + ae^{(a-b)x/n})^{n+3}} \\ \mu_4^{\tilde{P}_n}(x) &= \frac{n(n + 1)(n + 2)(n + 3) \cdot (a - b)^n e^{ax} (e^{(a-b)x/n} - 1)^4}{(-b + ae^{(a-b)x/n})^{n+4}}. \end{aligned}$$

Following [9], for (3), the result follows by simple computation.

Remark 1 An alternate approach to find moments is the moment-generating function which is $\tilde{P}_n(e^{\theta t}, x)$, and the moments are given by:

$$\begin{aligned} \mu_r^{\tilde{P}_n}(x) &= \left[\frac{\partial^r}{\partial \theta^r} \tilde{P}_n(e^{\theta t}, x) \right]_{\theta=0} \\ &= \left[\frac{\partial^r}{\partial \theta^r} \left\{ (a - b)^n e^{ax} [(-b + ae^{(a-b)x/n}) - \theta(e^{(a-b)x/n} - 1)]^{-n} \right\} \right]_{\theta=0}. \end{aligned}$$

Lemma 3 The central moments $T_r^{\tilde{P}_n}(x) = \tilde{P}_n((t - x)^r, x)$ are given below:

$$\begin{aligned} T_0^{\tilde{P}_n}(x) &= \frac{(a - b)^n e^{ax}}{(-b + ae^{(a-b)x/n})^n}, \\ T_1^{\tilde{P}_n}(x) &= \frac{n \cdot (a - b)^n e^{ax} (e^{(a-b)x/n} - 1)}{(-b + ae^{(a-b)x/n})^{n+1}} - \frac{x(a - b)^n e^{ax}}{(-b + ae^{(a-b)x/n})^n}, \\ T_2^{\tilde{P}_n}(x) &= \frac{n(n + 1) \cdot (a - b)^n e^{ax} (e^{(a-b)x/n} - 1)^2}{(-b + ae^{(a-b)x/n})^{n+2}} + \frac{x^2(a - b)^n e^{ax}}{(-b + ae^{(a-b)x/n})^n} \\ &\quad - \frac{2xn \cdot (a - b)^n e^{ax} (e^{(a-b)x/n} - 1)}{(-b + ae^{(a-b)x/n})^{n+1}}, \\ T_4^{\tilde{P}_n}(x) &= \frac{n(n + 1)(n + 2)(n + 3) \cdot (a - b)^n e^{ax} (e^{(a-b)x/n} - 1)^4}{(-b + ae^{(a-b)x/n})^{n+4}} \\ &\quad - 4x \frac{n(n + 1)(n + 2) \cdot (a - b)^n e^{ax} (e^{(a-b)x/n} - 1)^3}{(-b + ae^{(a-b)x/n})^{n+3}} \end{aligned}$$

$$\begin{aligned}
 &+6x^2 \frac{n(n+1) \cdot (a-b)^n e^{ax} (e^{(a-b)x/n} - 1)^2}{(-b + ae^{(a-b)x/n})^{n+2}} \\
 &-4x^3 \frac{n \cdot (a-b)^n e^{ax} (e^{(a-b)x/n} - 1)}{(-b + ae^{(a-b)x/n})^{n+1}} + x^4 \frac{(a-b)^n e^{ax}}{(-b + ae^{(a-b)x/n})^n}.
 \end{aligned}$$

Also, the following limits hold true:

$$\begin{aligned}
 \lim_{n \rightarrow \infty} T_0^{\tilde{P}_n}(x) &= x, \\
 \lim_{n \rightarrow \infty} nT_1^{\tilde{P}_n}(x) &= -\frac{(a+b)x^2}{2}, \\
 \lim_{n \rightarrow \infty} nT_2^{\tilde{P}_n}(x) &= x^2, \\
 \lim_{n \rightarrow \infty} n^2 T_4^{\tilde{P}_n}(x) &= 3x^4.
 \end{aligned}$$

In [2], a Korovkin-type theorem for the function e^{-kt} , $k = 0, 1, 2$, was considered for the class $C^*[0, \infty)$, which denote the linear space of real-valued continuous functions on $[0, \infty)$ with the property that $\lim_{x \rightarrow \infty} f(x)$ exists and is finite, equipped with uniform norm.

Theorem A ([7]) *Let $f \in C^*[0, \infty)$ and $L_n : C^*[0, \infty) \rightarrow C^*[0, \infty)$ be a sequence of linear positive operators and satisfies $\|L_n(e^{-it}, x) - e^{-ix}\|_\infty = a_i(n), i = 0, 1, 2$, where $a_i(n), i = 0, 1, 2$, tend to zero for n sufficiently large. Then, we have*

$$\|L_n f - f\|_\infty \leq \|f\|_\infty a_0(n) + (2 + a_0(n)) \cdot \omega^* \left(f, (a_0(n) + 2a_1(n) + a_2(n))^{1/2} \right),$$

where for every $\delta \geq 0$

$$\omega^*(f, \delta) = \sup_{\substack{x, t \geq 0 \\ |e^{-x} - e^{-t}| \leq \delta}} |f(x) - f(t)|.$$

3 Direct Results

The first result is the application of Theorem A to our operator (3):

Theorem 1 *For the sequence of modified Post-Widder operators $\tilde{P}_n : C^*[0, \infty) \rightarrow C^*[0, \infty)$, we have:*

1. For $a = 0$ and $b = -1$:

$$\|\tilde{P}_n f - f\|_{[0, \infty)} \leq 2\omega^*(f, \sqrt{a_2(n)}), f \in C^*[0, \infty).$$

2. For $a = 0$ and $b = -2$:

$$\|\tilde{P}_n f - f\|_{[0, \infty)} \leq 2\omega^*(f, \sqrt{2a_1(n)}), f \in C^*[0, \infty).$$

3. For $a = -1$ and $b = -2$, \tilde{P}_n is not an approximation method in $(C^*[0, \infty), \|\cdot\|_{[0, \infty)})$.

Proof

(1) In view of Lemma 1, we have

$$\tilde{P}_n(e^{\theta t}; x) = (1 - \theta(e^{x/n} - 1))^{-n}.$$

Obviously, if $\theta = 0, -1$, then we have $\tilde{P}_n(1; x) = 1$ and $\tilde{P}_n(e^{-t}; x) = e^{-x}$. Therefore by Theorem A, we immediately have $a_0(n) = a_1(n) = 0$. Let us consider $\theta = -2$, thus:

$$\tilde{P}_n(e^{-2t}; x) = (2e^{x/n} - 1)^{-n}$$

$$f_n(x) = (2e^{x/n} - 1)^{-n} - e^{-2x}, x \geq 0$$

Since $f_n(0) = f_n(\infty) = 0$, there exists a point $\xi_n \in (0, \infty)$ such that

$$\|f_n\| = f_n(\xi_n).$$

It follows that $f'_n(\xi_n) = 0$, that is:

$$e^{-2\xi_n} = e^{\frac{\xi_n}{n}} \left(2e^{\frac{\xi_n}{n}} - 1\right)^{-n-1}$$

and

$$f_n(\xi_n) = \left(2e^{\frac{\xi_n}{n}} - 1\right)^{-n} - e^{-2\xi_n} = \left(e^{\frac{\xi_n}{n}} - 1\right) \left(2e^{\frac{\xi_n}{n}} - 1\right)^{-1-n}.$$

Let $x_n := e^{\frac{\xi_n}{n}} - 1 > 0$. It follows that

$$f_n(\xi_n) = \frac{x_n}{(2x_n + 1)^{n+1}} \leq \min \left\{ x_n, \frac{1}{(2x_n + 1)^n} \right\} \rightarrow 0 \text{ as } n \rightarrow \infty.$$

This completes the proof of (1).

(2) Applying Lemma 1 for this case, we have

$$\tilde{P}_n(e^{\theta t}; x) = 2^n \left(2 - \theta(e^{2x/n} - 1)\right)^{-n}.$$

Obviously, if $\theta = 0, -2$, then we have $\tilde{P}_n(1; x) = 1$ and $\tilde{P}_n(e^{-2t}; x) = e^{-2x}$. Therefore by Theorem A, we immediately have $a_0(n) = a_2(n) = 0$. Let us consider $\theta = -1$, thus:

$$\tilde{P}_n(e^{-t}; x) = 2^n \left(e^{2x/n} + 1 \right)^{-n}$$

$$g_n(x) = 2^n \left(e^{2x/n} + 1 \right)^{-n} - e^{-x}, x \geq 0$$

Since $g_n(0) = g_n(\infty) = 0$, there exists a point $\zeta_n \in (0, \infty)$ such that

$$\|g_n\| = g_n(\zeta_n).$$

It follows that $g'_n(\zeta_n) = 0$, that is:

$$e^{-\zeta_n} = 2^{n+1} e^{\frac{2\zeta_n}{n}} \left(e^{\frac{2\zeta_n}{n}} + 1 \right)^{-n-1}$$

and

$$\begin{aligned} g_n(\zeta_n) &= 2^n \left(e^{2\zeta_n/n} + 1 \right)^{-n} - e^{-\zeta_n} = 2^n \left(e^{\frac{2\zeta_n}{n}} + 1 \right)^{-n-1} \left(1 - e^{\frac{2\zeta_n}{n}} \right) \\ &\leq \frac{2^n}{\left(e^{\frac{2\zeta_n}{n}} + 1 \right)^n} \end{aligned}$$

As $\zeta_n > 0$. It follows that

$$g_n(\zeta_n) \rightarrow 0 \text{ as } n \rightarrow \infty.$$

This completes the proof of (2).

- (3) Finally for $a = -1$ and $b = -2$, the operator \tilde{P}_n preserves e^{-x} and e^{-2x} and by Lemma 2, we have

$$\begin{aligned} \tilde{P}_n(1; x) &= e^{-x} \left(2 - e^{x/n} \right)^{-n} \\ &= 1 + x^2 \frac{1}{n} + \left(x^3 + \frac{x^4}{2} \right) \frac{1}{n^2} + \left(\frac{13x^4}{12} + x^5 + \frac{x^6}{6} \right) \frac{1}{n^3} \\ &\quad + \frac{1}{24} \left(30x^5 + 38x^6 + 12x^7 + x^8 \right) \frac{1}{n^4} + O[n^{-5}]. \end{aligned}$$

thus we do not have uniform convergence for $e_0 \in C^*[0, \infty)$.

Next, we prove the quantitative asymptotic formula.

Theorem 2 Let $f, f'' \in C^*[0, \infty)$, then, for $x \in [0, \infty)$, the following inequality holds:

$$\begin{aligned} & \left| n [\tilde{P}_n(f, x) - xf(x)] + \frac{(a+b)x^2}{2} f'(x) - \frac{x^2}{2} f''(x) \right| \\ & \leq |o_n(x)| |f(x)| + |p_n(x)| |f'(x)| + |q_n(x)| |f''(x)| \\ & \quad + \frac{1}{2} \left[2q_n(x) + x^2 + r_n(x) \right] \omega^*(f'', n^{-1/2}), \end{aligned}$$

where $o_n(x) := T_0^{\tilde{P}_n}(x) - x$, $p_n(x) := n T_1^{\tilde{P}_n}(x) + \frac{(a+b)x^2}{2}$, $q_n(x) := \frac{1}{2} (n T_2^{\tilde{P}_n}(x) - x^2)$ and $r_n(x) := [n^2 \tilde{P}_n((e^{-x} - e^{-t})^4, x)]^{1/2} \cdot [n^2 T_4^{\tilde{P}_n}(x)]^{1/2}$.

Proof By the Taylor’s formula, we have

$$f(t) = f(x) + (t-x)f'(x) + (t-x)^2 \frac{f''(\xi)}{2} + h(\xi, x)(t-x)^2,$$

where ξ lying between x and t and

$$h(\xi, x) := \frac{f''(\xi) - f''(x)}{2}$$

Applying the operator \tilde{P}_n to above equality and using Lemma 3, we can write that

$$\begin{aligned} & \left| \tilde{P}_n(f, x) - T_0^{\tilde{P}_n}(x)f(x) - T_1^{\tilde{P}_n}(x)f'(x) - \frac{1}{2} T_2^{\tilde{P}_n}(x)f''(x) \right| \\ & \leq \tilde{P}_n(|h(\xi, x)|(t-x)^2, x). \end{aligned}$$

Again using Lemma 3, we get

$$\begin{aligned} & \left| n [\tilde{P}_n(f, x) - xf(x)] + \frac{(a+b)x^2}{2} f'(x) - \frac{x^2}{2} f''(x) \right| \\ & \leq \left| T_0^{\tilde{P}_n}(x) - x \right| |f(x)| + \left| n T_1^{\tilde{P}_n}(x) + \frac{(a+b)x^2}{2} \right| |f'(x)| \\ & \quad + \frac{1}{2} \left| n T_2^{\tilde{P}_n}(x) - x^2 \right| |f''(x)| + \left| n \tilde{P}_n(h(\xi, x)(t-x)^2, x) \right|. \end{aligned}$$

Let $o_n(x) := T_0^{\tilde{P}_n}(x) - x$, $p_n(x) := n T_1^{\tilde{P}_n}(x) + \frac{(a+b)x^2}{2}$ and $q_n(x) := \frac{1}{2} (n T_2^{\tilde{P}_n}(x) - x^2)$.

Then:

$$\left| n [\tilde{P}_n(f, x) - xf(x)] + \frac{(a+b)x^2}{2} f'(x) - \frac{x^2}{2} f''(x) \right| \leq |o_n(x)| |f(x)| + |p_n(x)| |f'(x)| + |q_n(x)| |f''(x)| + \left| n \tilde{P}_n \left(h(\xi, x) (t-x)^2, x \right) \right|.$$

Also, from Lemma 3, we have $o_n(x) \rightarrow 0, p_n(x) \rightarrow 0$ and $q_n(x) \rightarrow 0$ for n sufficiently large. Now, we just have to compute the last estimate: $n \tilde{P}_n \left(h(\xi, x) (t-x)^2, x \right)$. Using the property of $\omega^*(\cdot, \delta) : |f(t) - f(x)| \leq \left(1 + \frac{(e^{-x} - e^{-t})^2}{\delta^2} \right) \omega^*(f, \delta), \delta > 0$, we get that

$$|h(\xi, x)| \leq \frac{1}{2} \left(1 + \frac{(e^{-x} - e^{-t})^2}{\delta^2} \right) \omega^*(f'', \delta).$$

Hence, we get

$$n \tilde{P}_n \left(|h(\xi, x)| (t-x)^2, x \right) \leq \frac{1}{2} n \omega^*(f'', \delta) T_2^{\tilde{P}_n}(x) + \frac{n}{2\delta^2} \omega^*(f'', \delta) \tilde{P}_n \left((e^{-x} - e^{-t})^2 (t-x)^2, x \right).$$

Applying Cauchy-Schwarz inequality, we obtain

$$n \tilde{P}_n \left(|h(\xi, x)| (t-x)^2, x \right) \leq \frac{1}{2} n \omega^*(f'', \delta) T_2^{\tilde{P}_n}(x) + \frac{n}{2\delta^2} \omega^*(f'', \delta) \left[\tilde{P}_n \left((e^{-x} - e^{-t})^4, x \right) \cdot T_4^{\tilde{P}_n}(x) \right]^{1/2}.$$

Considering

$$r_n(x) := \left[n^2 \tilde{P}_n \left((e^{-x} - e^{-t})^4, x \right) \right]^{1/2} \cdot \left[n^2 T_4^{\tilde{P}_n}(x) \right]^{1/2}.$$

and choosing $\delta = n^{-1/2}$, we finally get the desired result.

Remark 2 The convergence of modified Post-Widder operators \tilde{P}_n in the above theorem takes place for n sufficiently large. Using the software Mathematica, we find that

$$\begin{aligned} & \lim_{n \rightarrow \infty} n^2 \tilde{P}_n \left((e^{-x} - e^{-t})^4, x \right) \\ &= \lim_{n \rightarrow \infty} n^2 \left(\tilde{P}_n(e^{-4t}, x) - 4e^{-x} \tilde{P}_n(e^{-3t}, x) + 6e^{-2x} \tilde{P}_n(e^{-2t}, x) \right. \\ & \quad \left. - 4e^{-3x} \tilde{P}_n(e^{-t}, x) + e^{-4x} \right) \end{aligned}$$

$$\begin{aligned}
 &= \lim_{n \rightarrow \infty} n^2 \left((a-b)^n e^{ax} [(-b + ae^{(a-b)x/n}) + 4(e^{(a-b)x/n} - 1)]^{-n} \right. \\
 &\quad - 4e^{-x} (a-b)^n e^{ax} [(-b + ae^{(a-b)x/n}) + 3(e^{(a-b)x/n} - 1)]^{-n} \\
 &\quad + 6e^{-2x} (a-b)^n e^{ax} [(-b + ae^{(a-b)x/n}) + 2(e^{(a-b)x/n} - 1)]^{-n} \\
 &\quad - 4e^{-3x} (a-b)^n e^{ax} [(-b + ae^{(a-b)x/n}) + (e^{(a-b)x/n} - 1)]^{-n} \\
 &\quad \left. + e^{-4x} \frac{(a-b)^n e^{ax}}{(-b + ae^{(a-b)x/n})^n} \right) \\
 &= 3e^{-4x} x^4.
 \end{aligned}$$

4 Graphical Representation

This section deals with the graphical representation for some exponential function. We show through graphs the errors, using the software Mathematica. Let us consider the function $f(x) = e^{-4x}$ (Table 1).

- (i) For $n=20$, the approximation to the function f (Red color) by P_n (Gray), \tilde{P}_n [For $a = 0, b = -1$] (Black), and \tilde{P}_n [For $a = 0, b = -2$] (Blue) is illustrated in Figure 1:
- (ii) For $n=100$, the approximation to the function f (Red color) by P_n (Gray), \tilde{P}_n [For $a = 0, b = -1$] (Black), and \tilde{P}_n [For $a = 0, b = -2$] (Blue) is illustrated in Figure 2 (Table 2):

Conclusion. It is observed from the graphs and tables that if n is large, the error of approximation is reduced.

Table 1 Error of approximation for $P_n(f, x)$, \tilde{P}_n [For $a = 0, b = -1$] and \tilde{P}_n [For $a = 0, b = -2$] for $n = 20$

x	$ P_n(f, x) - f(x) $	$ \tilde{P}_n(f, x) - f(x) $ [For $a = 0, b = -1$]	$ \tilde{P}_n(f, x) - f(x) $ [For $a = 0, b = -2$]
1.5	0.00278303	0.00193077	0.00118913
1.8	0.00138755	0.000925397	0.000547718
2.0	0.000859734	0.000555895	0.000318939
2.2	0.000529645	0.000330751	0.000183254
2.4	0.000325609	0.000195621	0.000104262
2.6	0.00020031	0.000115328	0.0000588972
3.0	0.0000765738	0.0000400249	0.0000185426

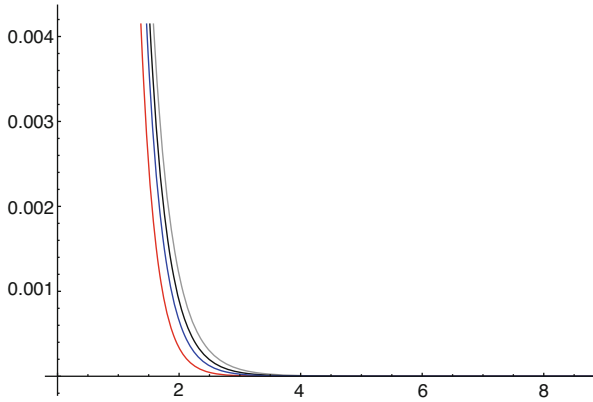


Fig. 1 The convergence of $P_n(f, x)$, \tilde{P}_n [For $a = 0, b = -1$] and \tilde{P}_n [For $a = 0, b = -2$] for $n = 20$ to $f(x)$

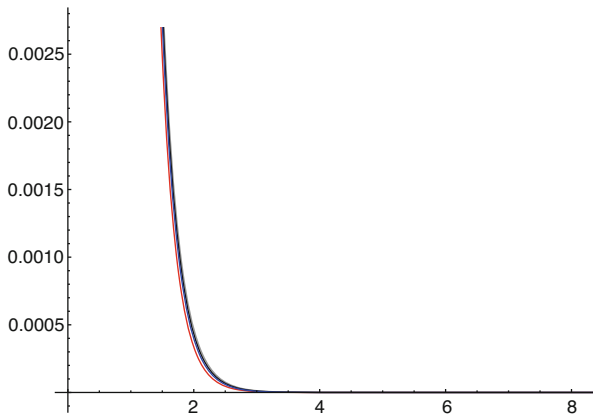


Fig. 2 The convergence of $P_n(f, x)$, \tilde{P}_n [For $a = 0, b = -1$] and \tilde{P}_n [For $a = 0, b = -2$] for $n = 100$ to $f(x)$

Table 2 Error of approximation for $P_n(f, x)$, \tilde{P}_n [For $a = 0, b = -1$] and \tilde{P}_n [For $a = 0, b = -2$] for $n = 100$

x	$ P_n(f, x) - f(x) $	$ \tilde{P}_n(f, x) - f(x) $ [For $a = 0, b = -1$]	$ \tilde{P}_n(f, x) - f(x) $ [For $a = 0, b = -2$]
1.5	0.000468474	0.000345398	0.000226351
1.8	0.000209554	0.000153159	0.0000994991
2.0	0.000119132	0.0000864777	0.0000557987
2.2	0.0000665923	0.0000479695	0.0000307167
2.4	0.00003673	0.0000262334	0.0000166568
2.6	0.0000200444	0.0000141821	0.0000089215
3.0	0.00000582864	0.00000403594	0.00000248561

References

1. T. Acar, A. Aral and H. Gonska, On Szász-Mirakyan operators preserving e^{2ax} , $a > 0$, H. Mediterr. J. Math. **14** (2017), no. 6, doi:10.1007/s00009-016-0804-7.
2. B. D. Boyanov and V. M. Veselinov, *A note on the approximation of functions in an infinite interval by linear positive operators*, Bull. Math. Soc. Sci. Math. Roum. **14** (1970), no. 62, 9–13.
3. V. Gupta and A. Aral, A note on Szász-Mirakyan-Kantorovich type operators preserving e^{-x} , Positivity **22**(2) (2018), 415–423. <https://doi.org/10.1007/s11117-017-0518-5>
4. V. Gupta and P. Maheshwari, Approximation with certain Post-Widder operators, Publ. Inst. Math.(Beograd), to appear.
5. V. Gupta and G. Tachev, Approximation with Positive Linear Operators and Linear Combinations, Series: Developments in Mathematics, Vol. **50** Springer, 2017
6. V. Gupta and G. Tachev, On approximation properties of Phillips operators preserving exponential functions, Mediterranean J. Math. **14** (4)(2017), Art. 177.
7. A. Holhoş, *The rate of approximation of functions in an infinite interval by positive linear operators*, Stud. Univ. Babeş-Bolyai, Math. **2**(2010), 133–142.
8. C. P. May, Saturation and inverse theorems for combinations of a class of exponential-type operators, Canad. J. Math., **28** (1976), 1224–1250.
9. L. Rempulska and M. Skorupka, On Approximation by Post-Widder and Stancu operators Preserving x^2 , Kyungpook Math. J. **49** (2009), 57-65.
10. D. V. Widder, The Laplace Transform, Princeton Mathematical Series, Princeton University Press, Princeton, N.J., 1941.

The Properties of Certain Linear and Nonlinear Differential Equations



Galina Filipuk and Alexander Chichurin

1 Introduction

The Schwarzian derivative is a differential operator that is invariant under all linear fractional transformations. It plays a significant role in the theory of modular forms, hypergeometric functions, univalent functions, and conformal mappings [1, 2]. It is defined by:

$$(Sf)(z) = \left(\frac{f''(z)}{f'(z)} \right)' - \frac{1}{2} \left(\frac{f''(z)}{f'(z)} \right)^2 = \frac{f'''(z)}{f'(z)} - \frac{3}{2} \left(\frac{f''(z)}{f'(z)} \right)^2.$$

The well-known relation between a second-order linear differential equation of the form:

$$y''(z) + Q(z)y(z) = 0$$

and the Schwarzian derivative of the ratio of two linearly independent solutions y_1, y_2 of the linear equation above is as follows:

$$(S\xi)(z) = 2Q(z),$$

G. Filipuk (✉)

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, Warsaw, 02-097, Poland

e-mail: G.Filipuk@mimuw.edu.pl

A. Chichurin

Institute of Mathematics and Computer Science, The John Paul II Catholic University of Lublin, ul. Konstantynów 1H, 20-708 Lublin, Poland

e-mail: achichurin@gmail.com

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41, https://doi.org/10.1007/978-3-030-02487-1_11

193

where $\xi = y_1/y_2$ and z is, in general, a complex variable. See [1, 2] for more details.

If we have a general second-order differential equation:

$$y''(z) + p(z)y'(z) + q(z)y(z) = 0, \quad (1)$$

then substituting $y(z) = \xi(z)y_1(z)$ with the condition that y_1 is also a solution of (1), we get an expression for ξ , y_1 and their derivatives (up to order 2 and 1, respectively). Differentiating again and eliminating y_1 , y_1' , we get that the function:

$$w(z) = (S\xi)(z)$$

satisfies

$$w(z) = \frac{1}{2}(4q(z) - p(z)^2 - 2p'(z)). \quad (2)$$

We call expression (2) the invariant for the second order linear differential equation (1).

In 1997, N.A. Lukashovich suggested the following method to study linear differential equations of the third order [3]. For a linear differential equation of the third order:

$$y'''(z) + p(z)y''(z) + q(z)y'(z) + r(z)y(z) = 0, \quad (3)$$

where the coefficients $p(z)$, $q(z)$, $r(z)$ are three times continuously differentiable functions with respect to the variable z , a similar procedure gives a second-order nonlinear differential equation [3–5] of the form:

$$(12w + b(z))w'' = 15w'^2 - h_0(z)w' - 8w^3 - h_1(z)w^2 - h_2(z)w - h_3(z). \quad (4)$$

Indeed, by putting

$$w(z) = S(\xi)(z),$$

where $\xi = y/y_1$ is the ratio of two linearly independent solutions of equation (3), we get equation (4) with

$$\begin{aligned} b &= 2(p^2 - 3q + 3p'), & h_1 &= 4b, \\ h_0 &= 8pp' - 6p'' - 12q' - 18pq + 4p^3 + 54r, \\ h_2 &= 12q'' - 8pp'' - 36r' + 10p'^2 - 24p'q + 12pq' + 2(p^2 - 3q)^2, \\ h_3 &= (q'' - 3r' + pr)b - (3q' + pq - 9r)(2p'' - q' + pq - 3r) \\ &\quad + 2(p' + p^2 - 2q)(pq' - p'q + q^2 - 3pr). \end{aligned} \quad (5)$$

In papers [3–5], the results are obtained for the Painlevé XXIII and XXV equations from the Ince list [6], which are nonlinear differential equations of the second order.

The generalization of the method for the linear differential equations of the fourth order is given in [7]. In papers [8–10], special classes of the fourth-order linear differential equations and the nonlinear fourth-order differential equations related via the Schwarzian derivative are considered and general solutions of both differential equations are found.

In paper [11], the generalization of the method for a special type of linear differential equations of the fifth order is given along with a computer realization of this method in *Mathematica* (www.wolfram.com).

Several questions arise. What happens if the second- and the third-order linear differential equations are related? What happens if we modify the function ξ to be the ratio of solutions of two different equations? The main objective of this paper is to answer these questions. We also note that the proofs of statements are computational; that is, the results can be verified by using any computer algebra system.

2 Main Results

In this section, we shall present several new results concerning relations between linear and nonlinear differential equations.

Theorem 1 *Let y be a solution of a third-order linear differential equation (3) and y_1 be a solution of a different third-order linear differential equation:*

$$y'''(z) + p_1(z)y''(z) + q_1(z)y'(z) + r_1(z)y(z) = 0.$$

If the function $w(z) = (S\xi)(z)$ with $\xi = y/y_1$ solves a nonlinear differential equation (4), then

$$p_1(z) = p(z), \quad q_1(z) = q(z), \quad r_1(z) = r(z)$$

and conditions (5) on b and h_i ($i = \overline{0, 3}$) hold.

Proof We substitute $w(z) = (S\xi)(z)$ into equation (4) with unknown coefficients and then replace ξ by the ratio of y and y_1 . Replacing the third- and higher-order derivatives of y and y_1 by using the linear equations, we collect the coefficients of y , y_1 , and their derivatives up to order 2. In the result, we obtain a system of equations on the coefficients of linear and nonlinear equations, from which we get the desired result. □

Theorem 2 *Let y be a solution of equation (3) and y_1 be a solution of a second-order linear differential equation of the form:*

$$y''(z) + q_1(z)y'(z) + r_1(z)y(z) = 0. \quad (6)$$

If the function $w(z) = (S\xi)(z)$ with $\xi = y/y_1$ solves the nonlinear differential equation (4), then we have conditions:

$$\begin{aligned} h_1 &= 4b, \quad b = 2(3p' + p^2 - 3q), \\ h_0 &= 2(-3(p'' + 2q') + p(4p' - 9q) + 2p^3 + 27r), \\ h_2 &= 2(p(6q' - 4p'') + 2p^2(p' - 3q) - 12qp' + 5p^2 + p^4 + 6q'' + 9q^2 - 18r'), \\ h_3 &= -3(3r - q')(-2p'' + q' + 3r) + p^2(-2qp' + 2q'' + q^2 - 6r') \\ &\quad + 2(q - p')(q(p' - 2q) - 3q'' + 9r') + 2p(q(-p'' - 3q' + 9r) + p'q') \\ &\quad + 2p^3(q' - 2r) \end{aligned}$$

on b and h_i ($i = \overline{0, 3}$) in (4) and two additional conditions on the coefficients of the linear equation (6):

$$r_1 = q - pq_1 + q_1^2 - q_1', \quad (7)$$

$$q_1'' = pq - p^2q_1 - qq_1 + 2p q_1^2 - q_1^3 - r - q_1p' + q' - 2pq_1' + 3q_1q_1'. \quad (8)$$

Proof We substitute $w(z) = (S\xi)(z)$ into equation (4) with unknown coefficients and then replace ξ by the ratio of y and y_1 . Replacing the third- and higher-order derivatives of y and the second- and higher-order derivatives of y_1 by using the linear equations, we collect the coefficients of y , y_1 , and their derivatives up to order 2 and order 1, respectively. In the result, we obtain a system of equations on the coefficients of linear and nonlinear equations, from which we get the desired result. \square

Theorem 3 *The third-order linear differential equation (3) has solutions satisfying (6) when the following conditions hold:*

$$r_1' = r - pr_1 + q_1r_1, \quad (9)$$

$$q_1' = q - pq_1 + q_1^2 - r_1. \quad (10)$$

Proof The proof is computational.

The next theorem says that if Theorem 3 holds (that is, solutions of the second-order linear equation are also solutions of the third-order linear equation), then the conditions of Theorem 2 are satisfied.

Theorem 4 *Conditions (7) and (8) are satisfied when conditions (9) and (10) hold.*

Proof We differentiate relation (9) and (10), and then substitute the obtained functions r_1'' and q_1'' , also r_1' and q_1' from equalities (9) and (10) into equations (7) and (8). After simplifications, we obtain two identities. \square

The symmetric square for the linear equation of second order is the third-order linear differential equation whose solutions are the products of the solutions of the second-order equation. More precisely, if y_1 and y_2 satisfy

$$y''(z) + q_1(z)y'(z) + q_2(z)y(z) = 0, \quad (11)$$

then $y = y_1 y_2$ satisfies

$$y'''(z) + 3q_1(z)y''(z) + (q_1'(z) + 4q_2(z) + 2q_1^2(z))y'(z) + 2(q_2'(z) + 2q_1(z)q_2(z))y(z) = 0. \quad (12)$$

The second-order linear equation has the invariant defined above. The third-order linear equation is connected to the nonlinear second-order equation. The following statement holds true.

Theorem 5 *The invariant $w = 2q_2 - q_1^2/2 - q_1'$ related to equation (11) solves the second-order nonlinear equation (4) for the symmetric square equation (12).*

Proof The proof is computational.

Theorem 6 *Assume that all solutions of the second-order linear equation (11) also solve the third-order linear equation (3). Then, the invariant for the second-order linear differential equation solves the second-order nonlinear differential equation (4) associated with (3).*

Proof From Theorem 3, we have

$$\begin{aligned} q &= pq_1 - q_1^2 + q_2 + q_1', \\ r &= pq_2 - q_1q_2 + q_2'. \end{aligned}$$

Next, we recalculate the nonlinear equation (4) for the third-order equation (3) with the conditions above and show that the invariant for the second-order equation solves it.

In particular, if we differentiate the second-order differential equation, we have the statement above. \square

Let us next replace y with y' in the second-order linear differential equation (11). In the result, we have the third-order linear equation of the form:

$$y'''(z) + q_1(z)y''(z) + q_2(z)y'(z) = 0. \quad (13)$$

Note that the second-order equation (11) does not solve it unless q_1 and q_2 are constants.

Theorem 7 *The invariant associated to the second-order equation (11) solves the second-order nonlinear differential equation (4) associated with (13) when*

$$\begin{aligned} & (6q_1' - 18q_2 + 4q_1^2)q_1''' - 9q_1''^2 - 12q_1^2q_2'' + 54q_2q_2'' \\ & + (4q_1^3 - 18q_1q_2 - 18q_1q_1' + 54q_2')q_1'' + (36q_2^2 + 78q_1q_2' - 18q_2'' - 8q_1^2q_2)q_1' \\ & - (8q_1^2 + 48q_2)q_1'^2 + 12q_1^3 + 4q_1^3q_2' - 18q_1q_2q_2' - 81q_2'^2 = 0. \end{aligned}$$

Proof Consider differential equation (4), (5) associated with (13). Then, the following relations on the coefficients hold:

$$p(z) = q_1(z), \quad q(z) = q_2(z), \quad r(z) = 0. \quad (14)$$

Next, we differentiate twice the invariant associated to the second-order equation (11). In the result, we obtain

$$w' = \frac{1}{2}(4q_2 - q_1^2 - 2q_1'), \quad w'' = 2q_2' - q_1q_1' - q_1'', \quad w''' = -q_1^{(3)} - q_1q_1'' - q_1'^2 + 2q_2''$$

and then we substitute these three functions into equations (4) and (5) with coefficients (14). After simplifications, we obtain an identity. \square

Finally, let us try to find an invariant associated to the first-order nonlinear differential equation of Riccati type:

$$y'(z) = a(z)y(z)^2 + b(z)y(z). \quad (15)$$

Let us take the ratio of two solutions of this equation and compute the Schwarzian derivative. Then, we have

$$w(z) = (S\xi)(z) = -\frac{1}{2}b(z)^2 - \frac{a'(z)}{a(z)}b(z) - \frac{3}{2}\frac{a'^2(z)}{a^2(z)} + \frac{a''(z)}{a(z)} + b'(z).$$

where $\xi = y/y_1$ and y, y_1 solve (15).

The general Riccati equation:

$$y'(z) = a(z)y(z)^2 + b(z)y(z) + c(z) \quad (16)$$

is linearizable. Indeed, substituting

$$y(z) = -\frac{1}{a(z)} \frac{v'(z)}{v(z)},$$

we get the following second-order linear differential equation:

$$v''(z) - \left(\frac{a'(z)}{a(z)} + b(z) \right) v'(z) + a(z)c(z)v(z) = 0.$$

Clearly,

$$v(z) = C \exp \left(\int -a(z)y(z)dz \right), \tag{17}$$

where C is an arbitrary constant. Since the invariant for the second-order linear differential equation is known, we can associate the following invariant for the general Riccati equation (16):

$$w(z) = -\frac{1}{2}b(z)^2 + 2a(z)c(z) - \frac{a'(z)}{a(z)}b(z) - \frac{3}{2}\frac{a'^2(z)}{a^2(z)} + \frac{a''(z)}{a(z)} + b'(z).$$

Indeed, substituting $w(z) = (S\xi)(z)$, where $\xi = v/v_1$ is the ratio of two solutions of the second-order linear equation related to the Riccati equation, and taking into account (17), we get the desired expression.

We also remark that if $y'(z) = y(z)^2/2 - f(z)$ (which is linearizable to the equation $u''(z) + f(z)/2u(z) = 0$ by using $y(z) = -2u'(z)/u(z)$ and $y(z) = w''(z)/w'(z)$, then $(Sw)(z) = f(z)$.

3 Discussion

It is interesting to obtain a discrete/difference analogue of the main results of this paper. It is an open problem to obtain a difference operator that has similar to the Schwarzian derivative invariance properties. One more research direction is to replace linear differential equations with nonlinear equations of second and higher order and to consider the Schwarzian derivative of the ratio of 2 solutions. This might give a new insight into the theory of some nonlinear special functions.

Acknowledgements GF is grateful to the organizers of the conference *Modern Mathematical Methods and High Performance Computing in Science and Technology—2018* for their invitation and the opportunity to present a recorded video lecture.

GF also acknowledges the support of the Alexander von Humboldt Foundation and the support of NCN OPUS 2017/25/B/BST1/00931.

References

1. Ahlfors, L. V.: Möbius transformations in several dimensions. Lecture notes at the University of Minnesota, Minneapolis (1981)
2. Dobrovolsky, V.A.: Essays on the development of the analytic theory of differential equations. Vishcha Shkola, Kiev (1974) (in Russian)

3. Lukashevich, N. A.: On the third order linear equations. *Differ. Equ.*, 35, 1384–1390 (1999)
4. Lukashevich, N. A., Martynov, I. P.: On the third order linear equations. Materials of the International Scientific Conference "Differential Equations and Their Applications", Grodno, Grodno State University, 78–85 (1998) (in Russian)
5. Lukashevich, N. A., Chichurin, A.V.: Differential equations of the first order. Belarusian State University, Minsk (1999) (in Russian)
6. Ince, E. L.: Ordinary differential equations. Dover Publications, New York (1956)
7. Chichurin, A.V.: The Chazy equation and linear equations of the Fuchs class. Publ. RUDN, Moscow (2003) (in Russian)
8. Chichurin, A., Stepaniuk G.: The computer method of construction of the general solution of the linear differential equation of the third order. *Studia i Materialy EWSIE*, 8, 17–27 (2014)
9. Chichurin, A.V., Stepaniuk, G.P.: General solution to the fourth order linear differential equation with coefficients satisfying the system of three first order differential equations. *Bulletin of Taras Shevchenko National University of Kyiv, Series: Physics & Mathematics*, 31, No.1, 29–34 (2014)
10. Chichurin, A.V.: The computer method of construction of the general solution of the nonlinear differential equation of the fourth order. *Studia i Materialy EWSIE*, 7, 39–47 (2014)
11. Chichurin, A.: Computer construction of the general solution of the linear differential equation of the fifth order. *Recent Developments in Mathematics and Informatics, Vol. I. Contemporary Mathematics and Computer Science*, KUL, Lublin, 19-36 (2016)

Fixed Points for (ϕ, ψ) -Contractions in Menger Probabilistic Metric Spaces



Vandana Tiwari and Tanmoy Som

1 Introduction

The concept of probabilistic metric space was initiated by Menger [13] in 1942. Menger probabilistic metric space (briefly, Menger PM-space) is a generalization of metric space in which distance between two points x and y , $d(x, y)$, is assigned by a distribution function $F_{x,y}$. Since then, many researchers extensively developed and expanded the study of PM-spaces in their pioneering works, e.g., [5, 19–23].

To prove existence and uniqueness of fixed point theorem in Menger PM-spaces, contraction is one of the basic tools. Sehgal and Bharucha-Reid [21] introduced probabilistic k -contraction and proved probabilistic version of classical Banach fixed point principle. Efforts have been made over the years to generalize and extend the k -contraction, like that the concept of φ -contraction, weak contraction, and generalized weak contraction, etc. in Menger PM-spaces. Some examples from the large existing results are [7–9, 16]. In other spaces, which are generalizations of usual metric spaces, such problems are also addressed by several authors, see [1–4, 6, 10–12, 14, 15, 18, 24]. Motivated by the very recent result [17], here, we have established the existence of a fixed point of a self-mapping which satisfies the (ϕ, ψ) -contraction.

Now, we give some basic definitions and preliminaries which will be needed to establish our main result. From here onward, let $\mathbb{R} = (-\infty, +\infty)$, $\mathbb{R}^+ = [0, +\infty)$, and \mathbb{N} be the set of all natural numbers.

V. Tiwari · T. Som (✉)

Department of Mathematical Sciences, Indian Institute of Technology (BHU), Varanasi-221005, U.P., India

e-mail: vandanatiwari.rs.apm12@itbhu.ac.in; tsom.apm@itbhu.ac.in

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41, https://doi.org/10.1007/978-3-030-02487-1_12

201

Definition 1 ([13]) A mapping $F : \mathbb{R} \rightarrow [0, 1]$ is called a distribution function if it is nondecreasing and left continuous with $\inf_{t \in \mathbb{R}} F(t) = 0$. If in addition, $F(0) = 0$, then F is called a distance distribution function.

Definition 2 ([13]) A distance distribution function F , satisfying $\lim_{t \rightarrow \infty} F(t) = 1$, is called a Menger distance distribution function. The set of all Menger distance distribution functions is denoted by D^+ . The space D^+ is partially ordered by the usual point-wise ordering of functions, that is, $F \leq G$ if and only if $F(t) \leq G(t)$, for all $t \in [0, \infty]$. The maximal element for D^+ in this order is the distance distribution function H , given by:

$$H(t) = \begin{cases} 0, & \text{if } t \leq 0 \\ 1, & \text{if } t > 0. \end{cases}$$

Definition 3 ([19]) A continuous t -norm T is a binary operation on $[0, 1]$, which satisfies the following conditions:

- (1) T is associative and commutative,
- (2) $T(a, 1) = a$, for all $a \in [0, 1]$,
- (3) $T(a, b) \leq T(c, d)$, whenever $a \leq c$ and $b \leq d$, for all $a, b, c, d \in [0, 1]$,
- (4) T is continuous.

For example:

- (a) The minimum t -norm, T_M , defined by $T_M(a, b) = \min\{a, b\}$; and (b) the product t -norm, T_P , defined by $T_P(a, b) = a \cdot b$, are two basic t -norms.

In 1942, Menger developed the theory of metric spaces and proposed a generalization of metric spaces called Menger probabilistic metric spaces (briefly, Menger PM-space).

Definition 4 ([13, 20]) A Menger PM-space is a triplet (X, F, T) , where X is a non-empty set, T is a t -norm, and $F : X \times X \rightarrow D^+$ be a mapping satisfying the following conditions (for $x, y \in X$, we denote $F(x, y)$ by $F_{x,y}$):

- (1) $F_{x,y}(t) = H(t)$, for all $x, y \in X$ and $t > 0$ if and only if $x = y$;
- (2) $F_{x,y}(t) = F_{y,x}(t)$, for all $x, y \in X$ and $t > 0$;
- (3) $F_{x,y}(s + t) \geq T(F_{x,z}(s), F_{z,y}(t))$, for all $x, y, z \in X$ and $s, t > 0$.

Let (X, F, T) be a Menger PM-space. Define (ϵ, λ) -neighborhood of $x \in X$ as $U_x(\epsilon, \lambda) = \{y \in X : F_{x,y}(\epsilon) > 1 - \lambda\}$, that is, the set of all points y in X for which the probability of the distance from x to y being less than ϵ is greater than $1 - \lambda$. Then, (ϵ, λ) -topology τ , induced by the family of (ϵ, λ) -neighborhoods $\{U_x(\epsilon, \lambda) : \epsilon > 0, \lambda \in (0, 1]\}$, is Hausdorff topology, if t -norm T satisfies $\limsup_{0 < t < 1} T(t, t) = 1$.

Definition 5 ([13, 20]) Let (X, F, T) be a Menger PM-space.

1. A sequence $\{x_n\}$ in (X, F, T) is said to converge to a point $x \in X$, written as $x_n \rightarrow x$, if given $\epsilon > 0, \lambda \in (0, 1]$ we can find $N_{\epsilon, \lambda} \in \mathbb{N}$ such that for all $n \geq N_{\epsilon, \lambda}, F_{x_n, x}(\epsilon) \geq 1 - \lambda$ holds.
2. A sequence $\{x_n\}$ in (X, F, T) is said to be a Cauchy sequence if for any given $\epsilon > 0$ and $\lambda \in (0, 1]$, there exists $N_{\epsilon, \lambda} \in \mathbb{N}$ such that $F_{x_n, x_m}(\epsilon) \geq 1 - \lambda$, whenever $m, n \geq N_{\epsilon, \lambda}$.
3. A Menger PM-space (X, F, T) is called complete if every Cauchy sequence $\{x_n\} \subset X$ is convergent to some point $x \in X$.

Lemma 1 ([17]) *If $*$ is a continuous t -norm, and $\{\alpha_n\}, \{\beta_n\}$, and $\{\gamma_n\}$ are sequences such that $\alpha_n \rightarrow \alpha, \gamma_n \rightarrow \gamma$ as $n \rightarrow \infty$, then $\limsup_{k \rightarrow \infty} (\alpha_k * \beta_k * \gamma_k) = \alpha * \limsup_{k \rightarrow \infty} \beta_k * \gamma$ and $\liminf_{k \rightarrow \infty} (\alpha_k * \beta_k * \gamma_k) = \alpha * \liminf_{k \rightarrow \infty} \beta_k * \gamma$.*

Lemma 2 ([17]) *Let $\{f(k, \cdot) : (0, \infty) \rightarrow (0, 1], k = 0, 1, 2, \dots\}$ be a sequence of functions such that $f(k, \cdot)$ is continuous and monotone increasing for each $k \geq 0$. Then, $\limsup_{k \rightarrow \infty} f(k, t)$ is a left-continuous function in t and $\liminf_{k \rightarrow \infty} f(k, t)$ is a right-continuous function in t .*

2 Main Results

Theorem 1 *Let (X, F, T) be a complete PM-space such that “ T ” is an arbitrary continuous t -norm and let $f : X \rightarrow X$ be a self-mapping satisfying the following condition*

$$\psi(F_{f_x, f_y}(t)) \leq \psi(F_{x, y}(t)) - \phi(F_{x, y}(t)), \tag{1}$$

where $\psi, \phi : (0, 1] \rightarrow [0, \infty)$ are two functions such that

- (i) ψ is monotone decreasing and continuous function with $\psi(s) = 0$ if and only if $s = 1$,
- (ii) ϕ is lower semicontinuous function with $\phi(s) = 0$ if and only if $s = 1$.

Then, f has a unique fixed point in X .

Proof Let $x_0 \in X$. We define a sequence $\{x_n\} \subset X$ such that $x_{n+1} = f x_n$, for each $n \geq 0$. If there exists a positive integer k such that $x_k = x_{k+1}$, then x_k is a fixed point of f . Hence, we shall assume that $x_n \neq x_{n+1}$, for all $n \geq 0$. Now, from (1)

$$\psi(F_{x_n, x_{n+1}}(t)) = \psi(F_{f x_{n-1}, f x_n}(t)) \leq \psi(F_{x_{n-1}, x_n}(t)) - \phi(F_{x_{n-1}, x_n}(t)). \tag{2}$$

Since ψ is monotone decreasing, we have that

$$F_{x_{n-1}, x_n}(t) \leq F_{x_n, x_{n+1}}(t).$$

Therefore, $\{F_{x_n, x_{n+1}}(t)\}$ is a monotone increasing sequence of nonnegative real numbers. Hence, there exists an $r > 0$ such that

$$\lim_{n \rightarrow \infty} F_{x_n, x_{n+1}}(t) = r.$$

Taking the limit as $n \rightarrow \infty$ in (2), we obtain

$$\psi(r) \leq \psi(r) - \phi(r),$$

which is a contradiction unless $r = 1$.

Hence

$$\lim_{n \rightarrow \infty} F_{x_n, x_{n+1}}(t) = 1. \tag{3}$$

Next, we show that $\{x_n\}$ is Cauchy sequence. If otherwise, there exist $\lambda, \epsilon > 0$ with $\lambda \in (0, 1)$ such that for each integer k , there are two integers $l(k)$ and $m(k)$ such that:

$$\begin{aligned} m(k) &> l(k) \geq k, \\ F_{x_{l(k)}, x_{m(k)}}(\epsilon) &\leq 1 - \lambda \text{ and} \\ F_{x_{l(k)}, x_{m(k)-1}}(\epsilon) &> 1 - \lambda. \end{aligned}$$

Now, by triangle inequality, for any s with $\frac{\epsilon}{2} > s > 0$ and for all $k > 0$, we have

$$\begin{aligned} 1 - \lambda &\geq F_{x_{l(k)}, x_{m(k)}}(\epsilon) \\ &\geq T(F_{x_{l(k)}, x_{l(k)+1}}(s), T(F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon - 2s), F_{x_{m(k)+1}, x_{m(k)}}(s))). \end{aligned} \tag{4}$$

For $t > 0$, we define the function:

$$h_1(t) = \limsup_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(t).$$

Taking \limsup on both the sides of (4), using (3) and the continuity of T , by Lemma 1 we conclude that

$$\begin{aligned} 1 - \lambda &\geq T(1, T(\limsup_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon - 2s), 1)) \\ &= T(1, \limsup_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon - 2s)) \\ &= \limsup_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon - 2s) \\ &= h_1(\epsilon - 2s). \end{aligned}$$

By an application of Lemma 2, h_1 is left continuous. Letting limit as $s \rightarrow 0$ in the above inequality, we obtain

$$h_1(\epsilon) = \limsup_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon) \leq 1 - \lambda. \tag{5}$$

Next, for all $t > 0$, we define the function:

$$h_2(t) = \liminf_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(t).$$

In above similar process, we can prove that

$$h_2(\epsilon) = \liminf_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon) \geq 1 - \lambda. \tag{6}$$

Combining (5) and (6), we get

$$\limsup_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon) \leq 1 - \lambda \leq \liminf_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon)$$

This implies that

$$\lim_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(t) = 1 - \lambda. \tag{7}$$

Again, by (5)

$$\limsup_{k \rightarrow \infty} F_{x_{l(k)}, x_{m(k)}}(\epsilon) \leq 1 - \lambda. \tag{8}$$

For $t > 0$, we define the function:

$$h_3(t) = \liminf_{k \rightarrow \infty} F_{x_{l(k)}, x_{m(k)}}(t).$$

Now for $s > 0$,

$$F_{x_{l(k)}, x_{m(k)}}(\epsilon + 2s) \geq T(F_{x_{l(k)}, x_{l(k)+1}}(s), T(F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon), F_{x_{m(k)+1}, x_{m(k)}}(s))).$$

Taking \liminf both the sides, we have

$$\liminf_{k \rightarrow \infty} F_{x_{l(k)}, x_{m(k)}}(\epsilon + 2s) \geq T(1, T(\liminf_{k \rightarrow \infty} F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon), 1)) = 1 - \lambda.$$

Thus, $h_3(\epsilon + 2s) \geq 1 - \lambda$.

Taking limit as $s \rightarrow 0$, we obtain

$$h_3(\epsilon) = \liminf_{k \rightarrow \infty} F_{x_{l(k)}, x_{m(k)}}(\epsilon) \geq 1 - \lambda. \tag{9}$$

Combining (8) and (9), we obtain

$$\lim_{k \rightarrow \infty} F_{x_{l(k)}, x_{m(k)}}(t) = 1 - \lambda. \tag{10}$$

Now

$$\psi(F_{x_{l(k)+1}, x_{m(k)+1}}(\epsilon)) \leq \psi(F_{x_{l(k)}, x_{m(k)}}(\epsilon)) - \phi(F_{x_{l(k)}, x_{m(k)}}(\epsilon)).$$

Taking limit as $k \rightarrow \infty$ and using (7) and (10), we obtain

$$\psi(1 - \lambda) \leq \psi(1 - \lambda) - \phi(1 - \lambda),$$

which is a contradiction.

Thus, $\{x_n\}$ is Cauchy sequence. Since X is complete, there exists $p \in X$ such that $x_n \rightarrow p$ as $n \rightarrow \infty$. Now:

$$\begin{aligned} \psi(F_{x_{n+1}, fp}(t)) &= \psi(F_{fx_n, fp}(t)) \\ &\leq \psi(F_{x_n, p}(t)) - \phi(F_{x_n, p}(t)). \end{aligned}$$

Taking limit as $n \rightarrow \infty$, we get

$$\psi(F_{p, fp}(t)) \leq \psi(F_{p, fp}(t)) - \phi(F_{p, fp}(t)) = 0,$$

which implies that $\phi(F_{p, fp}(t)) = 1$, that is,

$$F_{p, fp}(t) = 1 \text{ or } p = fp.$$

We next establish that fixed point is unique. Let p and q be two fixed points of f .

Putting $x = p$ and $y = q$ in (1),

$$\psi(F_{fp, fq}(t)) \leq \psi(F_{p, q}(t)) - \phi(F_{p, q}(t))$$

$$\text{or, } \psi(F_{p, q}(t)) \leq \psi(F_{p, q}(t)) - \phi(F_{p, q}(t))$$

$$\text{or, } \phi(F_{p, q}(t)) \leq 0,$$

or, equivalently, $\psi(F_{p, q}(t)) = 1$, that is, $p = q$.

The following example is in support of Theorem 1.

Example 1 Let $X = [0, 1]$. Define a function $F : X \times X \rightarrow D^+$ by:

$$F_{x,y}(t) = \begin{cases} 1, & \text{if } t \leq 0 \\ e^{-\frac{|x-y|}{t}}, & \text{if } t > 0. \end{cases}$$

for all $x, y \in X$. Then, (X, F, T) is a complete Menger probabilistic metric space, where “ T ” is product t -norm. Let $\psi, \phi : (0, 1] \rightarrow [0, \infty)$ be defined by:

$$\psi(s) = \frac{1}{s} - 1, \quad \phi(s) = \frac{1}{s} - \frac{1}{\sqrt{s}}, \quad \forall s \in (0, 1]. \tag{11}$$

Then, ψ and ϕ satisfy all the conditions of Theorem 1. Let the mapping $f : X \rightarrow X$ be defined by $fx = \frac{x^2}{4}$, for all $x \in X$.

Now, we shall show that f satisfies (1).

With the choices of ϕ and ψ as in (11), the inequality (1) has the form:

$$\frac{1}{F_{fx, fy}(t)} - 1 \leq \frac{1}{F_{x,y}(t)} - 1 - \frac{1}{F_{x,y}(t)} + \frac{1}{\sqrt{F_{x,y}(t)}},$$

that is,

$$F_{fx, fy}(t) \geq \sqrt{F_{x,y}(t)}.$$

Now,

$$\begin{aligned} F_{fx, fy}(t) &= e^{-\frac{|fx-fy|}{t}} \\ &= e^{-\frac{|x^2-y^2|}{4t}} \\ &= e^{-\left(\frac{|x-y|}{2t}\right)\left(\frac{|x+y|}{2}\right)} \\ &\geq e^{-\left(\frac{|x-y|}{2t}\right)} \\ &= \sqrt{F_{x,y}(t)}. \end{aligned}$$

Hence, all the conditions of Theorem (1) are satisfied.

Thus, 0 is the unique fixed point of f .

Acknowledgements The research work of the first author (Vandana Tiwari) is supported by the University Grant Commission (UGC) (No. 19-06/2011(i)EU-IV), India.

References

1. An, T.V., Chi, K.P., Karapinar, E., Thanh, T.D., An extension of generalized (ψ, ϕ) -weak contractions. *Int. J. Math. Math. Sci.* Article ID 431872, 11 pages (2012)
2. Aydi, H., Karapinar, E., Shatanawi, W., Coupled fixed point results for (ψ, ϕ) -weakly contractive condition in ordered partial metric spaces. *Comp. Math. with Appl.* 62, 4449–4460 (2011)
3. Aydi, H., Postolache, M., Shatanawi, W., Coupled fixed point results for (ψ, ϕ) -weakly contractive mappings in ordered G -metric spaces. *Comp. Math. with Appl.* 63, 298–309 (2012)
4. Berinde, V., Approximating fixed points of weak ϕ -contractions. *Fixed Point Theory.* 4, 131–142 (2003)
5. Choudhury, B.S., Das, K., A new contraction principle in Menger spaces. *Acta Math. Sin. Engl. Ser.* 24(8), 1379–1386 (2008)
6. Doric, D., Common fixed point for generalized (ψ, ϕ) -weak contractions. *Appl. Math. Lett.* 22, 1896–1900 (2009)
7. Dutta, P.N., Choudhury, B.S.: A generalization of contraction principle in metric spaces. *Fixed Point Theory and Appl.* 8, Article ID 406368 (2008)
8. Dutta, P.N., Choudhury, B.S., Das, K.: Some fixed point results in Menger spaces using a control function. *Surv. Math. Appl.* 4, 41–52 (2009)
9. Fang, J.-X.: On φ -contractions in probabilistic and fuzzy metric spaces. *Fuzz. Sets and Sys.* 267, 86–99 (2015)
10. Jamala, N., Sarwara, M., Imdad, M.: Fixed point results for generalized (ψ, ϕ) -weak contractions with an application to system of non-linear integral equations, *Trans. A. Razm. Math. Inst.* 171, 182–194 (2017)
11. Latif, A., Mongkolkeha, C., Sintunavarat, W.: Fixed point theorems for generalized α, β -weakly contraction mappings in metric spaces and applications, *Sci. World.* Article ID 784207, 14 pages (2014)

12. Luo, T.: Fuzzy (ψ, ϕ) -contractive mapping and fixed point theorem. *Appl. Math. Sci.* 8(148), 7375–7381 (2014)
13. Meneger, K.: Statistical metrics. *Proc. Nat. Acad. Sci. USA*, 28, 535–537 (1942)
14. Moradi, S., Farajzadeh, A.: On the fixed point of (ψ, ϕ) -weak and generalized (ψ, ϕ) -weak contraction mappings. *Appl. Math. Lett.* 25, 1257–1262 (2012)
15. Popescu, O.: Fixed point for $(\psi-\phi)$ -weak contractions. *Appl. Math. Lett.* 24, 1–4 (2011)
16. Rhoades, B.E.: Some theorems on weakly contractive maps. *Nonlinear Anal. TMA.* 47, 2683–2693 (2001)
17. Saha, P., Choudhury, B.S., Das, P.: Weak coupled coincidence point results having a partially ordering in fuzzy metric space. *Fuzzy. Inf. Eng.* 7, 1–18 (2016)
18. Samet, B., Vetro, C., Vetro, P.: Fixed point theorems for (α, ψ) -contractive type mappings. *Nonlinear Anal. TMA.* 75, 2154–2165 (2012)
19. Schweizer, B., Sklar, A.: Statistical metric spaces. *Pacific J. Math.* 10, 313–334 (1960)
20. Schweizer, B., Sklar, A.: *Probabilistic Metric Spaces*. Elsevier, New York, (1983)
21. Sehgal, V.M., Bharucha-Reid, A.T.: Fixed points of contraction mappings on PM-spaces. *Math. Syst. Theory.* 6, 97–102 (1972)
22. Tiwari, V., Som, T.: Fixed points for ϕ -contraction in Menger probabilistic generalized metric spaces, (Accepted in *Ann. Fuzzy Math. Inform.*)
23. Xiao, J.Z., Zhu, X.H., Cao, Y.F.: Common coupled fixed point results for probabilistic φ -contractions in Menger spaces. *Nonlinear Anal. TMA.* 74, 4589–4600 (2011)
24. Zhang, Q., Song, Y.: Fixed point theory for generalized ϕ -weak contractions. *Appl. Math. Lett.* 22, 75–78 (2009)

A Novel Canonical Duality Theory for Solving 3-D Topology Optimization Problems



David Gao and Elaf Jaafar Ali

1 Introduction

Topology optimization is a powerful tool for optimal design in multidisciplinary fields of optics, electronics, structural, bio-, and nano-mechanics. Mathematically speaking, this tool is based on finite element method such that the coupled variational/optimization problems in computational mechanics can be formulated as certain mixed integer nonlinear programming (MINLP) problems [18]. Due to the integer constraint, traditional theory and methods in continuous optimization can't be applied for solving topology optimization problems. Therefore, most MINLP problems are considered to be NP-hard (*nondeterministic polynomial-time hard*) in global optimization and computer science [23]. During the past forty years, many approximate methods have been developed for solving topology optimization problems, these include homogenization method [3, 4], density-based method [2], solid isotropic material with penalization (SIMP) [41, 42, 53], level set approximation [38, 43], evolutionary structural optimization (ESO) [50, 51], and

© Springer International Publishing, AG 2018
V.K. Singh, D.Y. Gao, A. Fisher (eds). *Emerging Trends in Applied Mathematics and High-Performance Computing*

D. Gao (✉) · E. J. Ali
School of Science and Technology, Federation University Australia, Mt Helen, VIC 3353,
Australia
e-mail: d.gao@federation.edu.au

© Springer Nature Switzerland AG 2019
V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_13

bidirectional evolutionary structural optimization (BESO) [30, 39, 40]. Currently, the popular commercial software products used in topology optimization are based on SIMP and ESO/BESO methods [31, 36, 46, 52]. However, these approximate methods can't mathematically guarantee the global convergence. Also, they usually suffer from having different intrinsic disadvantages, such as slow convergence, the gray-scale elements, and checkerboards patterns, etc. [6, 44, 45].

Canonical duality theory (CDT) is a methodological theory, which was developed from Gao and Strang's original work in 1989 on finite deformation mechanics [28]. The key feature of this theory is that by using certain canonical strain measure, general nonconvex/nonsmooth potential variational problems can be equivalently reformulated as a pure (stress-based only) complementary energy variational principle [11]. The associated triality theory provides extremality criteria for both global and local optimal solutions, which can be used to develop powerful algorithms for solving general nonconvex variational problems [12]. This pure complementary energy variational principle solved a well-known open problem in nonlinear elasticity and is known as the Gao principle in literature [35]. Based on this principle, a canonical dual finite element method was proposed in 1996 for large deformation nonconvex/nonsmooth mechanics [9]. Applications have been given to post-buckling problems of large deformed beams [1], nonconvex variational problems [24], and phase transitions in solids [29]. It was discovered by Gao in 2007 that by simply using a canonical measure $\epsilon(x) = x(x - 1) = 0$, the 0-1 integer constraint $x \in \{0, 1\}$ in general nonconvex minimization problems can be equivalently converted to a unified concave maximization problem in continuous space, which can be solved deterministically to obtain global optimal solution in polynomial time [14]. Therefore, this pure complementary energy principle plays a fundamental role not only in computational nonlinear mechanics, but also in discrete optimization [25]. Most recently, Gao proved that the topology optimization should be formulated as a bi-level mixed integer nonlinear programming problem (BL-MINLP) [18, 20]. The upper-level optimization of this BL-MINLP is actually equivalent to the well-known Knapsack problem, which can be solved analytically by the CDT [20]. The review articles [15, 27] and the newly published book [23] provide comprehensive reviews and applications of the canonical duality theory in multidisciplinary fields of mathematical modeling, engineering mechanics, nonconvex analysis, global optimization, and computational science.

The main goal of this paper is to apply the canonical duality theory for solving 3-dimensional benchmark problems in topology optimization. In the next section, we first review Gao's recent work why the topology optimization should be formulated as a bi-level mixed integer nonlinear programming problem. A basic mathematical mistake in topology optimization modeling is explicitly addressed. A canonical penalty-duality method for solving this Knapsack problem is presented in Section 3, which is actually the so-called β -perturbation method first proposed in global optimization [25] and recently in topology optimization [18]. Section 4 reveals for the first time the unified relation between this canonical penalty-duality method in integer programming and Gao's pure complementary energy principle in nonlinear elasticity. Section 5 provides 3-D finite element analysis and the associated

canonical penalty-duality (CPD) algorithm. The volume evolutionary method and computational complexity of this CPD algorithm are discussed. Applications to 3-D benchmark problems are provided in Section 6. The paper is ended with concluding remarks and open problems. Mathematical mistakes in the popular methods are explicitly addressed. Also, general modeling and conceptual mistakes in engineering optimization are discussed based on reviewers' comments.

2 Mathematical Problems for 3-D Topology Optimization

The minimum total potential energy principle provides a theoretical foundation for all mathematical problems in computational solid mechanics. For general 3-D nonlinear elasticity, the total potential energy has the following standard form:

$$\Pi(\mathbf{u}, \rho) = \int_{\Omega} \left(W(\nabla \mathbf{u})\rho + \mathbf{u} \cdot \mathbf{b}\rho \right) d\Omega - \int_{\Gamma_t} \mathbf{u} \cdot \mathbf{t} d\Gamma, \quad (1)$$

where $\mathbf{u} : \Omega \rightarrow \mathbb{R}^3$ is a displacement vector field, \mathbf{b} is a given body force vector, \mathbf{t} is a given surface traction on the boundary $\Gamma_t \subset \partial\Omega$, and the dot-product $\mathbf{u} \cdot \mathbf{t} = \mathbf{u}^T \mathbf{t}$. In this paper, the stored energy density $W(\mathbf{F})$ is an *objective function* (see Remark 4) of the deformation gradient $\mathbf{F} = \nabla \mathbf{u}$. In topology optimization, the mass density $\rho : \Omega \rightarrow \{0, 1\}$ is the design variable, which takes $\rho(\mathbf{x}) = 1$ at a solid material point $\mathbf{x} \in \Omega$, while $\rho(\mathbf{x}) = 0$ at a void point $\mathbf{x} \in \Omega$. Additionally, it must satisfy the so-called knapsack condition:

$$\int_{\Omega} \rho(\mathbf{x}) d\Omega \leq V_c, \quad (2)$$

where $V_c > 0$ is a desired volume bound.

By using finite element method, the whole design domain Ω is meshed with n disjointed finite elements $\{\Omega_e\}$. In each element, the unknown variables can be numerically written as $\mathbf{u}(\mathbf{x}) = \mathbf{N}(\mathbf{x})\mathbf{u}_e$, $\rho(\mathbf{x}) = \rho_e \in \{0, 1\} \quad \forall \mathbf{x} \in \Omega_e$, where $\mathbf{N}(\mathbf{x})$ is a given interpolation matrix, and \mathbf{u}_e is a nodal displacement vector. Let $\mathcal{U}_a \subset \mathbb{R}^m$ be a kinetically admissible space, in which certain deformation conditions are given, v_e represents the volume of the e -th element Ω_e , and $\mathbf{v} = \{v_e\} \in \mathbb{R}^n$. Then, the admissible design space can be discretized as a discrete set:

$$\mathcal{Z}_a = \left\{ \boldsymbol{\rho} = \{\rho_e\} \in \mathbb{R}^n \mid \rho_e \in \{0, 1\} \quad \forall e = 1, \dots, n, \quad \boldsymbol{\rho}^T \mathbf{v} = \sum_{e=1}^n \rho_e v_e \leq V_c \right\} \quad (3)$$

and on $\mathcal{U}_a \times \mathcal{Z}_a$, the total potential energy functional can be numerically reformulated as a real-valued function:

$$\Pi_h(\mathbf{u}, \boldsymbol{\rho}) = C(\boldsymbol{\rho}, \mathbf{u}) - \mathbf{u}^T \mathbf{f}, \quad (4)$$

where

$$C(\boldsymbol{\rho}, \mathbf{u}) = \boldsymbol{\rho}^T \mathbf{c}(\mathbf{u}),$$

in which

$$\mathbf{c}(\mathbf{u}) = \left\{ \int_{\Omega_e} [W(\nabla \mathbf{N}(\mathbf{x})\mathbf{u}_e) - \mathbf{b}^T \mathbf{N}(\mathbf{x})\mathbf{u}_e] d\Omega \right\} \in \mathbb{R}^n, \quad (5)$$

and

$$\mathbf{f} = \left\{ \int_{\Gamma_e} \mathbf{N}(\mathbf{x})^T \mathbf{t}(\mathbf{x}) d\Gamma \right\} \in \mathbb{R}^m.$$

By the facts that the topology optimization is a combination of both variational analysis on a continuous space \mathcal{U}_a and optimal design on a discrete space \mathcal{Z}_a , it can't be simply formulated in a traditional variational form. Instead, a general problem of topology optimization should be proposed as a bi-level programming [20]:

$$(\mathcal{P}_{bl}) : \quad \min\{\Phi(\boldsymbol{\rho}, \mathbf{u}) \mid \boldsymbol{\rho} \in \mathcal{Z}_a, \mathbf{u} \in \mathcal{U}_a\}, \quad (6)$$

$$\text{s.t. } \mathbf{u} \in \arg \min_{\mathbf{v} \in \mathcal{U}_a} \Pi_h(\mathbf{v}, \boldsymbol{\rho}), \quad (7)$$

where $\Phi(\boldsymbol{\rho}, \mathbf{u})$ represents the upper-level cost function, and $\boldsymbol{\rho} \in \mathcal{Z}_a$ is the upper-level variable. Similarly, $\Pi_h(\mathbf{u}, \boldsymbol{\rho})$ represents the lower-level cost function and $\mathbf{u} \in \mathcal{U}_a$ is the lower-level variable. The cost function $\Phi(\boldsymbol{\rho}, \mathbf{u})$ depends on both particular problems and numerical methods. It can be $\Phi(\boldsymbol{\rho}^p, \mathbf{u}) = \mathbf{f}^T \mathbf{u} - \mathbf{c}(\mathbf{u})^T \boldsymbol{\rho}^p$ for any given parameter $p \geq 1$, or simply $\Phi(\boldsymbol{\rho}, \mathbf{u}) = -\boldsymbol{\rho}^T \mathbf{c}(\mathbf{u})$.

Since the topology optimization is a design-analysis process, it is reasonable to use the alternative iteration method [20] for solving the challenging topology optimization problem (\mathcal{P}_{bl}) , that is:

- (i) For a given design variable $\boldsymbol{\rho}_{k-1} \in \mathcal{Z}_a$, solving the lower-level optimization (7) for

$$\mathbf{u}_k = \arg \min\{\Pi_h(\mathbf{u}, \boldsymbol{\rho}_{k-1}) \mid \mathbf{u} \in \mathcal{U}_a\} \quad (8)$$

- (ii) For the given $\mathbf{c}_u = \mathbf{c}(\mathbf{u}_k)$, solve the upper-level optimization problem (6) for

$$\boldsymbol{\rho}_k = \arg \min\{\Phi(\boldsymbol{\rho}, \mathbf{u}_k) \mid \boldsymbol{\rho} \in \mathcal{Z}_a\}. \quad (9)$$

The upper-level problem (9) is actually equivalent to the well-known Knapsack problem in its most simple (linear) form:

$$(\mathcal{P}_u) : \quad \min\{P_u(\boldsymbol{\rho}) = -\mathbf{c}_u^T \boldsymbol{\rho} \mid \boldsymbol{\rho}^T \mathbf{v} \leq V_c, \boldsymbol{\rho} \in \{0, 1\}^n\}, \quad (10)$$

which makes a perfect sense in topology optimization, i.e., among all elements $\{\Omega_e\}$, one should keep those stored more strain energy. Knapsack problems appear extensively in multidisciplinary fields of operations research, decision science, and engineering design problems. Due to the integer constraint, even this most simple linear knapsack problem is listed as one of Karp’s 21 NP-complete problems [33]. However, by using the canonical duality theory, this challenging problem can be solved easily to obtain global optimal solution.

For linear elastic structures without the body force, the stored energy C is a quadratic function of \mathbf{u} :

$$C(\boldsymbol{\rho}, \mathbf{u}) = \frac{1}{2} \mathbf{u}^T \mathbf{K}(\boldsymbol{\rho}) \mathbf{u}, \tag{11}$$

where $\mathbf{K}(\boldsymbol{\rho}) = \{\rho_e \mathbf{K}_e\} \in \mathbb{R}^{n \times n}$ is the overall stiffness matrix, obtained by assembling the sub-matrix $\rho_e \mathbf{K}_e$ for each element Ω_e . For any given $\boldsymbol{\rho} \in \mathcal{Z}_a$, the displacement variable can be obtained analytically by solving the linear equilibrium equation $\mathbf{K}(\boldsymbol{\rho}) \mathbf{u} = \mathbf{f}$. Thus, the topology optimization for linear elastic structures can be simply formulated as:

$$(\mathcal{P}_{le}) : \quad \min \left\{ \mathbf{f}^T \mathbf{u} - \frac{1}{2} \mathbf{u}^T \mathbf{K}(\boldsymbol{\rho}) \mathbf{u} \mid \mathbf{K}(\boldsymbol{\rho}) \mathbf{u} = \mathbf{f}, \mathbf{u} \in \mathcal{U}_a, \boldsymbol{\rho} \in \mathcal{Z}_a \right\}. \tag{12}$$

Remark 1 (On Compliance Minimization Problem) In the literature, topology optimization for linear elastic structures is usually formulated as a compliance minimization problem (see [36] and the problem (P) in [47]²):

$$(P) : \quad \min_{\boldsymbol{\rho} \in \mathbb{R}^n, \mathbf{u} \in \mathcal{U}_a} \frac{1}{2} \mathbf{f}^T \mathbf{u} \quad s.t. \quad \mathbf{K}(\boldsymbol{\rho}) \mathbf{u} = \mathbf{f}, \boldsymbol{\rho} \in \{0, 1\}^n, \boldsymbol{\rho}^T \mathbf{v} \leq V_c. \tag{13}$$

Clearly, if the displacement is replaced by $\mathbf{u} = [\mathbf{K}(\boldsymbol{\rho})]^{-1} \mathbf{f}$, this problem can be written as:

$$(P_c) : \quad \min \left\{ P_c(\boldsymbol{\rho}) = \frac{1}{2} \mathbf{f}^T [\mathbf{K}(\boldsymbol{\rho})]^{-1} \mathbf{f} \mid \mathbf{K}(\boldsymbol{\rho}) \text{ is invertible for all } \boldsymbol{\rho} \in \mathcal{Z}_a \right\}. \tag{14}$$

which is equivalent to (\mathcal{P}_{le}) under the regularity condition, i.e., $[\mathbf{K}(\boldsymbol{\rho})]^{-1}$ exists for all $\boldsymbol{\rho} \in \mathcal{Z}_a$. However, instead of \mathbf{u} the given external force in the cost function of (P) is replaced by $\mathbf{f} = \mathbf{K} \mathbf{u}$ such that (P) is commonly written in the so-called minimization of strain energy (see [45]):

$$(P_s) : \quad \min \left\{ \frac{1}{2} \mathbf{u}^T \mathbf{K}(\boldsymbol{\rho}) \mathbf{u} \mid \mathbf{K}(\boldsymbol{\rho}) \mathbf{u} = \mathbf{f}, \boldsymbol{\rho} \in \mathcal{Z}_a, \mathbf{u} \in \mathcal{U}_a \right\}, \tag{15}$$

²The linear inequality constraint $\mathbf{A} \boldsymbol{\rho} \leq \mathbf{b}$ in [36] is ignored in this paper.

One can see immediately that (P_s) contradicts (P_{le}) in the sense that the alternative iteration for solving (P_c) leads to an anti-Knapsack problem:

$$\min \mathbf{c}_u^T \boldsymbol{\rho}, \quad s.t. \quad \boldsymbol{\rho} \in \{0, 1\}^n, \quad \boldsymbol{\rho}^T \mathbf{v} \leq V_c. \quad (16)$$

By the fact that $\mathbf{c}_u = \mathbf{c}(\mathbf{u}_k) \in \mathbb{R}_+^n := \{\mathbf{c} \in \mathbb{R}^n \mid \mathbf{c} \geq \mathbf{0}\}$ is a nonnegative vector for any given \mathbf{u}_k , this problem has only a trivial solution. Therefore, the alternative iteration is not allowed for solving (P_s) . In continuum physics, the linear scalar-valued function $\mathbf{u}^T \mathbf{f} \in \mathbb{R}$ is called the external (or input) energy, which is not an objective function (see Remark 4). Since \mathbf{f} is a given force, it can't be replaced by $\mathbf{K}(\boldsymbol{\rho})\mathbf{u}$. Although the cost function $P_c(\boldsymbol{\rho})$ can be called as the mean compliance, it is not an objective function either. Thus, the problem (P_c) works only for those problems that $\mathbf{u}(\boldsymbol{\rho})$ can be uniquely determined. Its complementary form:

$$(P^c) : \quad \max \left\{ \frac{1}{2} \mathbf{u}^T \mathbf{K}(\boldsymbol{\rho})\mathbf{u} \mid \mathbf{K}(\boldsymbol{\rho})\mathbf{u} = \mathbf{f}, \quad \boldsymbol{\rho} \in \mathcal{Z}_a \right\} \quad (17)$$

can be called a maximum stiffness problem, which is equivalent to (P_{le}) in the sense that both problems produce the same results by the alternative iteration method. Therefore, it is a conceptual mistake to call the strain energy $\frac{1}{2} \mathbf{u}^T \mathbf{K}(\boldsymbol{\rho})\mathbf{u}$ as the mean compliance and (P_s) as the compliance minimization.³ The problem (P_s) has been used as a mathematical model for many approximation methods, including the SIMP and BESO. Additionally, some conceptual mistakes in the compliance minimization and mathematical modeling are also addressed in Remark 4.

3 Canonical Dual Solution to Knapsack Problem

The canonical duality theory for solving general integer programming problems was first proposed by Gao in 2007 [14]. Applications to topology optimization have been given recently in [18, 20]. In this paper, we present this theory in a different way, i.e., instead of the canonical measure in \mathbb{R}^{n+1} , we introduce a canonical measure in \mathbb{R}^n :

$$\boldsymbol{\varepsilon} = \Lambda(\boldsymbol{\rho}) = \boldsymbol{\rho} \circ \boldsymbol{\rho} - \boldsymbol{\rho} \in \mathbb{R}^n \quad (18)$$

³Due to this conceptual mistake, the general problem for topology optimization was originally formulated as a double-min optimization (P_{bl}) in [18]. Although this model is equivalent to a knapsack problem for linear elastic structures under the condition $\mathbf{f} = \mathbf{K}(\boldsymbol{\rho})\mathbf{u}$, it contradicts the popular theory in topology optimization.

and the associated super-potential:

$$\Psi(\boldsymbol{\varepsilon}) = \begin{cases} 0 & \text{if } \boldsymbol{\varepsilon} \in \mathbb{R}_-^n := \{\boldsymbol{\varepsilon} \in \mathbb{R}^n \mid \boldsymbol{\varepsilon} \leq \mathbf{0}\} \\ +\infty & \text{otherwise,} \end{cases} \quad (19)$$

such that the integer constraint in the Knapsack problem (\mathcal{P}_u) can be relaxed by the following canonical form:

$$\min \left\{ \Pi_u(\boldsymbol{\rho}) = \Psi(\Lambda(\boldsymbol{\rho})) - \mathbf{c}_u^T \boldsymbol{\rho} \mid \boldsymbol{\rho}^T \mathbf{v} \leq V_c, \boldsymbol{\rho} \in \mathbb{R}^n \right\}. \quad (20)$$

This is a nonsmooth minimization problem in \mathbb{R}^n with only one linear inequality constraint. The classical Lagrangian for this inequality-constrained problem is

$$L(\boldsymbol{\rho}, \tau) = \Psi(\Lambda(\boldsymbol{\rho})) - \mathbf{c}_u^T \boldsymbol{\rho} + \tau(\boldsymbol{\rho}^T \mathbf{v} - V_c), \quad (21)$$

and the canonical minimization problem (20) is equivalent to the following min-max problem:

$$\min_{\boldsymbol{\rho} \in \mathbb{R}^n} \max_{\tau \in \mathbb{R}} L(\boldsymbol{\rho}, \tau) \quad \text{s.t.} \quad \tau \geq 0. \quad (22)$$

According to the Karush-Kuhn-Tucker theory in inequality-constrained optimization, the Lagrange multiplier τ should satisfy the following KKT conditions:

$$\tau(\boldsymbol{\rho}^T \mathbf{v} - V_c) = 0, \quad \tau \geq 0, \quad \boldsymbol{\rho}^T \mathbf{v} - V_c \leq 0. \quad (23)$$

The first equality $\tau(\boldsymbol{\rho}^T \mathbf{v} - V_c) = 0$ is the so-called *complementarity condition*. It is well known that to solve the complementarity problems is not an easy task, even for linear complementarity problems [32]. Also, the Lagrange multiplier has to satisfy the constraint qualification $\tau \geq 0$. Therefore, the classical Lagrange multiplier theory can be essentially used for linear equality-constrained optimization problems [34]. This is one of main reasons why the canonical duality theory was developed.

By the fact that the super-potential $\Psi(\boldsymbol{\varepsilon})$ is a convex, lower-semicontinuous function (l.s.c), its sub-differential is a positive cone \mathbb{R}_+^n [12]:

$$\partial \Psi(\boldsymbol{\varepsilon}) = \begin{cases} \{\boldsymbol{\sigma}\} \in \mathbb{R}_+^n & \text{if } \boldsymbol{\varepsilon} \leq \mathbf{0} \in \mathbb{R}_-^n \\ \emptyset & \text{otherwise.} \end{cases} \quad (24)$$

Using Fenchel transformation, the conjugate function of $\Psi(\boldsymbol{\varepsilon})$ can be uniquely defined by (see [12]):

$$\Psi^\sharp(\boldsymbol{\sigma}) = \sup_{\boldsymbol{\varepsilon} \in \mathbb{R}^n} \{\boldsymbol{\varepsilon}^T \boldsymbol{\sigma} - \Psi(\boldsymbol{\varepsilon})\} = \begin{cases} 0 & \text{if } \boldsymbol{\sigma} \in \mathbb{R}_+^n, \\ +\infty & \text{otherwise,} \end{cases} \quad (25)$$

which can be viewed as a *super complementary energy* [8]. By the theory of convex analysis, we have the following *canonical duality relations* [14]:

$$\Psi(\boldsymbol{\varepsilon}) + \Psi^\sharp(\boldsymbol{\sigma}) = \boldsymbol{\varepsilon}^T \boldsymbol{\sigma} \Leftrightarrow \boldsymbol{\sigma} \in \partial \Psi(\boldsymbol{\varepsilon}) \Leftrightarrow \boldsymbol{\varepsilon} \in \partial \Psi^\sharp(\boldsymbol{\sigma}). \quad (26)$$

By the Fenchel-Young equality $\Psi(\boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^T \boldsymbol{\sigma} - \Psi^\sharp(\boldsymbol{\sigma})$, the Lagrangian $L(\boldsymbol{\rho}, \tau)$ can be written in the following form:

$$\Xi(\boldsymbol{\rho}, \boldsymbol{\sigma}, \tau) = G_{ap}(\boldsymbol{\rho}, \boldsymbol{\sigma}) - \boldsymbol{\rho}^T \boldsymbol{\sigma} - \Psi^\sharp(\boldsymbol{\sigma}) - \boldsymbol{\rho}^T \mathbf{c}_u + \tau(\boldsymbol{\rho}^T \mathbf{v} - V_c). \quad (27)$$

This is the Gao-Strang total complementary function for the Knapsack problem, in which $G_{ap}(\boldsymbol{\rho}, \boldsymbol{\sigma}) = \boldsymbol{\sigma}^T (\boldsymbol{\rho} \circ \boldsymbol{\rho})$ is the so-called *complementary gap function*. Clearly, if $\boldsymbol{\sigma} \in \mathbb{R}_+^n$, this gap function is convex and $G_{ap}(\boldsymbol{\rho}, \boldsymbol{\sigma}) \geq 0 \quad \forall \boldsymbol{\rho} \in \mathbb{R}^n$. Let

$$\mathcal{S}_a^+ = \{\boldsymbol{\zeta} = \{\boldsymbol{\sigma}, \tau\} \in \mathbb{R}^{n+1} \mid \boldsymbol{\sigma} > \mathbf{0} \in \mathbb{R}^n, \tau \geq 0\}. \quad (28)$$

Then on \mathcal{S}_a , we have

$$\Xi(\boldsymbol{\rho}, \boldsymbol{\zeta}) = \boldsymbol{\sigma}^T (\boldsymbol{\rho} \circ \boldsymbol{\rho} - \boldsymbol{\rho}) - \boldsymbol{\rho}^T \mathbf{c}_u + \tau(\boldsymbol{\rho}^T \mathbf{v} - V_c) \quad (29)$$

and for any given $\boldsymbol{\zeta} \in \mathcal{S}_a^+$, the canonical dual function can be obtained by:

$$P_u^d(\boldsymbol{\zeta}) = \min_{\boldsymbol{\rho} \in \mathbb{R}^n} \Xi(\boldsymbol{\rho}, \boldsymbol{\zeta}) = -\frac{1}{4} \boldsymbol{\tau}_u^T(\boldsymbol{\zeta}) \mathbf{G}(\boldsymbol{\sigma})^{-1} \boldsymbol{\tau}_u(\boldsymbol{\zeta}) - \tau V_c, \quad (30)$$

where

$$\mathbf{G}(\boldsymbol{\sigma}) = \text{Diag}(\boldsymbol{\sigma}), \quad \boldsymbol{\tau}_u = \boldsymbol{\sigma} + \mathbf{c}_u - \tau \mathbf{v}.$$

This canonical dual function is the so-called *pure complementary energy* in nonlinear elasticity, first proposed by Gao in 1999 [11], where $\boldsymbol{\tau}_u$ and $\boldsymbol{\sigma}$ are corresponding to the first and second Piola-Kirchhoff stresses, respectively. Thus, the canonical dual problem of the Knapsack problem can be proposed in the following:

$$(\mathcal{P}_u^d) : \quad \max \left\{ P_u^d(\boldsymbol{\zeta}) \mid \boldsymbol{\zeta} \in \mathcal{S}_a^+ \right\}. \quad (31)$$

Theorem 1 (Canonical Dual Solution for Knapsack Problem [18]) For any given $\mathbf{u}_k \in \mathcal{U}_a$ and $V_c > 0$, if $\bar{\boldsymbol{\zeta}} = (\bar{\boldsymbol{\sigma}}, \bar{\tau}) \in \mathcal{S}_a^+$ is a solution to (\mathcal{P}_u^d) , then

$$\bar{\boldsymbol{\rho}} = \frac{1}{2} \mathbf{G}(\bar{\boldsymbol{\sigma}})^{-1} \boldsymbol{\tau}_u(\bar{\boldsymbol{\zeta}}) \quad (32)$$

is a global minimum solution to the Knapsack problem (\mathcal{P}_u) and

$$P_u(\bar{\boldsymbol{\rho}}) = \min_{\boldsymbol{\rho} \in \mathbb{R}^n} P_u(\boldsymbol{\rho}) = \Xi(\bar{\boldsymbol{\rho}}, \bar{\boldsymbol{\zeta}}) = \max_{\boldsymbol{\zeta} \in \mathcal{S}_a^+} P_u^d(\boldsymbol{\zeta}) = P_u^d(\bar{\boldsymbol{\zeta}}). \quad (33)$$

Proof By the convexity of the super-potential $\Psi(\boldsymbol{\epsilon})$, we have $\Psi^{**}(\boldsymbol{\epsilon}) = \Psi(\boldsymbol{\epsilon})$. Thus,

$$L(\boldsymbol{\rho}, \tau) = \sup_{\boldsymbol{\sigma} \in \mathbb{R}^n} \Xi(\boldsymbol{\rho}, \boldsymbol{\sigma}, \tau) \quad \forall \boldsymbol{\rho} \in \mathbb{R}^n, \quad \tau \in \mathbb{R}. \quad (34)$$

It is easy to show that for any given $\boldsymbol{\rho} \in \mathbb{R}^n$, $\tau \in \mathbb{R}$, the supremum condition is governed by $\Lambda(\boldsymbol{\rho}) \in \partial\Psi^*(\boldsymbol{\sigma})$. By the canonical duality relations given in (26), we have the equivalent relations:

$$\Lambda(\boldsymbol{\rho})^T \boldsymbol{\sigma} = \boldsymbol{\sigma}^T (\boldsymbol{\rho} \circ \boldsymbol{\rho} - \boldsymbol{\rho}) = 0 \quad \Leftrightarrow \quad \boldsymbol{\sigma} \in \mathbb{R}_+^n \quad \Leftrightarrow \quad \Lambda(\boldsymbol{\rho}) = (\boldsymbol{\rho} \circ \boldsymbol{\rho} - \boldsymbol{\rho}) \in \mathbb{R}_-^n. \quad (35)$$

This is exactly equivalent to the KKT conditions of the canonical problem for the inequality condition $\Lambda(\boldsymbol{\rho}) \in \mathbb{R}_-^n$. Thus, if $\bar{\boldsymbol{\zeta}} \in \mathcal{S}_a^+$ is a KKT solution to (\mathcal{P}_u^d) , then $\bar{\boldsymbol{\sigma}} > \mathbf{0}$ and the complementarity condition in (35) leads to $\bar{\boldsymbol{\rho}} \circ \bar{\boldsymbol{\rho}} - \bar{\boldsymbol{\rho}} = \mathbf{0}$, i.e., $\bar{\boldsymbol{\rho}} \in \{0, 1\}^n$. It is easy to prove that for a given $\bar{\boldsymbol{\zeta}}$, the equality (32) is exactly the criticality condition $\nabla_{\boldsymbol{\rho}} \Xi(\bar{\boldsymbol{\rho}}, \bar{\boldsymbol{\zeta}}) = \mathbf{0}$. Therefore, the vector $\bar{\boldsymbol{\rho}} \in \{0, 1\}^n$ defined by (32) is a solution to the Knapsack problem (\mathcal{P}_u) . According to Gao and Strang [28] that the total complementary function $\Xi(\boldsymbol{\rho}, \boldsymbol{\zeta})$ is a saddle function on $\mathbb{R}^n \times \mathcal{S}_a^+$, then

$$\min_{\boldsymbol{\rho} \in \mathbb{R}^n} P_u(\boldsymbol{\rho}) = \min_{\boldsymbol{\rho} \in \mathbb{R}^n} \max_{\boldsymbol{\zeta} \in \mathcal{S}_a^+} \Xi(\boldsymbol{\rho}, \boldsymbol{\zeta}) = \max_{\boldsymbol{\zeta} \in \mathcal{S}_a^+} \min_{\boldsymbol{\rho} \in \mathbb{R}^n} \Xi(\boldsymbol{\rho}, \boldsymbol{\zeta}) = \max_{\boldsymbol{\zeta} \in \mathcal{S}_a^+} P_u^d(\boldsymbol{\zeta}). \quad (36)$$

The complementary-dual equality (33) can be proved by the canonical duality relations. \square

This theorem shows that the so-called NP-hard Knapsack problem is canonically dual to a concave maximization problem (\mathcal{P}_u^d) in continuous space, which is much easier than the 0-1 programming problem (\mathcal{P}_u) in discrete space. Whence the canonical dual solution $\bar{\boldsymbol{\zeta}}$ is obtained, the solution to the Knapsack problem can be given analytically by (32).

4 Pure Complementary Energy Principle and Perturbed Solution

Based on Theorem 1, a perturbed solution for the Knapsack problem has been proposed recently in [18, 20]. This section demonstrates the relation of this solution with the pure complementary energy principle in nonlinear elasticity discovered by Gao in 1997–1999 [10, 11].

In terms of the deformation $\boldsymbol{\chi} = \mathbf{u} + \mathbf{x}$, the total potential energy variational principle for general large deformation problems can also be written in the following form:

$$(\mathcal{P}_{\boldsymbol{\chi}}) : \quad \inf_{\boldsymbol{\chi} \in \mathcal{X}_a} \Pi(\boldsymbol{\chi}) = \int_{\Omega} [W(\nabla \boldsymbol{\chi}) - \boldsymbol{\chi} \cdot \mathbf{b}] \rho d\Omega - \int_{\Gamma_f} \boldsymbol{\chi} \cdot \mathbf{t} d\Gamma, \quad (37)$$

where \mathcal{X}_a is a kinetically admissible deformation space, in which the boundary condition $\boldsymbol{\chi}(\mathbf{x}) = \mathbf{0}$ is given on $\Gamma_{\boldsymbol{\chi}}$. It is well known that the stored energy $W(\mathbf{F})$ is usually a nonconvex function of the deformation gradient $\mathbf{F} = \nabla \boldsymbol{\chi} = \nabla \mathbf{u} + \mathbf{I}$ in order to model complicated phenomena, such as phase transitions and post-buckling. By the fact that $W(\mathbf{F})$ must be an objective function [37], there exists a real-valued function $\Psi(\mathbf{C})$ such that $W(\mathbf{F}) = \Psi(\mathbf{F}^T \mathbf{F})$ (see [5]). For most reasonable materials (say the St. Venant-Kirchhoff material [22]), the function $\Psi(\mathbf{C})$ is a usually convex function of the Cauchy strain measure $\mathbf{C} = \mathbf{F}^T \mathbf{F}$ such that its complementary energy density can be uniquely defined by the Legendre transformation:

$$\Psi^*(\mathbf{S}) = \{ \text{tr}(\mathbf{C} \cdot \mathbf{S}) - \Psi(\mathbf{C}) \mid \mathbf{S} = \nabla \Psi(\mathbf{C}) \}. \quad (38)$$

Therefore, a pure complementary energy variational principle was obtained by Gao in 1999 [11, 12]:

Theorem 2 (Pure Complementary Energy Principle for Nonlinear Elasticity [11])

For any given external force field $\mathbf{b}(\mathbf{x})$ in Ω and $\mathbf{t}(\mathbf{x})$ on Γ_t , if $\boldsymbol{\tau}(\mathbf{x})$ is a statically admissible stress field, that is:

$$\boldsymbol{\tau} \in \mathcal{T}_a := \left\{ \boldsymbol{\tau}(\mathbf{x}) : \Omega \rightarrow \mathbb{R}^{3 \times 3} \mid -\nabla \cdot \boldsymbol{\tau} = \mathbf{b} \quad \forall \mathbf{x} \in \Omega, \quad \mathbf{n} \cdot \boldsymbol{\tau} = \mathbf{t} \quad \forall \mathbf{x} \in \Gamma_t \right\}, \quad (39)$$

and $\bar{\mathbf{S}}$ is a critical point of the pure complementary energy:

$$\Pi^d(\mathbf{S}) = - \int_{\Omega} \left[\frac{1}{4} \text{tr}(\boldsymbol{\tau} \cdot \mathbf{S}^{-1} \cdot \boldsymbol{\tau}) + \Psi^*(\mathbf{S}) \right] \rho \, d\Omega, \quad (40)$$

then the deformation field $\bar{\boldsymbol{\chi}}(\mathbf{x})$ defined by:

$$\bar{\boldsymbol{\chi}}(\mathbf{x}) = \frac{1}{2} \int_{\mathbf{x}_0}^{\mathbf{x}} \boldsymbol{\tau} \cdot \bar{\mathbf{S}}^{-1} \, d\mathbf{x} \quad (41)$$

along any path from $\mathbf{x}_0 \in \Gamma_{\boldsymbol{\chi}}$ to $\mathbf{x} \in \Omega$ is a critical point of the total potential energy $\Pi(\boldsymbol{\chi})$ and $\Pi(\bar{\boldsymbol{\chi}}) = \Pi^d(\bar{\mathbf{S}})$. Moreover, if $\bar{\mathbf{S}}(\mathbf{x}) \succ 0 \quad \forall \mathbf{x} \in \Omega$, then $\bar{\boldsymbol{\chi}}$ is a global minimizer of $\Pi(\boldsymbol{\chi})$.

It is easy to prove that the criticality condition $\delta \Pi_{\boldsymbol{\chi}}^d(\mathbf{S}) = 0$ is governed by the so-called canonical dual algebraic equation [12]:

$$4\mathbf{S} \cdot [\nabla \Psi^*(\mathbf{S})] \cdot \mathbf{S} = \boldsymbol{\tau}^T \cdot \boldsymbol{\tau}. \quad (42)$$

For certain materials, this algebraic equation can be solved analytically to obtain all possible solutions [24]. Particularly, for the St Venant-Kirchhoff material, this tensor equation could have at most 27 solutions at each material point \mathbf{x} , but

only one positive-definite $\mathbf{S}(\mathbf{x}) \succ 0 \quad \forall \mathbf{x} \in \Omega$, which leads to the global minimum solution $\bar{\chi}(\mathbf{x})$ [22]. The pure complementary energy principle solved a well-known open problem in large deformation mechanics and is known as the Gao principle in literature (see [35]). This principle plays an important role not only in large deformation theory and nonconvex variational analysis but also in global optimization and computational science. Indeed, Theorem 1 is simply an application of this principle as if we consider the quadratic operator $\boldsymbol{\varepsilon}(\boldsymbol{\rho})$ as the Cauchy strain measure $\mathbf{C}(\boldsymbol{\chi})$, then the canonical dual $\boldsymbol{\sigma} \in \partial\Psi(\boldsymbol{\varepsilon})$ is corresponding to the second Piola-Kirchhoff stress $\mathbf{S} = \nabla\Psi(\mathbf{C})$, while $\boldsymbol{\tau}_u$ is corresponding to the first Piola-Kirchhoff stress $\boldsymbol{\tau}$. By the fact that $\Psi^\sharp(\boldsymbol{\sigma})$ is nonsmooth, the associated canonical dual algebraic equation (42) should be governed by the KKT conditions (35). In order to solve this problem, a β -perturbation method was proposed in 2010 for solving general integer programming problems [25] and recently for solving the topology optimization problems [18].

According to the canonical duality theory for mathematical modeling [20], the integer constraint $\boldsymbol{\rho} \in \{0, 1\}^n$ in the Knapsack problem (\mathcal{P}_u) is a constitutive condition, while $\boldsymbol{\rho} \cdot \mathbf{v} \leq V_c$ is a geometrical constraint. Thus, by using the so-called pan-penalty functions:

$$W(\boldsymbol{\rho}) = \begin{cases} 0 & \text{if } \boldsymbol{\rho} \in \{0, 1\}^n \\ +\infty & \text{otherwise,} \end{cases} \quad F(\boldsymbol{\rho}) = \begin{cases} \mathbf{c}_u \cdot \boldsymbol{\rho} & \text{if } \boldsymbol{\rho} \cdot \mathbf{v} \leq V_c \\ -\infty & \text{otherwise,} \end{cases} \quad (43)$$

the Knapsack problem (\mathcal{P}_u) can be equivalently written in Gao-Strang's unconstrained form [28]:

$$\min \{W(\boldsymbol{\rho}) - F(\boldsymbol{\rho}) \mid \boldsymbol{\rho} \in \mathbb{R}^n\}. \quad (44)$$

By introducing a penalty parameter $\beta > 0$ and a Lagrange multiplier $\tau \geq 0$, these two pan-penalty functions can have the following relaxations:

$$W_\beta(\boldsymbol{\rho}) = \beta \|\boldsymbol{\rho} \circ \boldsymbol{\rho} - \boldsymbol{\rho}\|^2, \quad F_\tau(\boldsymbol{\rho}) = \mathbf{c}_u \cdot \boldsymbol{\rho} - \tau(\boldsymbol{\rho} \cdot \mathbf{v} - V_c). \quad (45)$$

It is easy to prove that

$$W(\boldsymbol{\rho}) = \lim_{\beta \rightarrow \infty} W_\beta(\boldsymbol{\rho}), \quad F(\boldsymbol{\rho}) = \min_{\tau \geq 0} F_\tau(\boldsymbol{\rho}) \quad \forall \boldsymbol{\rho} \in \mathbb{R}^n. \quad (46)$$

Thus, the Knapsack problem can be relaxed by the so-called penalty-duality approach:

$$\min_{\boldsymbol{\rho} \in \mathbb{R}^n} \max_{\tau \geq 0} \{L_\beta(\boldsymbol{\rho}, \tau) = W_\beta(\boldsymbol{\rho}) - \mathbf{c}_u \cdot \boldsymbol{\rho} + \tau(\boldsymbol{\rho} \cdot \mathbf{v} - V_c)\}. \quad (47)$$

Since the penalty function $W_\beta(\boldsymbol{\rho})$ is nonconvex, by using the canonical transformation $W_\beta(\boldsymbol{\rho}) = \Psi_\beta(\Lambda(\boldsymbol{\rho}))$, we have $\Psi_\beta(\boldsymbol{\varepsilon}) = \beta \|\boldsymbol{\varepsilon}\|^2$, which is a convex

quadratic function. Its Legendre conjugate is simply $\Psi_\beta^*(\sigma) = \frac{1}{4}\beta^{-1}\|\sigma\|^2$. Thus, the Gao and Strang total complementary optimization problem for the penalty-duality approach (47) can be given by [18]:

$$\min_{\rho \in \mathbb{R}^n} \max_{\zeta \in \mathcal{S}_a^+} \left\{ \Xi_\beta(\rho, \zeta) = (\rho \circ \rho - \rho) \cdot \sigma - \frac{1}{4}\beta^{-1}\|\sigma\|^2 - \mathbf{c}_u \cdot \rho + \tau(\rho \cdot \mathbf{v} - V_c) \right\}. \quad (48)$$

For any given $\beta > 0$ and $\zeta = \{\sigma, \tau\} \in \mathcal{S}_a^+$, a canonical penalty-duality (CPD) function can be obtained as:

$$P_\beta^d(\zeta) = \min_{\rho \in \mathbb{R}^n} \Xi_\beta(\rho, \zeta) = P_u^d(\sigma, \tau) - \frac{1}{4}\beta^{-1}\|\sigma\|^2, \quad (49)$$

which is exactly the so-called β -perturbed canonical dual function presented in [18, 20]. It was proved by Theorem 7 in [25] that there exists a $\beta_c > 0$ such that for any given $\beta \geq \beta_c$, both the CPD problem:

$$(\mathcal{P}_\beta^d) : \max\{P_\beta^d(\zeta) \mid \zeta \in \mathcal{S}_a^+\} \quad (50)$$

and the problem (\mathcal{P}_u^d) have the same solution set. Since $\Psi_\beta^*(\sigma)$ is a quadratic function, the corresponding canonical dual algebraic equation (42) is a coupled cubic algebraic system:

$$2\beta^{-1}\sigma_e^3 + \sigma_e^2 = (\tau v_e - c_e)^2, \quad e = 1, \dots, n, \quad (51)$$

$$\sum_{e=1}^n \frac{1}{2} \frac{v_e}{\sigma_e} (\sigma_e - v_e \tau + c_e) - V_c = 0. \quad (52)$$

It was proved in [12, 14] that for any given $\beta > 0$, $\tau \geq 0$, and $\mathbf{c}_u = \{c_e(\mathbf{u}_e)\}$ such that $\theta_e = \tau v_e - c_e(\mathbf{u}_e) \neq 0$, $e = 1, \dots, n$, the canonical dual algebraic equation (51) has a unique positive real solution:

$$\sigma_e = \frac{1}{12}\beta[-1 + \phi_e(\tau) + \phi_e^c(\tau)] > 0, \quad e = 1, \dots, n \quad (53)$$

where

$$\phi_e(\zeta) = \eta^{-1/3} \left[2\theta_e^2 - \eta + 2i\sqrt{\theta_e^2(\eta - \theta_e^2)} \right]^{1/3}, \quad \eta = \frac{\beta^2}{27},$$

and ϕ_e^c is the complex conjugate of ϕ_e , i.e., $\phi_e \phi_e^c = 1$. Thus, a canonical penalty-duality algorithm has been proposed recently for solving general topology optimization problems [18, 20].

5 CPD Algorithm for 3-D Topology Optimization

For three-dimensional linear elastic structures, we simply use cubic 8-node hexahedral elements $\{\Omega_e\}$, each element contains 24 degrees of freedom corresponding to the displacements in x-y-z directions (each node has three degrees of freedom) as shown in Figure 1. Thus, the displacement interpolation matrix is $\mathbf{N} = [N_1 \ N_2 \ \dots \ N_8]$ and

$$N_i = \begin{bmatrix} N_i & 0 & 0 \\ 0 & N_i & 0 \\ 0 & 0 & N_i \end{bmatrix}. \tag{54}$$

The shape functions $N_i = N_i(\xi_1, \xi_2, \xi_3), i = 1, \dots, 8$ are derived by:

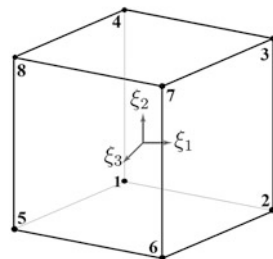
$$\begin{aligned} N_1 &= \frac{1}{8}(1 - \xi_1)(1 - \xi_2)(1 - \xi_3), & N_2 &= \frac{1}{8}(1 + \xi_1)(1 - \xi_2)(1 - \xi_3), \\ N_3 &= \frac{1}{8}(1 + \xi_1)(1 + \xi_2)(1 - \xi_3), & N_4 &= \frac{1}{8}(1 - \xi_1)(1 + \xi_2)(1 - \xi_3), \\ N_5 &= \frac{1}{8}(1 - \xi_1)(1 - \xi_2)(1 + \xi_3), & N_6 &= \frac{1}{8}(1 + \xi_1)(1 - \xi_2)(1 + \xi_3), \\ N_7 &= \frac{1}{8}(1 + \xi_1)(1 + \xi_2)(1 + \xi_3), & N_8 &= \frac{1}{8}(1 - \xi_1)(1 + \xi_2)(1 + \xi_3), \end{aligned}$$

in which $\xi_1, \xi_2,$ and ξ_3 are the natural coordinates of the i^{th} node. The nodal displacement vector \mathbf{u}_e is given by:

$$\mathbf{u}_e^T = [u_1^e \ u_2^e \ \dots \ u_8^e],$$

where $u_i^e = (x_i^e, y_i^e, z_i^e) \in \mathbb{R}^3, i = 1, \dots, 8,$ are the displacement components at node i . The components B_i of strain-displacement matrix $\mathbf{B} = [B_1 \ B_2 \ \dots \ B_8],$ which relates the strain ε and the nodal displacement $\mathbf{u}_e (\varepsilon = \mathbf{B}\mathbf{u}_e),$ are defined as:

Fig. 1 The hexahedron element—eight nodes



$$\mathbf{B}_i = \begin{bmatrix} \frac{\partial N_i}{\partial x} & 0 & 0 \\ 0 & \frac{\partial N_i}{\partial y} & 0 \\ 0 & 0 & \frac{\partial N_i}{\partial z} \\ \frac{\partial N_i}{\partial y} & \frac{\partial N_i}{\partial x} & 0 \\ \frac{\partial N_i}{\partial z} & 0 & \frac{\partial N_i}{\partial x} \\ 0 & \frac{\partial N_i}{\partial z} & \frac{\partial N_i}{\partial y} \end{bmatrix}. \tag{55}$$

Hooke’s law for isotropic materials in constitutive matrix form is given by:

$$\mathbf{H} = \frac{E}{(1 + \nu)(1 - 2\nu)} \begin{bmatrix} 1 - \nu & \nu & \nu & 0 & 0 & 0 \\ \nu & 1 - \nu & \nu & 0 & 0 & 0 \\ \nu & \nu & 1 - \nu & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1-2\nu}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1-2\nu}{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1-2\nu}{2} \end{bmatrix}, \tag{56}$$

where E is the Young’s modulus and ν is the Poisson’s ratio of the isotropic material. The stiffness matrix of the structure in CPD algorithm is given by:

$$\mathbf{K}(\boldsymbol{\rho}) = \sum_{e=1}^n (E_{min} + (E - E_{min})\rho_e)\mathbf{K}_e, \tag{57}$$

where E_{min} must be small enough (usually let $E_{min} = 10 - 9E$) to avoid singularity in computation and \mathbf{K}_e is defined as:

$$\mathbf{K}_e = \int_{-1}^1 \int_{-1}^1 \int_{-1}^1 \mathbf{B}^T \mathbf{H} \mathbf{B} \, d\xi_1 d\xi_2 d\xi_3. \tag{58}$$

Based on the canonical duality theory, an evolutionary canonical penalty-duality (CPD) algorithm⁴ for solving the topology optimization problem [18] can be presented in the following.

Canonical Penalty-Duality Algorithm for Topology Optimization (CPD)

1. Initialization:

Choose a suitable initial volume reduction rate $\mu < 1$.

Let $\boldsymbol{\rho}^0 = \{1\} \in \mathbb{R}^n$.

Given an initial value $\tau^0 > 0$, an initial volume $V_\gamma = \mu V_0$.

Given a perturbation parameter $\beta > 10$, error allowances ω_1 and ω_2 , in which ω_1 is a termination criterion.

⁴This algorithm was called the CDT algorithm in [18]. Since a new CDT algorithm without β perturbation has been developed, this algorithm based on the canonical penalty-duality method should be called CPD algorithm.

Let $\gamma = 0$ and compute

$$\mathbf{u}^0 = \mathbf{K}^{-1}(\boldsymbol{\rho}^0)\mathbf{f}(\boldsymbol{\rho}^0), \quad \mathbf{c}^0 = \mathbf{c}(\mathbf{u}^0) = \mathbf{u}^{0T} \mathbf{K}(\boldsymbol{\rho}^0)\mathbf{u}^0.$$

2. Let $k = 1$.
3. Compute $\boldsymbol{\zeta}_k = \{\boldsymbol{\sigma}^k, \tau^k\}$ by:

$$\sigma_e^k = \frac{1}{6}\beta[-1 + \phi_e(\tau^{k-1}) + \phi_e^c(\tau^{k-1})], \quad e = 1, \dots, n.$$

$$\tau^k = \frac{\sum_{e=1}^n v_e(1 + c_e^\gamma/\sigma_e^k) - 2V_\gamma}{\sum_{e=1}^n v_e^2/\sigma_e^k}.$$

4. If

$$\Delta = |P_u^d(\boldsymbol{\sigma}^k, \tau^k) - P_u^d(\boldsymbol{\sigma}^{k-1}, \tau^{k-1})| > \omega_1, \tag{59}$$

then let $k = k + 1$, go to Step 3; otherwise, continue.

5. Compute $\boldsymbol{\rho}^{\gamma+1} = \{\rho_e^{\gamma+1}\}$ and $\mathbf{u}^{\gamma+1}$ by:

$$\rho_e^{\gamma+1} = \frac{1}{2}[1 - (\tau^k v_e - c_e^\gamma)/\sigma_e^k], \quad e = 1, \dots, n.$$

$$\mathbf{u}^{\gamma+1} = \mathbf{K}(\boldsymbol{\rho}^{\gamma+1})^{-1}\mathbf{f}(\boldsymbol{\rho}^{\gamma+1}).$$

6. If $|\boldsymbol{\rho}^{\gamma+1} - \boldsymbol{\rho}^\gamma| \leq \omega_2$ and $V_\gamma \leq V_c$, then stop; otherwise, continue.
7. Let $V_{\gamma+1} = \mu V_\gamma$, $\tau^0 = \tau^k$, and $\gamma = \gamma + 1$, go to step 2.

Remark 2 (Volume Evolutionary Method and Computational Complexity) By Theorem 1, we know that for any given desired volume $V_c > 0$, the optimal solution $\bar{\boldsymbol{\rho}}$ can be analytically obtained by (32) in terms of its canonical dual solution in continuous space. By the fact that the topology optimization problem (\mathcal{P}_{bl}) is a coupled nonconvex minimization, numerical optimization depends sensitively on the the initial volume V_0 . If $\mu_c = V_c/V_0 \ll 1$, any given iteration method could lead to unreasonable numerical solutions. In order to resolve this problem, a volume decreasing control parameter $\mu \in (\mu_c, 1)$ was introduced in [18] to produce a volume sequence $V_\gamma = \mu V_{\gamma-1}$ ($\gamma = 1, \dots, \gamma_c$) such that $V_{\gamma_c} = V_c$ and for any given $V_\gamma \in [V_c, V_0]$, the problem (\mathcal{P}_{bl}) is replaced by:

$$(\mathcal{P}_{bl})^\gamma : \min \left\{ \mathbf{f}^T \mathbf{u} - C_p(\boldsymbol{\rho}, \mathbf{u}) \mid \boldsymbol{\rho} \in \{0, 1\}^n, \quad \mathbf{v}^T \boldsymbol{\rho} \leq V_\gamma \right\}, \tag{60}$$

$$\text{s.t. } \mathbf{u}(\boldsymbol{\rho}) = \arg \min \{ \Pi_h(\mathbf{v}, \boldsymbol{\rho}) \mid \mathbf{v} \in \mathcal{U}_a \}. \tag{61}$$

The initial values for solving this γ -th problem are $V_{\gamma-1}$, $\mathbf{u}_{\gamma-1}$, $\boldsymbol{\rho}_{\gamma-1}$. Theoretically speaking, for any given sequence $\{V_\gamma\}$ we should have

$$(\mathcal{P}_{bl}) = \lim_{\gamma \rightarrow \gamma_c} (\mathcal{P}_{bl})^\gamma. \quad (62)$$

Numerically, different volume sequence $\{V_\gamma\}$ may produce totally different structural topology as long as the alternative iteration is used. This is intrinsic difficulty for all coupled bi-level optimal design problems.

The original idea of this sequential volume decreasing technique is from an evolutionary method for solving optimal shape design problems (see Chapter 7, [12]). It was realized recently that the same idea was used in the ESO and BESO methods. But, these two methods are not polynomial-time algorithm. By the facts that there are only two loops in the CPD algorithm, i.e., the γ -loop and the k -loop, and the canonical dual solution is analytically given in the k -loop, the main computing is the $m \times m$ matrix inversion in the γ -loop. The complexity for the Gauss-Jordan elimination is $O(m^3)$. Therefore, the CPD is a polynomial-time algorithm.

6 Applications to 3-D Benchmark Problems

In order to demonstrate the novelty of the CPD algorithm for solving 3D topology optimization problems, our numerical results are compared with the two popular methods: BESO and SIMP. The algorithm for the soft-kill BESO is from [31].⁵ A modified SIMP algorithm without filter is used according to [36]. The parameters used in BESO and SIMP are: the minimum radius $r_{\min} = 1.5$, the evolutionary rate $er = 0.05$, and the penalization power $p = 3$. Young's modulus and Poisson's ratio of the material are taken as $E = 1$ and $\nu = 0.3$, respectively. The initial value for τ used in CPD is $\tau^0 = 1$. We take the design domain $V_0 = 1$, the initial design variable $\boldsymbol{\rho}^0 = \{1\}$ for both CPD and BESO algorithms. All computations are performed by a computer with Processor Intel Core I7-4790, CPU 3.60GHz, and memory 16.0 GB.

6.1 Cantilever Beam Problems

For this benchmark problem, we present results based on three types of mesh resolutions with two types of loading conditions.

⁵According to Professor Y.M. Xie at RMIT, this BESO code was poorly implemented and has never been used for any of their further research simply because it was extremely slow compared to their other BESO codes. Therefore, the comparison for computing time between CPD and BESO provided in this section may not show the reality if the other commercial BESO codes are used.

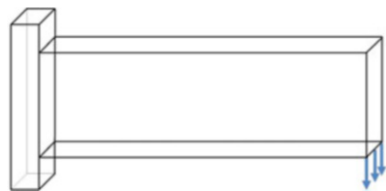
6.1.1 Uniformly Distributed Load with $60 \times 20 \times 4$ Meshes

First, let us consider the cantilever beam with uniformly distributed load at the right end as illustrated in Figure 2. The target volume and termination criterion for CPD, BESO, and SIMP are selected as $V_c = 0.3$ and $\omega_1 = 10^{-6}$, respectively. For both CPD and BESO methods, we take the volume evolution rate $\mu = 0.89$, the perturbation parameter for CPD is $\beta = 4000$. The results are reported in Table 1.⁶

Figure 3 shows the convergence of compliances produced by all the three methods. As we can see that the SIMP provides an upper bound approach since this method is based on the minimization of the compliance, i.e., the problem (P). By Remark 1, we know that this problem violates the minimum total potential energy principle, the SIMP converges in a strange way, i.e., the structures produced by the SIMP at the beginning are broken until $It. = 15$ (see Figure 3), which is physically unreasonable. Dually, both the CPD and BESO provide lower bound approaches. It is reasonable to believe that the main idea of the BESO is similar to the Knapsack problem, i.e., at each volume iteration, to eliminate elements which stored less strain energy by simply using comparison method. By the fact that the same volume evolutionary rate μ is adopted, the results obtained by the CPD and BESO are very close to each other (see also Figure 4). However, the CPD is almost 100 times faster than the BESO method since the BESO is not a polynomial-time algorithm.

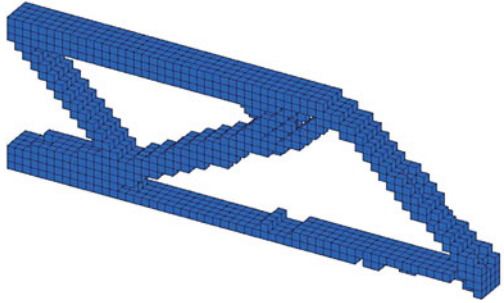
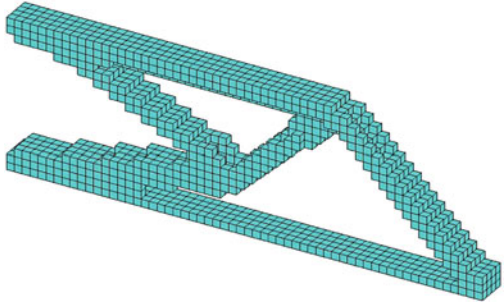
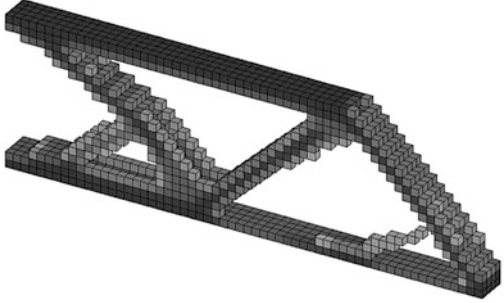
The optimal structures produced by the CPD with $\omega_1 = 10^{-6}$ and with different values of μ and β are summarized in Table 2. Also, the target compliances during the iterations for all CPD examples are reported in Figure 5 with different values of μ and β . The results show that the CPD algorithm sensitively depends on the volume evolution parameter μ , but not the penalty parameter β . The comparison for volume evolutions by CPD and BESO is given in Figure 6, which shows as expected that the BESO method also sensitively depends on the volume evolutionary rate μ . For a fixed $\beta = 4000$, the convergence of the CPD is more stable and faster than the BESO. The C-Iteration curve for BESO jumps for every given μ , which could be the so-called “chaotic convergence curves” addressed by G. I. N. Rozvany in [41].

Fig. 2 Cantilever beam with uniformly distributed load in the right end



⁶The so-called compliance in this section is actually a doubled strain energy, i.e., $c = 2C(\rho, \mathbf{u})$ as used in [36].

Table 1 Structures produced by CPD, BESO, and SIMP for cantilever beam ($60 \times 20 \times 4$)

Method	Details	Structure
CPD	$C = 1973.028$ It. = 23 Time= 27.1204	
BESO	$C = 1771.3694$ It. = 154 Time= 2392.9594	
SIMP	$C = 2416.6333$ It. = 200 Time= 98.7545	

6.1.2 Uniformly Distributed Load with $120 \times 50 \times 8$ Mesh Resolution

Now, let us consider the same loaded beam as shown in Figure 2 but with a finer mesh resolution of $120 \times 50 \times 8$. In this example, the target volume fraction and termination criterion for all procedures are assumed to be $V_c = 0.3$ and $\omega_1 = 10^{-6}$, respectively. The initial volume reduction rate for both CPD and BESO is $\mu = 0.935$. The perturbation parameter for CPD is $\beta = 7000$. The optimal topologies produced by CPD, BESO, and SIMP methods are reported in Table 3. As we can see, the CPD is about five times faster than the SIMP and almost 100 times faster than the BESO method.

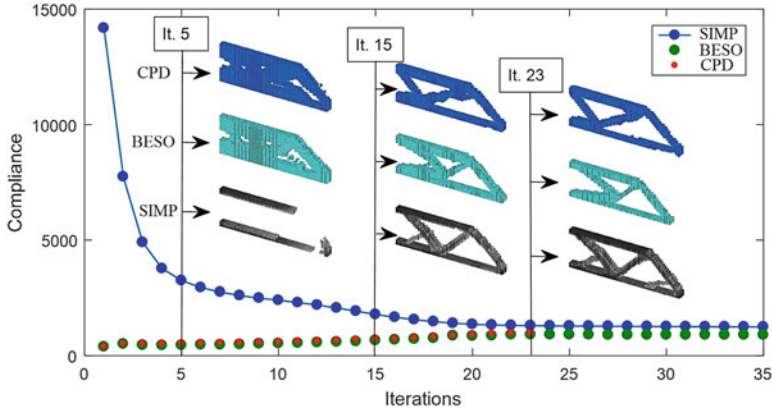


Fig. 3 Convergence test for CPD, BESO, and SIMP

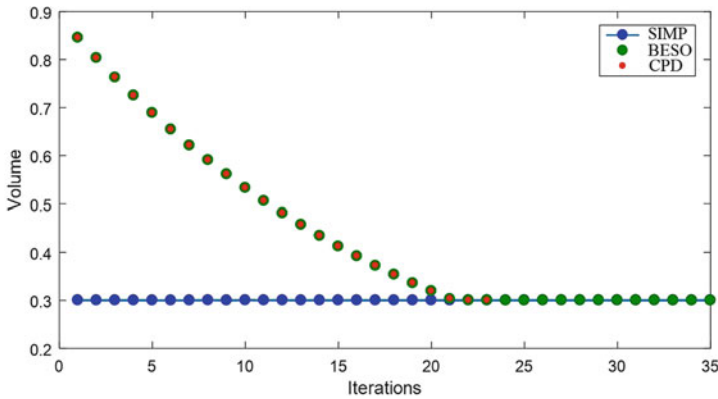


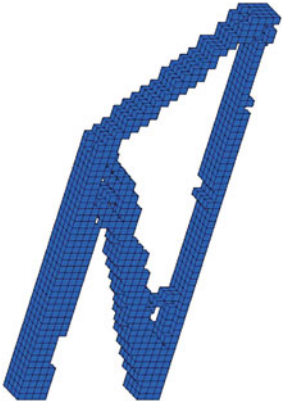
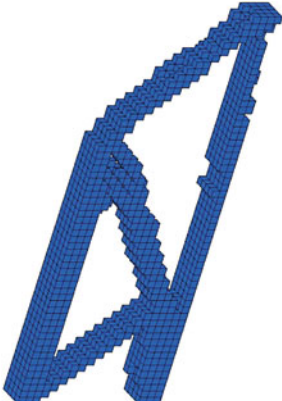
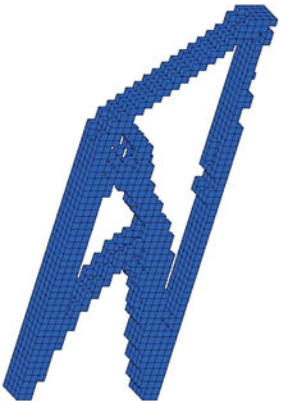
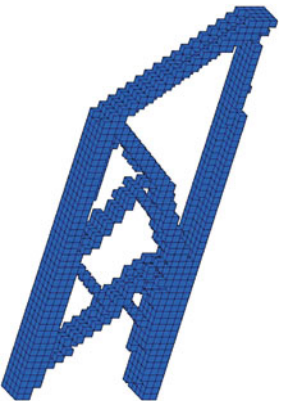
Fig. 4 Comparison of volume variations for CPD, BESO, and SIMP

If we choose $\omega_1 = 0.001$, the computing times (iterations) for CPD, BESO, and SIMP are 0.97 (24), 24.67 (44), and 4.3 (1000) hours, respectively. Actually, the SIMP failed to reach the given precision. If we increase $\omega_1 = 0.01$, the SIMP takes 3.14 hours with 742 iterations to satisfy the given precision. Our numerical results show that the CPD method can produce very good results with much less computing time. For a given very small $\omega_1 = 10^{-16}$, Table 4 shows the effects of the parameters of μ , β , and V_c on the computing time of the CPD method.

6.1.3 Beam with a Central Load and $40 \times 20 \times 20$ Meshes

In this example, the beam is subjected to a central load at its right end (see Figure 7). We let $V_c = 0.095$, $\omega_1 = 0.001$, $\beta = 7000$, and $\mu = 0.888$.

Table 2 Optimal structures produced by CPD with different values of μ and β

Details	Structure	Details	Structure
$\mu = 0.88$ $\beta = 4000$ $C = 2182.78$ It. =22 Time=29.44		$\mu = 0.89$ $\beta = 90000$ $C = 1973.02$ It. =23 Time=30.69	
$\mu = 0.9$ $\beta = 4000$ $C = 1920.68$ It. =23 Time=30.87		$\mu = 0.92$ $\beta = 90000$ $C = 1832.59$ It. =23 Time=33.73	

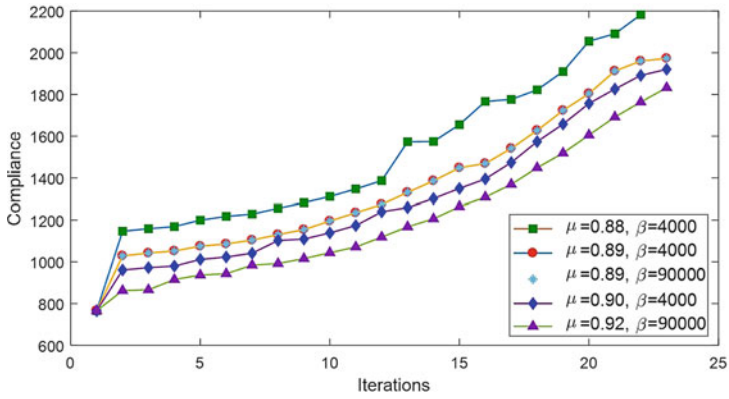


Fig. 5 Convergence tests for CPD method at different values of μ and β

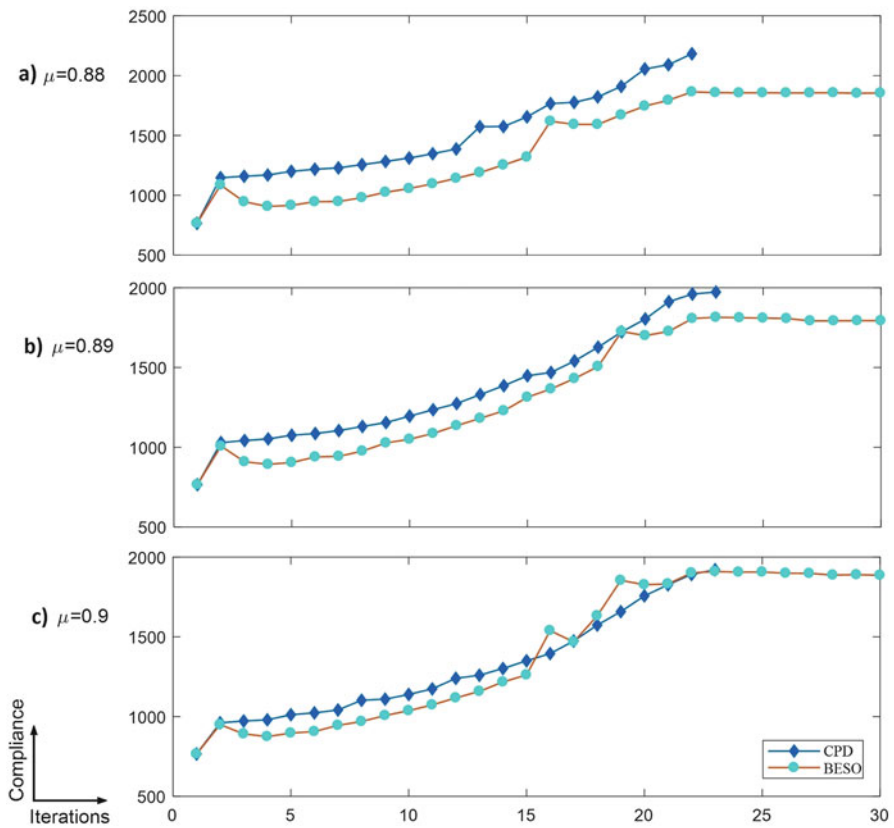
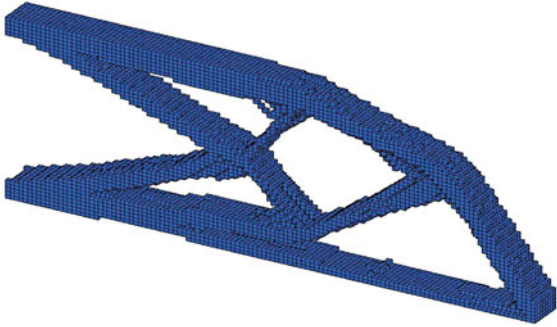
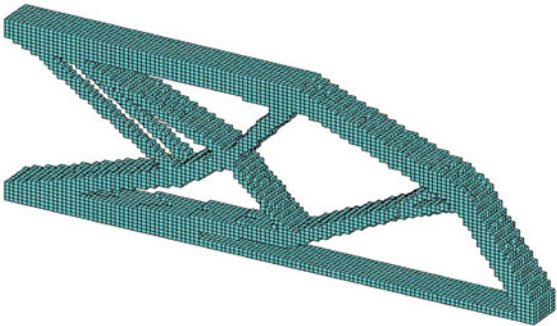
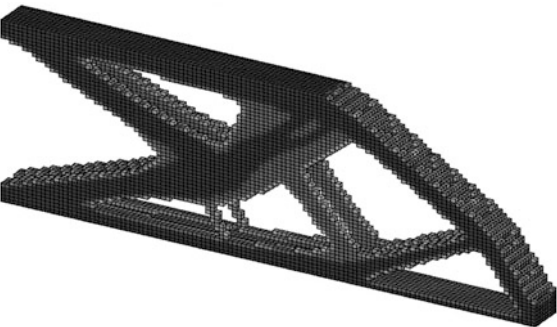


Fig. 6 Convergence test for CPD and BESO with different μ .

Table 3 Topology optimization for cantilever beam ($120 \times 50 \times 8$)

Method	Details	Structure
CPD	$C = 1644.0886$ It. =24 Time=3611.23	
BESO	$C = 1605.1102$ It. =200 Time=342751.96	
SIMP	$C = 1835.4106$ It. =1000 Time=15041.06	

The topology optimized structures produced by CPD, SIMP, and BESO methods are summarized in Table 5. Compared with the SIMP method, we can see that by using only 20% of computing time, the CPD can produce global optimal solution, which is better than that produced by the BESO, but with only 8% of computing time. We should point out that for the given $\omega_1 = 0.001$, the SIMP method failed to converge in 1000 iterations (the so-called “change” $\Delta = 0.0061 > \omega_1$).

Table 4 Effects of μ , β , and V_c to the final results by CPD method ($\omega_1 = 10^{-16}$)





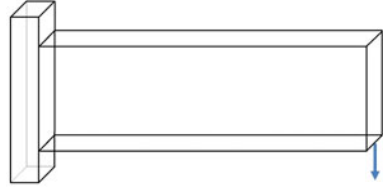
<p>$\mu = 0.935$, $\beta = 3000$, $V_c = 0.3$ $C = 1632.959$, It. =25, Time=3022.029</p> 	<p>$\mu = 0.935$, $\beta = 7000$, $V_c = 0.18$ $C = 2669.980$, It. =34, Time=5040.6647</p> 
<p>$\mu = 0.98$, $\beta = 7000$, $V_c = 0.3$ $C = 1635.922$, It. =25, Time=3531.3235</p> 	<p>$\mu = 0.98$, $\beta = 7000$, $V_c = 0.18$ $C = 2892.914$, It. =35, Time=4853.3776</p> 

Fig. 7 Design domain for cantilever beam with a central load in the right end



6.2 MBB Beam

The second benchmark problem is the 3-D Messerschmitt- \ddot{B} olkow-Blohm (MBB) beam. Two examples with different loading and boundary conditions are illustrated.

6.2.1 Example 1

The MBB beam design for this example is illustrated in Figure 8. In this example, we use $40 \times 20 \times 20$ mesh resolution, $V_c = 0.1$, and $\omega_1 = 0.001$. The initial volume reduction rate and perturbation parameter are $\mu = 0.89$ and $\beta = 5000$, respectively.

Table 6 summarizes the optimal topologies by using CPD, BESO, and SIMP methods. Compared with the BESO method, we see again that the CPD produces a mechanically sound structure and takes only 12.6% of computing time. Also, the SIMP method failed to converge for this example and the result presented in Table 6 is only the output of the 1000th iteration when $\Delta = 0.039 > \omega_1$.

6.2.2 Example 2

In this example, the MBB beam is supported horizontally in its four bottom corners under central load as shown in Figure 9. The mesh resolution is $60 \times 10 \times 10$, the target volume is $V_c = 0.155$. The initial volume reduction rate and perturbation parameter are defined as $\mu = 0.943$ and $\beta = 7250$, respectively.

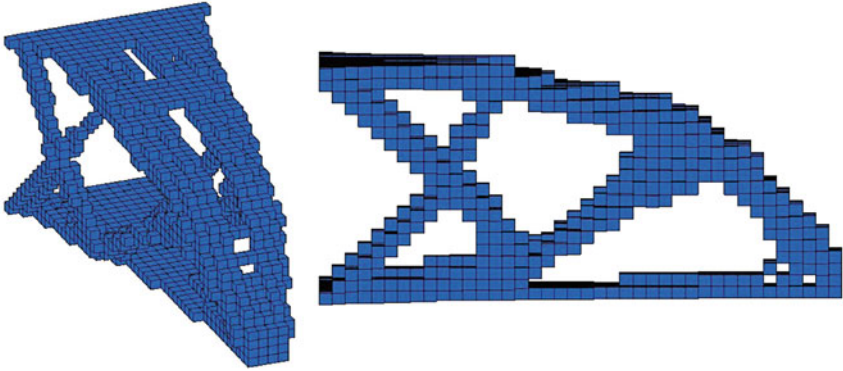
The topology optimized structures produced by CPD, BESO, and SIMP with $\omega_1 = 10^{-5}$ are reported in Table 7. Once again, we can see that without using any artificial techniques, the CPD produces mechanically sound integer density distribution but the computing time is only 3.3% of that used by the BESO.

6.3 Cantilever Beam with a Given Hole

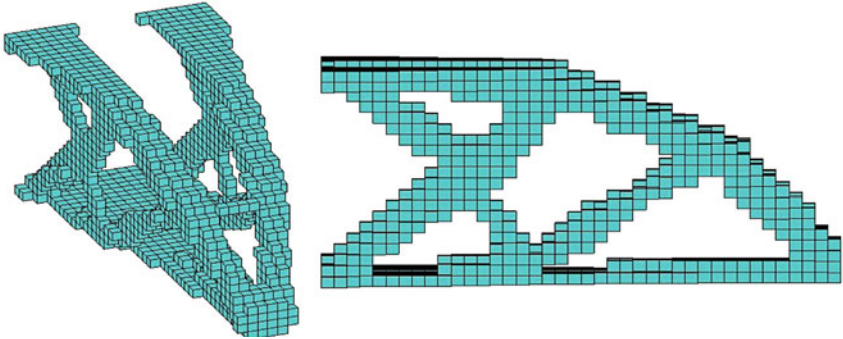
In real-world applications, the desired structures are usually subjected to certain design constraints such that some elements are required to be either solid or void. Now, let us consider the cantilever beam with a given hole as illustrated in Figure 10.

Table 5 Topologies of the cantilever beam with a central load in the right end

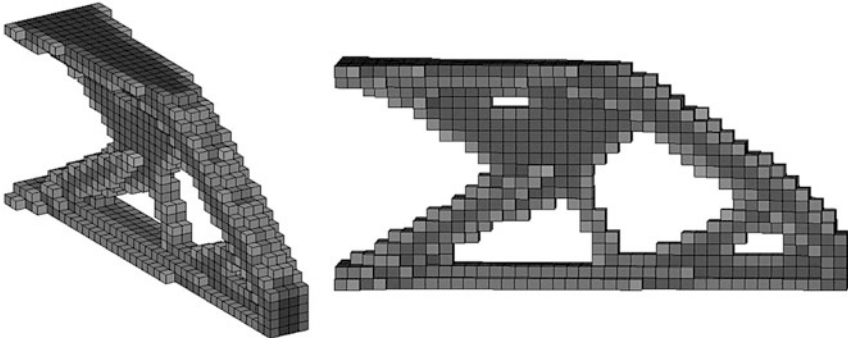
CPD: $C = 20.564$, It. =45, Time=959.7215



BESO: $C = 20.1533$, It. =53, Time=11461.128



SIMP: $C = 25.7285$, It. =1000, Time=4788.4762



We use mesh resolution $70 \times 30 \times 6$ and parameters $V_c = 0.5$, $\beta = 7000$, $\mu = 0.94$, and $\omega_1 = 0.001$.

The optimal topologies produced by CPD, BESO, and SIMP are summarized in Table 8. The results show clearly that the CPD method is significantly faster than

Fig. 8 MBB beam with uniformly distributed central load

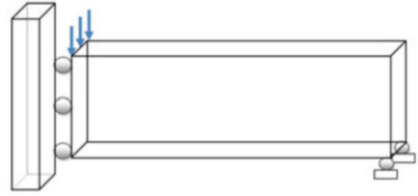
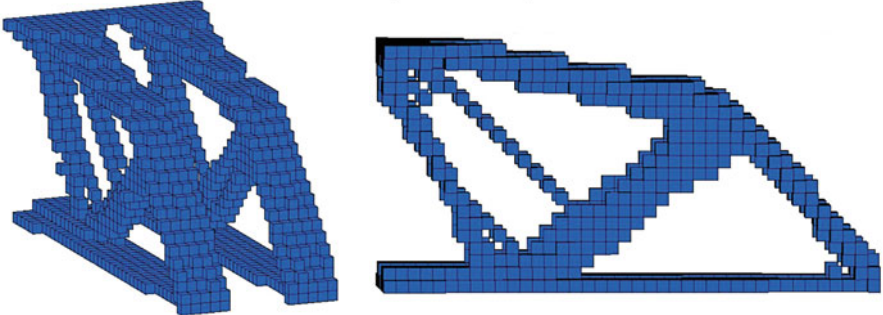
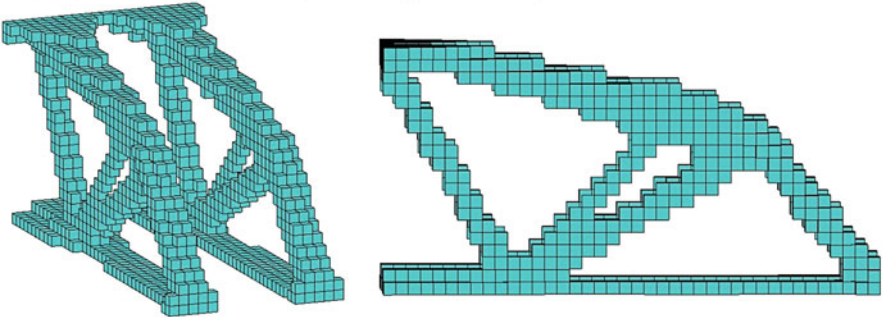


Table 6 Results for 3-D MBB beam with uniformly distributed load

CPD: $C = 7662.5989$, It. =46, Time=1249.1267



BESO: $C = 7745.955$, It. =55, Time=9899.0921



SIMP: $C = 12434.8629$, It. =1000, Time=5801.0065

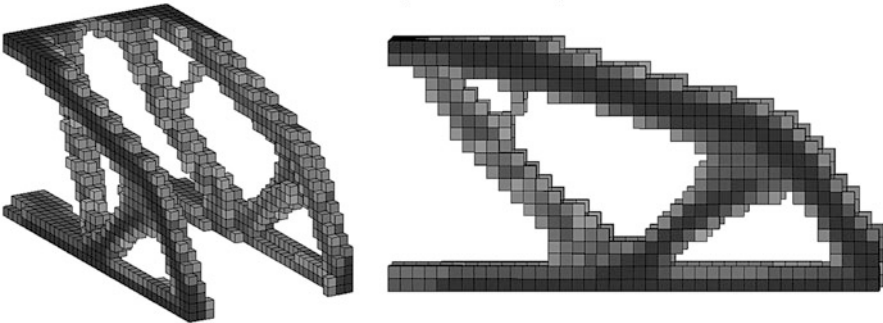
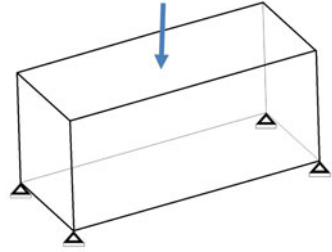


Fig. 9 3-D MBB beam with a central load



both BESO and SIMP. Again, the SIMP failed to converge in 1000 iterations and the “Change” $\Delta = 0.011 > \omega_1$ at the last iteration.

6.4 3D Wheel Problem

The 3D wheel design problem is constrained by planar joint on the corners with a downward point load in the center of the bottom as shown in Figure 11. The mesh resolution for this problem is $40 \times 20 \times 40$. The target volume is $V_c = 0.2$ and the parameters used are $\beta = 150$, $\mu = 0.94$, and $\omega_1 = 10^{-5}$. The optimal topologies produced by CPD, BESO, and SIMP are reported in Table 9. We can see that the CPD takes only about 18% and 32% of computing times by BESO and SIMP, respectively. Once again, the SIMP failed to converge in 1000 iterations and the “Change” $\Delta = 0.0006 > \omega_1$ at the last iteration.

For a given very small termination criterion $\omega_1 = 10^{-16}$ and for mesh resolution $30 \times 20 \times 30$, Table 10 shows effects of the parameters μ and V_c on the topology optimized results by CPD.

7 Concluding Remarks and Open Problems

We have presented a novel canonical penalty-duality method for solving challenging topology optimization problems. The relation between the CPD method for solving 0-1 integer programming problems and the pure complementary energy principle in nonlinear elasticity is revealed for the first time. Applications are demonstrated by 3-D linear elastic structural topology optimization problems. By the fact that the integer density distribution is obtained analytically, it should be considered as the global optimal solution at each volume iteration. Generally speaking, the so-called compliance produced by the CPD is higher than those by BESO for most of tested problems except for the MBB beam and the cantilever beam with a given hole. The possible reason is that certain artificial techniques such as the so-called soft-kill, filter, and sensitivity are used by the BESO method. The following remarks are important for understanding these popular methods and conceptual mistakes in topology optimization.

Table 7 Structures for 3-D MBB beam with a central load

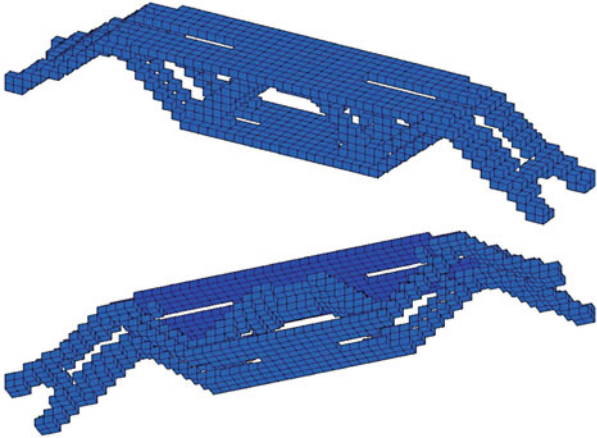
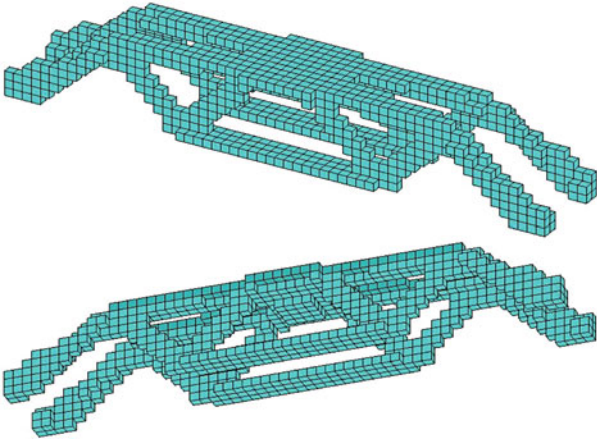
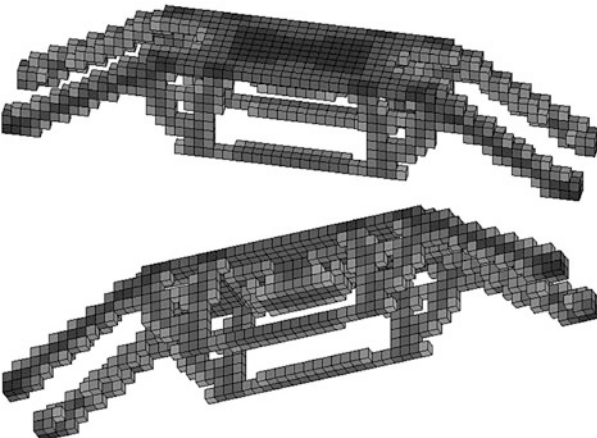
Method	Details	Structure
CPD	$C = 19.5313$ It. = 37 Time=48.2646	
BESO	$C = 20.1132$ It. =57 Time=1458.488	
SIMP	$C = 41.4099$ It. =95 Time=366.4988	

Fig. 10 Design domain for cantilever beam with a given hole

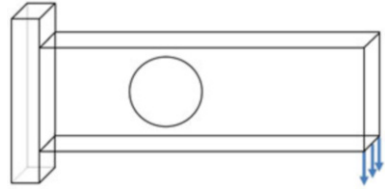
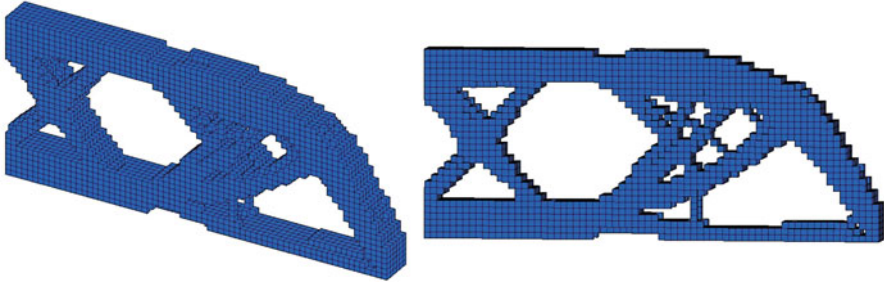
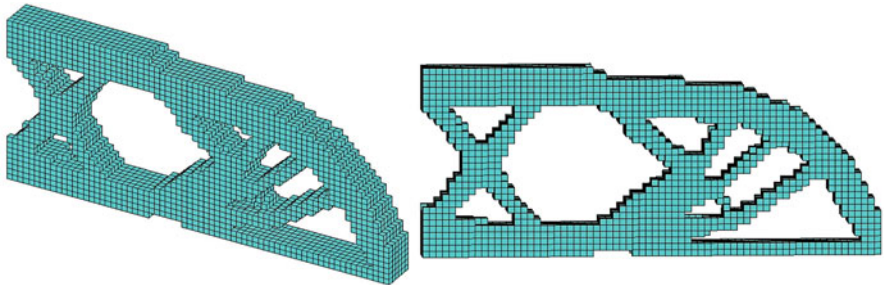


Table 8 Topology optimized structures for cantilever beam with a given hole

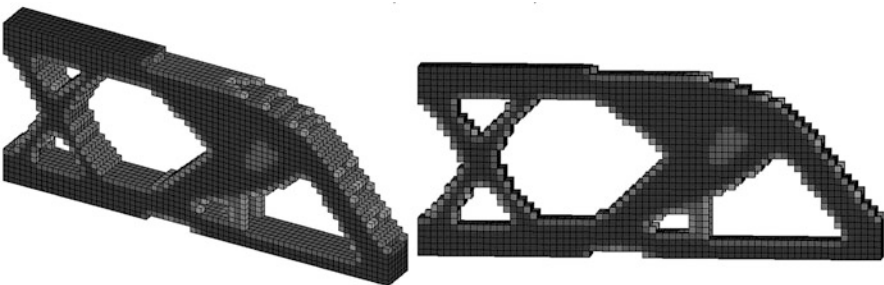
CPD: $C = 910.0918$, $It. = 14$, $Time = 74.61$



BESO: $C = 916.3248$, $It. = 21$, $Time = 1669.5059$



SIMP: $C = 997.1556$, $It. = 1000$, $Time = 1932.7697$



Remark 3 (On Penalty-Duality, SIMP, and BESO Methods) It is well-known that the Lagrange multiplier method can be used essentially for solving convex problem with equality constraints. The Lagrange multiplier must be a solution to the Lagrangian dual problem (see the Lagrange Multiplier’s Law in [12], page 36). For

Fig. 11 3D wheel problem

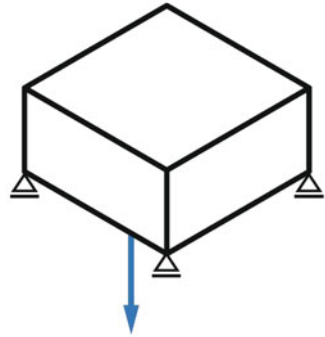


Table 9 Topology optimized results for 3D wheel problem ($40 \times 20 \times 40$) by CPD (left), BESO (middle), and SIMP (right)

$C = 3.6164$, It. =32 Time=6716.1433	$C = 3.6136$, It. =52 Time=37417.5089	$C = 3.7943$, It. =1000 Time=20574.8348

Table 10 Topology optimized results by CPD for 3D wheel problem ($30 \times 20 \times 30$) with two different views

$\mu = 0.88$, $V_c = 0.06$ $C = 5.7296$, It. =55 Time=2324.0445	$\mu = 0.88$, $V_c = 0.1$ $C = 4.2936$, It. =44 Time=1888.6451	$\mu = 0.92$, $V_c = 0.1$ $C = 4.3048$, It. =45 Time=1823.7826

inequality constraint, the Lagrange multiplier must satisfy the KKT conditions. The penalty method can be used for solving problems with both equality and inequality constraints, but the iteration method must be used. By the facts that the penalty parameter is hard to control during the iterations and in principle, needs to be large enough for the penalty function to be truly effective, which on the other hand, may cause numerical instabilities, the penalty method was becoming disreputable after the *augmented Lagrange multiplier method* was proposed in 1970s and 1980s. The augmented Lagrange multiplier method is simply the combination of the Lagrange multiplier method and the penalty method, which has been actively studied for more than 40 years. But, this method can be used mainly for solving linearly constrained problems since any simple nonlinear constraint could lead to a nonconvex minimization problem [34].

For example, let us consider the knapsack problem (\mathcal{P}_u). As we know that by using the canonical measure $\Lambda(\boldsymbol{\rho}) = \boldsymbol{\rho} \circ \boldsymbol{\rho} - \boldsymbol{\rho}$, the 0-1 integer constraint $\boldsymbol{\rho} \in \{0, 1\}^n$ can be equivalently written in equality $\boldsymbol{\rho} \circ \boldsymbol{\rho} - \boldsymbol{\rho} = \mathbf{0}$. Even for this most simple quadratic nonlinear equality constraint, its penalty function $W_\beta = \beta \|\boldsymbol{\rho} \circ \boldsymbol{\rho} - \boldsymbol{\rho}\|^2$ is a nonconvex function! In order to solve this nonconvex optimization problem, the canonical duality theory has to be used as discussed in Section 4. The idea for this penalty-duality method was originally from Gao's PhD thesis [7]. By Theorem 1, the canonical dual variable $\boldsymbol{\sigma}$ is exactly the Lagrange multiplier to the canonical equality constraint $\boldsymbol{\varepsilon} = \Lambda(\boldsymbol{\rho}) = \boldsymbol{\rho} \circ \boldsymbol{\rho} - \boldsymbol{\rho} = \mathbf{0}$, and the penalty parameter β is theoretically not necessary for the canonical duality approach. But, by this parameter, the canonical dual solution can be analytically and uniquely obtained. By Theorem 7 in [25], there exists a $\beta_c > 0$ such that for any given $\beta \geq \beta_c$, this analytical solution solves the canonical dual problem (\mathcal{P}_u^d), therefore, the parameter β is not arbitrary and no iteration is needed for solving the β -perturbed canonical dual problem (\mathcal{P}_β^d).

The mathematical model for the SIMP is formulated as a box-constrained minimization problem:

$$(P_{sp}) : \min \left\{ \frac{1}{2} \mathbf{u}^T \mathbf{K}(\boldsymbol{\rho}^p) \mathbf{u} \mid \mathbf{K}(\boldsymbol{\rho}^p) \mathbf{u} = \mathbf{f}, \mathbf{u} \in \mathcal{U}_a, \boldsymbol{\rho} \in \mathcal{Z}_b \right\}, \quad (63)$$

where $p > 0$ is a given parameter, and

$$\mathcal{Z}_b = \{ \boldsymbol{\rho} \in \mathbb{R}^n \mid \boldsymbol{\rho}^T \mathbf{v} \leq V_c, \boldsymbol{\rho} \in (0, 1]^n \}.$$

By the fact that $\boldsymbol{\rho}^p = \boldsymbol{\rho} \quad \forall p \in \mathbb{R}, \quad \forall \boldsymbol{\rho} \in \{0, 1\}^n$, the problem (P_{sp}) is obtained from (P_s) by artificially replacing the integer constraint $\boldsymbol{\rho} \in \{0, 1\}^n$ in \mathcal{Z}_a with the box constraint $\boldsymbol{\rho} \in (0, 1]^n$. Therefore, the SIMP is not a mathematically correct penalty method for solving the integer-constrained problem (P_s) and p is not a correct penalty parameter. By Remark 1, we know that the alternative iteration can't be used for solving (P_{sp}) and the target function must be written in terms of $\boldsymbol{\rho}$ only, i.e., $P_c(\boldsymbol{\rho}^p) = \frac{1}{2} \mathbf{f}^T [\mathbf{K}(\boldsymbol{\rho}^p)]^{-1} \mathbf{f}$, which is not a coercive function and, for any given $p > 1$, its extrema are usually located on the boundary of \mathcal{Z}_b (see [20]).

Therefore, unless some artificial techniques are adopted, any mathematically correct approximations to (P_{sp}) can't produce reasonable solutions to either (P_c) or (P_s) . Indeed, from all examples presented above, the SIMP produces only gray-scaled topology, and from Figure 3 we can see clearly that during the first 15 iterations, the structures produced by SIMP are broken, which are both mathematically and physically unacceptable. Also, the so-called magic number $p = 3$ works only for certain homogeneous material/structures. For general composite structures, the global min of $P_c(\rho^3)$ can't be integers [20].

The optimization problem of BESO as formulated in [30] is posed in the form of minimization of mean compliance, i.e., the problem (P) . Since the alternative iteration is adopted by BESO, and by Remark 1 this alternative iteration leads to an anti-Knapsack problem, the BESO should theoretically produce only trivial solution at each volume evolution. However, instead of solving the anti-Knapsack problem (16), a comparison method is used to determine whether an element needs to be added to or removed from the structure, which is actually a direct method for solving the knapsack problem (P_u) . This is the reason why the numerical results obtained by BESO are similar to that by CPD. But, the direct method is not a polynomial-time algorithm. Due to the combinatorial complexity, this popular method is computationally expensive and be used only for small-sized problems. This is the very reason that the knapsack problem was considered as NP-complete for all existing direct approaches.

Remark 4 (On Compliance, Objectivity, and Modeling in Engineering Optimization) By Wikipedia (see <https://en.wikipedia.org/wiki/Stiffness>), the concept of “compliance” in mechanical science is defined as the inverse of stiffness, i.e., if the stiffness of an elastic bar is k , then the compliance should be $c = 1/k$, which is also called the flexibility. In 3-D linear elasticity, the stiffness is the Hooke tensor \mathbf{K} , which is associated with the strain energy $W(\boldsymbol{\varepsilon}) = \frac{1}{2} \boldsymbol{\varepsilon} : \mathbf{K} : \boldsymbol{\varepsilon}$; while the compliance is $\mathbf{C} = \mathbf{K}^{-1}$, which is associated with the complementary energy $W^*(\boldsymbol{\sigma}) = \frac{1}{2} \boldsymbol{\sigma} : \mathbf{K}^{-1} : \boldsymbol{\sigma}$. All these are well written in textbooks. However, in topology optimization literature, the linear function $F(\mathbf{u}) = \mathbf{u}^T \mathbf{f}$ is called the compliance. Mathematically speaking, the inner product $\mathbf{u}^T \mathbf{f}$ is a scalar, while the compliance \mathbf{C} is a matrix; physically, the scalar-valued function $F(\mathbf{u})$ represents the external (or input) energy, while the compliance matrix \mathbf{C} depends on the material of structure, which is related to the internal energy $W^*(\boldsymbol{\sigma})$. Therefore, they are two totally different concepts, mixed using these terminologies could lead to serious confusions in multidisciplinary research⁷. Also, the well-defined stiffness and compliance are mainly for linear elasticity. For nonlinear elasticity or plasticity, the strain energy is nonlinear and the complementary energy can't be explicitly defined. For nonconvex $W(\boldsymbol{\varepsilon})$, the complementary energy is not unique. In these

⁷Indeed, since the first author was told that the strain energy is also called the compliance in topology optimization and (P_c) is a correct model for topology optimization, the general problem (P_{bi}) was originally formulated as a minimum total potential energy so that using $\mathbf{f} = \mathbf{K}(\rho)\bar{\mathbf{u}}$, $\min\{\Pi_h(\bar{\mathbf{u}}, \rho) \mid \rho \in \mathcal{Z}_a\} = \min\{-\frac{1}{2}\mathbf{c}(\mathbf{u})\rho^T \mid \rho \in \mathcal{Z}_a\}$ is a knapsack problem [18].

cases, even if the stiffness can be defined by the Hessian matrix $\mathbf{K}(\boldsymbol{\epsilon}) = \nabla^2 W(\boldsymbol{\epsilon})$, the compliance \mathbf{C} can't be well defined since $\mathbf{K}(\boldsymbol{\epsilon})$ could be singular even for the so-called G-quasiconvex materials [19].

Objectivity is a central concept in our daily life, related to reality and truth. According to Wikipedia, the objectivity in philosophy means the state or quality of being true even outside a subject's individual biases, interpretations, feelings, and imaginings.⁸ In science, the objectivity is often attributed to the property of scientific measurement, as the accuracy of a measurement can be tested independent from the individual scientist who first reports it.⁹ In continuum mechanics, it is well known that a real-valued function $W(\boldsymbol{\epsilon})$ is called to be objective if and only if $W(\boldsymbol{\epsilon}) = W(\mathbf{R}\boldsymbol{\epsilon})$ for any given rotation tensor $\mathbf{R} \in SO(3)$, i.e., $W(\boldsymbol{\epsilon})$ must be an invariant under rigid rotation (see [5], and Chapter 6 [12]). The duality relation $\boldsymbol{\epsilon}^* = \nabla W(\boldsymbol{\epsilon})$ is called the constitutive law, which is independent of any particularly given problem. Clearly, any linear function is not objective. The objectivity lays a foundation for mathematical modeling. In order to emphasize its importance, the objectivity is also called the principle of frame indifference in continuum physics [49].

Unfortunately, this fundamentally important concept has been mistakenly used in optimization literature with other functions, such as the target, cost, energy, and utility functions, etc.¹⁰ As a result, the general optimization problem has been proposed as:

$$\min f(x), \quad s.t. \quad g(x) \leq 0, \quad (64)$$

and the arbitrarily given $f(x)$ is called objective function,¹¹ which is even allowed to be a linear function. Clearly, this general problem is artificial. Without detailed information on the functions $f(x)$ and $g(x)$, it is impossible to have powerful theory and method for solving this artificially given problem. It turns out that many nonconvex/nonsmooth optimization problems are considered to be NP-hard.

In linguistics, a grammatically correct sentence should be composed by at least three components: subject, object, and a predicate. Based on this rule and the canonical duality principle [12], a unified mathematical problem for multi-scale complex systems was proposed by Gao in [16]:

$$(\mathcal{P}_g) : \quad \min\{\Pi(\mathbf{u}) = W(\mathbf{D}\mathbf{u}) - F(\mathbf{u}) \mid \mathbf{u} \in \mathcal{U}_c\}, \quad (65)$$

where $W(\boldsymbol{\epsilon}) : \mathcal{E}_a \rightarrow \mathbb{R}$ is an objective function such that the internal duality relation $\boldsymbol{\epsilon}^* = \nabla W(\boldsymbol{\epsilon})$ is governed by the constitutive law, its domain \mathcal{E}_a contains

⁸[https://en.wikipedia.org/wiki/Objectivity_\(philosophy\)](https://en.wikipedia.org/wiki/Objectivity_(philosophy)).

⁹[https://en.wikipedia.org/wiki/Objectivity_\(science\)](https://en.wikipedia.org/wiki/Objectivity_(science)).

¹⁰http://en.wikipedia.org/wiki/Mathematical_optimization.

¹¹This terminology is used mainly in the English literature. The function $f(x)$ is correctly called the target function in Chinese and Japanese literature.

only physical constraints (such as the incompressibility and plastic yield conditions [8]), which depends on mathematical modeling; $F(\mathbf{u}) : \mathcal{U}_a \rightarrow \mathbb{R}$ is a subjective function such that the external duality relation $\mathbf{u}^* = \nabla F(\mathbf{u}) = \mathbf{f}$ is a given input (or source), its domain \mathcal{U}_a contains only geometrical constraints (such as boundary and initial conditions), which depends on each given problem; $\mathbf{D} : \mathcal{U}_a \rightarrow \mathcal{E}_a$ is a linear operator which links the two spaces \mathcal{U}_a and \mathcal{E}_a with different physical scales; the feasible space is defined by $\mathcal{U}_c = \{\mathbf{u} \in \mathcal{U}_a \mid \mathbf{D}\mathbf{u} \in \mathcal{E}_a\}$. The predicate in (\mathcal{P}_g) is the operator “—” and the difference $\Pi(\mathbf{u})$ is called the target function in general problems. The object and subject are in balance only at the optimal states.

The unified form (\mathcal{P}_g) covers general constrained nonconvex/nonsmooth/discrete variational and optimization problems in multi-scale complex systems [23, 29]. Since the input \mathbf{f} does not depend on the output \mathbf{u} , the subjective function $F(\mathbf{u})$ must be linear. Dually, the objective function $W(\boldsymbol{\varepsilon})$ must be nonlinear such that there exist an objective measure $\boldsymbol{\xi} = \Lambda(\mathbf{u})$ and a convex function $\Psi(\boldsymbol{\xi})$, the canonical transformation $W(\mathbf{D}\mathbf{u}) = \Psi(\Lambda(\mathbf{u}))$ holds for most real-world systems. This is the reason why the canonical duality theory was naturally developed and can be used to solve general challenging problems in multidisciplinary fields. However, since the objectivity has been misused in optimization community, this theory was mistakenly challenged by M.D. Voisei and C. Zălinescu (cf. [23]). By oppositely choosing linear functions for $W(\boldsymbol{\varepsilon})$ and nonlinear functions for $F(\mathbf{u})$, they produced a list of “counterexamples” and concluded: “a correction of this theory is impossible without falling into trivial.” The conceptual mistakes in their challenges revealed at least two important truths: 1) there exists a huge gap between optimization and mechanics; and 2) incorrectly using the well-defined concepts can lead to ridiculous arguments. Interested readers are recommended to read recent papers [17] for further discussion.

For continuous systems, the necessary optimality condition for the general problem (\mathcal{P}_g) leads to an abstract equilibrium equation:

$$\mathbf{D}^* \partial_{\boldsymbol{\varepsilon}} W(\mathbf{D}\mathbf{u}) = \mathbf{f}. \quad (66)$$

It is linear if the objective function $W(\boldsymbol{\varepsilon})$ is quadratic. This abstract equation includes almost all well-known equilibrium problems in textbooks from partial differential equations in mathematical physics to algebraic systems in numerical analysis and optimization [48]¹². In mathematical economics, if the output $\mathbf{u} \in \mathcal{U}_a \subset \mathbb{R}^n$ represents product of a manufacture company, the input \mathbf{f} can be considered as the market price of \mathbf{u} , then the subjective function $F(\mathbf{u}) = \mathbf{u}^T \mathbf{f}$ in this example is the total income of the company. The products are produced by workers $\boldsymbol{\varepsilon} = \mathbf{D}\mathbf{u}$ and $\mathbf{D} \in \mathbb{R}^{m \times n}$ is a cooperation matrix. The workers are paid by salary $\boldsymbol{\varepsilon}^* = \nabla W(\boldsymbol{\varepsilon})$ and the objective function $W(\boldsymbol{\varepsilon})$ is the total cost. Thus, the

¹²The celebrated textbook *Introduction to Applied Mathematics* by Gil Strang is a required course for all engineering graduate students at MIT. Also, the well-known MIT online teaching program was started from this course.

optimization problem (\mathcal{P}_g) is to minimize the total loss $\Pi(\mathbf{u})$ under certain given constraints in \mathcal{U}_c . A comprehensive review on modeling, problems, and NP-hardness in multi-scale optimization is given in [21].

In summary, the theoretical results presented in this paper show that the canonical duality theory is indeed an important methodological theory not only for solving the most challenging topology optimization problems but also for correctly understanding and modeling multi-scale problems in complex systems. The numerical results verified that the CPD method can produce mechanically sound optimal topology, also it is much more powerful than the popular SIMP and BESO methods. Specific conclusions are given below:

1. The mathematical model for general topology optimization should be formulated as a bi-level mixed integer nonlinear programming problem (\mathcal{P}_{bl}). This model works for both linearly and nonlinearly deformed elastoplastic structures.
2. The alternative iteration is allowed for solving (\mathcal{P}_{bl}), which leads to a knapsack problem for linear elastic structures. The CPD is a polynomial-time algorithm, which can solve (\mathcal{P}_{bl}) to obtain global optimal solution at each volume iteration.
3. The pure complementary energy principle is a special application of the canonical duality theory in nonlinear elasticity. This principle plays an important role not only in nonconvex analysis and computational mechanics but also in topology optimization, especially for large deformed structures.
4. Unless a magic method is proposed, the volume evolution is necessary for solving (\mathcal{P}_{bl}) if $\mu_c = V_c/V_0 \ll 1$. But, the global optimal solution depends sensitively on the evolutionary rate $\mu \in [\mu_c, 1)$.
5. The compliance minimization problem (P) should be written in the form of (P_c) instead of the minimum strain energy form (P_s). The problem (P_c) is actually a single-level reduction of (\mathcal{P}_{bl}) for linear elasticity. Alternative iteration for solving (P_s) leads to an anti-knapsack problem.
6. The SIMP is not a mathematically correct penalty method for solving either (P) or (P_c). Even if the magic number $p = 3$ works for certain material/structures, this method can't produce correct integer solutions.
7. Although the BESO is posed in the form of minimization of mean compliance, it is actually a direct method for solving a knapsack problem at each volume reduction. For small-scale problems, BESO can produce reasonable results much better than by SIMP. But, it is time consuming for large-scale topology optimization problems since the direct method is not a polynomial-time algorithm.

By the fact that the canonical duality is a basic principle in mathematics and natural sciences, the canonical duality theory plays a versatile role in multidisciplinary research. As indicated in the monograph [12] (page 399), applications of this methodological theory have three aspects:

- (1) To check the validity and completeness of the existence theorems;
- (2) To develop new (dual) theories and methods based upon the known ones;
- (3) To predict the new systems and possible theories by the triality principles and its sequential extensions.

This paper is just a simple application of the canonical duality theory for linear elastic topology optimization. The canonical penalty-duality method for solving general nonlinear-constrained problems and a 66-line Matlab code for topology optimization are given in the coming paper [26]. The canonical duality theory is particularly useful for studying nonconvex, nonsmooth, nonconservative large deformed dynamical systems [13]. Therefore, the future works include the CPD method for solving general topology optimization problems of large deformed elastoplastic structures subjected to dynamical loads. The main open problems include the optimal parameter μ in order to ensure the fast convergence rate with the optimal results, and the existence and uniqueness of the global optimization solution for a given design domain V_c .

Acknowledgements This research is supported by the US Air Force Office for Scientific Research (AFOSR) under the grants FA2386-16-1-4082 and FA9550-17-1-0151. The authors would like to express their sincere gratitude to Professor Y.M. Xie at RMIT for providing his BES03D code in Python and for his important comments and suggestions.

References

1. Ali, E.J. and Gao, D.Y. (2017). Improved canonical dual finite element method and algorithm for post buckling analysis of nonlinear gao beam, *Canonical Duality-Triality: Unified Theory and Methodology for Multidisciplinary Study*, D.Y. Gao, N. Ruan and V. Latorre (Eds). Springer, New York, pp. 277–290.
2. Bendsøpe, M. P. (1989.). Optimal shape design as a material distribution problem. *Structural Optimization*, 1, 193–202.
3. Bendsøpe, M. P. and Kikuchi, N. (1988). Generating optimal topologies in structural design using a homogenization method. *Computer Methods in Applied Mechanics and Engineering*, 71(2), 197–224.
4. Bendsøpe, M. P. and Sigmund, O. (2004). *Topological optimization: theory, methods and applications*. Berlin: Springer-Verlag, 370.
5. Ciarlet, P.G. (1988). *Mathematical Elasticity*, Volume 1: Three Dimensional Elasticity. North-Holland, 449pp.
6. Díaz, A. and Sigmund, O. (1995). Checkerboard patterns in layout optimization. *Structural Optimization*, 10(1), 40–45.
7. Gao, D.Y. (1986). *On Complementary-Dual Principles in Elastoplastic Systems and Pan-Penalty Finite Element Method*, PhD Thesis, Tsinghua University.
8. Gao, D.Y. (1988). Panpenalty finite element programming for limit analysis, *Computers & Structures*, 28, 749–755.
9. Gao, D.Y. (1996). Complementary finite-element method for finite deformation nonsmooth mechanics, *Journal of Engineering Mathematics*, 30(3), 339–353.
10. Gao, D.Y. (1997). Dual extremum principles in finite deformation theory with applications to post-buckling analysis of extended nonlinear beam theory, *Appl. Mech. Rev.*, 50(11), S64–S71.
11. Gao, D.Y. (1999). Pure complementary energy principle and triality theory in finite elasticity, *Mech. Res. Comm.*, 26(1), 31–37.
12. Gao, D.Y. (2000). *Duality Principles in Nonconvex Systems: Theory, Methods and Applications*, Springer, London/New York/Boston, xviii + 454pp.
13. Gao, D.Y. (2001). Complementarity, polarity and triality in nonsmooth, nonconvex and nonconservative Hamilton systems, *Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences*, 359, 2347–2367.

14. Gao, D.Y. (2007). Solutions and optimality criteria to box constrained nonconvex minimization problems. *Journal of Industrial & Management Optimization*, 3(2) 293–304.
15. Gao, D.Y. (2009). Canonical duality theory: unified understanding and generalized solutions for global optimization. *Comput. & Chem. Eng.* 33, 1964–1972.
16. Gao, D.Y. (2016). On unified modeling, theory, and method for solving multi-scale global optimization problems, in *Numerical Computations: Theory And Algorithms*, (Editors) Y. D. Sergeyev, D. E. Kvasov and M. S. Mukhametzhanov, AIP Conference Proceedings 1776, 020005.
17. Gao, D.Y. (2016). On unified modeling, canonical duality-triality theory, challenges and breakthrough in optimization, <https://arxiv.org/abs/1605.05534> .
18. Gao, D.Y. (2017). Canonical Duality Theory for Topology Optimization, *Canonical Duality-Triality: Unified Theory and Methodology for Multidisciplinary Study*, D.Y. Gao, N. Ruan and V. Latorre (Eds). Springer, New York, pp.263–276.
19. Gao, D.Y. (2017). Analytical solution to large deformation problems governed by generalized neo-Hookean model, in *Canonical Duality Theory: Unified Methodology for Multidisciplinary Studies*, DY Gao, V. Latorre and N. Ruan (Eds). Springer, pp.49–68.
20. Gao, D.Y. (2017). On Topology Optimization and Canonical Duality Solution. Plenary Lecture at *Int. Conf. Mathematics, Trends and Development*, 28–30 Dec. 2017, Cairo, Egypt, and Opening Address at *Int. Conf. on Modern Mathematical Methods and High Performance Computing in Science and Technology*, 4–6, January, 2018, New Delhi, India. Online first at <https://arxiv.org/abs/1712.02919>, to appear in *Computer Methods in Applied Mechanics and Engineering*.
21. Gao, D.Y. (2018). Canonical duality-triality: Unified understanding modeling, problems, and NP-hardness in multi-scale optimization. In *Emerging Trends in Applied Mathematics and High-Performance Computing*, V.K. Singh, D.Y. Gao and A. Fisher (eds), Springer, New York.
22. Gao, DY and Hajilarov, E. (2016). On analytic solutions to 3-d finite deformation problems governed by St Venant-Kirchhoff material. in *Canonical Duality Theory: Unified Methodology for Multidisciplinary Studies*, DY Gao, V. Latorre and N. Ruan (Eds). Springer, 69–88.
23. Gao, D.Y., V. Latorre, and N. Ruan (2017). *Canonical Duality Theory: Unified Methodology for Multidisciplinary Study*, Springer, New York, 377pp.
24. Gao, D.Y., Ogden, R.W. (2008). Multi-solutions to non-convex variational problems with implications for phase transitions and numerical computation. *Q. J. Mech. Appl. Math.* 61, 497–522.
25. Gao, D.Y. and Ruan, N. (2010). Solutions to quadratic minimization problems with box and integer constraints. *J. Glob. Optim.*, 47, 463–484.
26. Gao, D.Y. and Ruan, N. (2018). On canonical penalty-duality method for solving nonlinear constrained problems and a 66-line Matlab code for topology optimization. To appear.
27. Gao, D.Y. and Sherali, H.D. (2009). Canonical duality theory: Connection between nonconvex mechanics and global optimization, in *Advances in Appl. Mathematics and Global Optimization*, 257–326, Springer.
28. Gao, D.Y. and Strang, G.(1989). Geometric nonlinearity: Potential energy, complementary energy, and the gap function. *Quart. Appl. Math.*, 47(3), 487–504.
29. Gao, D.Y., Yu, H.F. (2008). Multi-scale modelling and canonical dual finite element method in phase transitions of solids. *Int. J. Solids Struct.* 45, 3660–3673.
30. Huang, X. and Xie, Y.M. (2007). Convergent and mesh-independent solutions for the bi-directional evolutionary structural optimization method. *Finite Elements in Analysis and Design*, 43(14) 1039–1049.
31. Huang, R. and Huang, X. (2011). Matlab implementation of 3D topology optimization using BESO. *Incorporating Sustainable Practice in Mechanics of Structures and Materials*, 813–818.
32. Isac, G. *Complementarity Problems*. Springer, 1992.
33. Karp, R. (1972). Reducibility among combinatorial problems. In: Miller, R.E., Thatcher, J.W. (eds.) *Complexity of Computer Computations*, Plenum Press, New York, 85–103.

34. Latorre, V. and Gao, D.Y. (2016). Canonical duality for solving general nonconvex constrained problems. *Optimization Letters*, 10(8), 1763–1779.
35. Li, S.F. and Gupta, A. (2006). On dual configuration forces, *J. of Elasticity*, 84, 13–31.
36. Liu, K. and Tovar, A. (2014). An efficient 3D topology optimization code written in Matlab. *Struct Multidisc Optim*, 50, 1175–1196.
37. Marsden, J.E. and Hughes, T.J.R.(1983). *Mathematical Foundations of Elasticity*, Prentice-Hall.
38. Osher, S. and Sethian, J.A. (1988). Fronts propagating with curvature-dependent speed: algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79(1), 12–49.
39. Querin, O. M., Steven, G.P. and Xie, Y.M. (2000). Evolutionary Structural optimization using an additive algorithm. *Finite Element in Analysis and Design*, 34(3–4), 291–308.
40. Querin, O.M., Young V., Steven, G.P. and Xie, Y.M. (2000). Computational Efficiency and validation of bi-directional evolutionary structural optimization. *Comput Methods Applied Mechanical Engineering*, 189(2), 559–573.
41. Rozvany, G.I.N. (2009). A critical review of established methods of structural topology optimization. *Structural and Multidisciplinary Optimization*, 37(3), 217–237.
42. Rozvany, G.I.N., Zhou, M. and Birker, T. (1992). Generalized shape optimization without homogenization. *Structural Optimization*, 4(3), 250–252.
43. Sethian, J.A. (1999). Level set methods and fast marching methods: evolving interfaces in computational geometry, fluid mechanics, computer version and material science. *Cambridge, UK: Cambridge University Press*, 12–49.
44. Sigmund, O. and Petersson, J. (1998). Numerical instabilities in topology optimization: A survey on procedures dealing with checkerboards, mesh-dependencies and local minima. *Structural Optimization*, 16(1), 68–75.
45. Sigmund, O. and Maute, K. (2013). Topology optimization approaches: a comparative review. *Structural and Multidisciplinary Optimization*, 48(6), 1031–1055.
46. Sigmund, O. (2001). A 99 line topology optimization code written in matlab. *Struct Multidisc Optim*, 21(2), 120–127.
47. Stolpe, M. and Bendsoe, M.P. (2011). Global optima for the Zhou–Rozvany problem, *Struct Multidisc Optim*, 43, 151–164.
48. Strang, G. (1986). *Introduction to Applied Mathematics*, Wellesley-Cambridge Press.
49. Truesdell, C.A. and Noll, W. (1992). *The Non-Linear Field Theories of Mechanics*, Second Edition, 591 pages. Springer-Verlag, Berlin-Heidelberg-New York.
50. Xie, Y.M. and Steven, G.P. (1993). A simple evolutionary procedure for structural optimization. *Comput Struct*, 49(5), 885–896.
51. Xie, Y.M. and Steven, G.P. (1997). Evolutionary structural optimization. *London: Springer*.
52. Zuo, Z.H. and Xie, Y.M. (2015). A simple and compact Python code for complex 3D topology optimization. *Advances in Engineering Software*, 85, 1–11.
53. Zhou, M. and Rozvany, G.I.N. (1991). The COC algorithm, Part II: Topological geometrical and generalized shape optimization. *Computer Methods in Applied Mechanics and Engineering*, 89(1), 309–336.

Part II
High Performance and Scientific
Computing

High Performance Computing: Challenges and Risks for the Future



Michael M. Resch, Thomas Boenisch, Michael Gienger, and Bastian Koller

1 Introduction

High performance computing (HPC) is facing considerable challenges in the coming years. While for some time processor and system architectures were considered the most crucial problems [6, 7], today challenges go well beyond the mere discussion of the hardware. In this paper, we will present and discuss the most recent trends in hardware development and will explore how these trends will influence HPC, and what challenges and opportunities come with it.

Even though the question of architecture is perhaps no longer the most pressing one, the first problem that HPC will have to handle in the future is a hardware problem: the end of Moore's law [5]. The prediction of Moore in 1965, that we would be able to cram ever more transistors on the same surface, with the numbers doubling every 12 – later Moore shifted this to 18 – months, did hold for about 50 years and predictions were given in 2015 that it might hold for another decade.¹ Currently, there is doubt whether it can be extended even in the coming few years. Traditional assumptions no longer hold and no longer can be considered to be the main driving factor for hardware improvement. We will explore what impact this might have on HPC.

¹2015 International Technology Roadmap for Semiconductors (ITRS): http://www.semiconductors.org/main/2015_international_technology_roadmap_for_semiconductors_itrs/.

M. M. Resch (✉) · T. Boenisch · M. Gienger · B. Koller
High Performance Computing Center Stuttgart (HLRS), University of Stuttgart, Nobelstrasse 19,
70569 Stuttgart, Germany
e-mail: resch@hlrs.de; boenisch@hlrs.de; gienger@hlrs.de; koller@hlrs.de

With traditional hardware improvement making it much more difficult to squeeze more performance from any given HPC architecture, the focus of attention is slowly shifting towards software and towards mathematical methods for HPC simulation. Over the last decades, various investigations show that mathematical methods – at least for the solution of systems of equations – have contributed substantially to the increase in performance [4]. Algorithms will play a more crucial role and this paper will explore how this will happen and what the key questions from the point of view of HPC will be.

Another trend that has a huge impact on HPC is what can best be described as *Big Data*. Even though usage of data goes way beyond the original idea of handling large amount of data, the term Big Data still in a sense is useful as it describes well how HPC may be overwhelmed by the data problem in the coming years. HPC may well become a smaller part of a larger digital infrastructure that is focusing around data rather than around compute power. We will address how this will impact HPC.

2 The End of Moore's Law

2.1 Technical Limits

Technically, HPC is facing the end of a development that used to be called Moore's law. Processor clock frequencies, which carried the main load of speeding up hardware cannot be further increased – a fact that can be seen in processor industry already since about 2004. Clock frequencies in standard processors as used in HPC systems reached a peak of about four GHz and are currently hovering at about two to three GHz. An increase in clock frequency would increase leakage and hence, make cooling of such a processor difficult if not impossible.

The way to get to more performance is parallelism and multicore processors have hence become a standard. Parallelism was introduced decades ago, but became the key technology for HPC only in the last two decades. A typical HPC processor is currently based on 16 to 32 cores – each of which is using internal parallelism in functional units to boost the theoretical number of operations per clock cycle. In order to avoid a too-high power consumption of such processors, the clock frequency is typically lower than it could theoretically be. Power budgets have become a main issue when it comes to processor design.

The so-called accelerators provide solutions that rely on simplified core architectures, but allow to push the number of cores on a single chip to extremes. Thousands of cores on a single accelerator chip allow much higher theoretical peak performance than standard processors. The classical accelerator approach aims at high peak performance based on simplified cores running at lower clock frequencies. As a result, higher peak performance seems to be achievable within a still reasonable power budget. Unfortunately, such accelerators leave the users

with the need to adapt their codes to a new architecture and a new programming model. The costs for such adaptation are high, given that there is still no agreed open standard for the programming of various accelerators. At the same time, the parallelism of accelerators typically requires single instruction multiple data (SIMD) parallelism as described by Flynn [2].

For some time, it was expected that these two developments (parallelism in processors and in system architectures) would keep pushing HPC to ever higher levels of performance with the Exaflop being the target for 2020. Ever large systems based on a mix of standard processors and accelerators were built. A most recent update of the TOP500² list shows this trend over the last decade. However, recent investigations show that the required reduction of feature size on chips – necessary to further increase the number of cores on a chip – is no longer to be expected over the next decade [1]. We have to explore what this will mean to HPC and how we can react to this in terms of architectures.

2.2 What Does It Mean for Architectures?

First of all, we have not yet reached the end of the line. Transistors keep shrinking for a number of years still – even if it will only be for five to eight years. This will further push the degree of parallelism we will see in standard processors (moving from tens of cores to probably even 100 cores) and accelerators (moving from thousands of cores to potentially ten thousand cores). The basic concept of HPC architectures will theoretically not have to change during this time frame. Large-scale systems will potentially host a mix of standard processors and accelerators. Accelerators may be as different as a graphic card or a vector processor, but they will most certainly be included in any top 20 systems in the world over the next decade.

Given the further progress to be expected, such architecture will be able to reach a peak performance of about one Exaflop most likely after 2020 – as hardware vendors and national large-scale HPC projects keep adapting their roadmaps shifting the first Exaflop to ever later dates. The exact year is hard to predict as the race for this prestigious number is not only motivated by technical goals but at least as much by political considerations. It would lead us too far away from the purpose of this paper if we discussed the political implications of the end of Moore's law. However, it suffices to say that the level of disappointment in the political arena will increase with ever higher budgets that are necessary to buy faster systems – given that only an increase in budget allows to further increase peak performance at a certain point as increase in speed is only possible by increasing the number of compute nodes available in a single system.

²The Top 500 list: www.top500.org.

3 How Do We Move Forward in HPC Architectures?

Assuming that an increased processor speed is no longer possible and hence won't provide the necessary increase in peak performance, there are two potential ways out.

First, we can – and actually already do – further increase the number of processors in an HPC system. This will give higher peak performance but will require higher budgets and will make it more difficult to program such systems. Furthermore, this will increase our need for space, power, and cooling. As a side effect, the core of costs for HPC is shifting from investment costs to infrastructure costs and operational costs. As of today, operational and infrastructure costs can be even higher than investment costs. Consequently, this will become worse in the coming years.

Second, we can simply give up the notion that higher peak performance is better. This is not to say that peak performance is at all irrelevant. The Exaflop will allow to tackle some of the most challenging problems in science and will help to achieve new breakthroughs. However, it is about time that we admit that transistor technology has reached a level at which it is simply not viable to go any further. While costs are exploding, scientists struggle with getting sustained performance from ever more complex systems. To admit that a technology has matured and no longer offers huge increases in performance is not an unusual thing to say when it comes to modern technologies. Automotive industry has long given up on maximum speed or maximum horse power as the key performance indicator for cars. Aerospace industry has built the Concorde to push the maximum speed of a civil aircraft beyond the speed of sound but stopped following that path and finally even terminated the ill-fated Concorde.

For HPC, this means that we need to define a different metric to measure the power of an HPC system. New approaches for benchmarking like HPCG³ or HPGMG⁴ are available, but suffer all from the same weakness: they only measure performance for one typical method but give no real indication of the overall level of sustained performance a user may expect from a system. On the other hand, any comprehensive suite of benchmarks would make it virtually impossible to run all benchmarks to get to a reasonable evaluation of the system. Already today, the Linpack benchmark as used in the TOP500 takes way too long for a public computing center to be considered a reasonable benchmarking approach. It is mostly only run to make sure that there is an entry to the TOP500 list.

³www.hpcg-benchmark.org.

⁴<https://hpgmg.org>.

4 The Opportunities of Software and Algorithms

With hardware not offering the same potential increase in performance as over the last decades, the focus of attention starts to shift to other components of an HPC simulation. Basically, three aspects have to be considered and show some potential for the user and for computing centers: programming models, power efficiency, and new algorithms.

4.1 *Programming Models*

Current programming models were developed when parallel computers were using hundreds or at most thousands of cores. The standardization, both for message passing (MPI) and for shared memory parallelism (OpenMP), did make sense at the time when they were developed and were a great step forward since they provided a standardized way of writing portable programs for parallel systems. However, today these approaches are increasingly facing challenges, both for their implementation and for their users.

Given the hierarchical nature of modern HPC systems with a cascade of logical units like racks, nodes, and processors, what is required are models that reflect this hierarchy. One approach could certainly be – and actually it is used already widely in the community – a combination of MPI and OpenMP often described as hybrid programming. However, there are two drawbacks for this model. First, it simplifies the view of the architecture to two levels – inside a shared memory node and across several such nodes. With architectures that show several levels of parallelism, this is a mismatch of programming model and hardware architecture. Second, it does not make the life of programmers much easier given that process management – as is required by MPI – is still extremely difficult when the number of nodes is in the tens of thousands.

4.2 *Power Efficiency*

Over the last years, a lot of effort has been invested in evaluating the potential on saving power in HPC simulations. Given that the power consumption of a processor depends on the usage of the processor and hence, on the programming style of the user, investigations were started to find out about power saving programming models. The key finding of such investigations can easily be summarized. It turns out that all kind of communication is causing most power consumption and that hence a basic rule for power saving is: do not communicate.

Now, this is exactly what we want to do when we optimize our parallel programs. Both at the node level and in internode communication, we try to avoid communication as much as possible. Hence, power saving and performance

increase coincide favorably such that we can focus in our optimization on avoiding communication and will both increase performance and reduce power consumption.

However, avoiding communication is not always possible. Densely coupled problems require communication, and a decoupling of computational parts of a program is not always feasible. Sometimes, it even leads to wrong results as the software-wise decoupling of a problem that is physically densely coupled leads to non-convergence or effects like oscillations in the solution caused by the decoupling method[8].

4.3 New Algorithms

Finally, we need to explore the potential of algorithms to better exploit existing hardware architectures. As we have seen above, algorithms did contribute substantially to the increase in performance for HPC simulations. The speedup through algorithms is comparable to the speedup through clock frequency increase over several decades. Multigrid methods are orders of magnitudes faster than the classical methods with which we started computations back in the 1950s.

The key factor in designing algorithms will be to consider the rule: do not communicate. Modern algorithms will hence have to devise ways of providing locality as much as possible without losing the overall convergence for the solution of the simulation problem. It does make sense to look for ways that might even increase the number of floating point operations if such an increase comes with a substantial decrease in communication.

5 The Impact of Data on HPC

Even though a traditional look at HPC already shows some dramatic changes, there is something that might be even more important for HPC. Considering the current trends, we find that HPC is going to be just a small part of something bigger – which is Data. It is meanwhile well accepted that the value is in the data and not so much in the simulation itself. This is also true for data not coming from a simulation.

Sources of data are manifold: from the traces each person is generating each day using systems in the Internet, when shopping, communicating, watching movies, or visiting other web pages. There are business data operations, which are digitally available and stored for years. There is an increasing amount of sensors everywhere especially, powered by the Internet of Things, going from production lines to personal homes. Smart meters are a good example for that. And last but not least, there is an increasing amount of data generated by technical and scientific simulations.

Several companies already own a huge amount of data which they cannot handle or analyze anymore in a traditional way. Very often such data are only stored for

later analysis, but get lost when storage technology changes. In many cases, data are collected but are not available or cannot be analyzed sufficiently when problems arise. A typical example is the new technology of Industrie 4.0, as it is called in Germany. With production machinery collecting data and communicating among each other, the factory of the future turns into a gigantic pool of processing and communication. The heap of data created requires analysis not only for the sake of improving production but also to be able to detect errors and identify sources of problems. As a consequence, we see an increasing number of problems, where an increased size of data to be analyzed requires increased computational capacity. On the other hand, HPC creates very large data sets that require data analytics capabilities that go way beyond traditional visualization.

5.1 What Does this Mean for HPC?

In most cases today, data analytics or big data are widely disconnected from HPC. Both technologies keep pushing their limits and are developed in different departments both in research and in industry. However, in many cases there is a convergence visible or might be beneficial in the near future. Even in cases where this is not visible today. Fraud detection is a well-known application field of data analytics meanwhile. Banks for example are using systems to detect the misuse of credit cards. They have only very little time to decide about a transaction. To meet those requirements, HPC technology is used to speed up the analytics process.

For many business cases where data analytics is done in large in-memory databases, nobody is thinking about HPC today. However, the next step after the analysis of business data could be the question of making the next step in business logic. In many cases, this could lead to the requirement of large simulations and parameter studies, which will naturally require HPC systems. This is already visible in railroad companies, where in case of delays simulations are used to decide between different opportunities to improve the current traffic situation.

The increasing use and number of linked sensors is another area where data volumes are exploding. This leads to the idea of in-time analytics to detect events before they occur for example with machine learning technologies. In several cases, this leads to new insights and the information about existing dependencies. The interest of further understanding and analysis is leading to the requirement of solving inverse problems to get further insight into the system behavior. This will increase the requirement for HPC in the future.

Another example with even higher impact on HPC is the usage of sensors to detect major events, which might lead into disasters like earthquakes and tsunamis [3]. After the detection of an event, an urgent simulation is required to understand the potential effects of this event. It needs to be decided how critical the event is and what should be recommended for the public and for officials. Therefore, urgent computing is required also in the frame of HPC, which changes the operation models of an HPC center drastically.

The simulation field itself is causing a data problem, too. With the compute power available today, the resulting data sets are often huge, and the current methodology to analyze the results does not work in many cases as it often does not scale, either due to manual invention or as the used hardware or software provides limits. One way to reduce the amount of data to be stored is the usage of lossy data compression. There are efforts going on to learn, for which use cases which compression method and what factor of information loss is acceptable [9]. In other fields, like meteorology, measurement data shall be integrated into the simulation. This leads to questions of data assimilation and how to perform that in an HPC environment.

In several fields, the question of some automatic data pre-analysis is coming up. Again in meteorology, it could become interesting to use the current methods from data analytics to detect events in the large result data sets. This would significantly reduce the amount of time required for the analysis steps. In CFD as another example, researchers are developing methods to further analyze and categorize turbulence. Actual methods require reading the whole time, depending on simulation data about ten times. Other methods under development will increase the data volume by a factor of ten and more before reducing the amount data in the end.

5.2 The Impact on HPC Environments and Architectures

The described development will have a significant impact on architectures and on HPC environments. In several cases, a direct connection to include up-to-date input data into the on-going simulations will require a change in the HPC environment setup and will require solving new security issues. Additionally, there is the upcoming requirement for urgent computing, which needs to be solved administratively as well as technically as many HPC systems are not prepared for such a requirement. However, as most of today's systems are publicly funded, a future community requirement cannot be ignored and must be handled.

The requirement of handling large amounts of data in an HPC system requires some changes in the architecture or system setup. The solution width is quite high and might go from the integration of a data-intensive computing platform into the HPC environment up to a fully integrated system, which is able to handle huge amounts of data with a high speed as well as compute-intensive jobs. Fortunately, there are some recent and on-going developments in the memory and storage hardware section, which can help here in future systems. The integration of burst buffers as a fast intermediate storage to speed up I/O is only a first step into this direction. The on-going development of low latency, high bandwidth storage devices of the NVRAM class will provide further support and has the potential to act as a game changer for the integration of data analytics and HPC.

6 Conclusions

Summarizing our findings, we clearly identify three major trends. First, Moore's law is coming to an end, which will make it much more difficult to achieve higher peak performance by hardware improvement. As a result of this, the focus of attention will dramatically shift from hardware to software and algorithms. This will cover programming models as well as mathematical methods and potentially even the types of methods we will employ on modern HPC architectures. As this happens, mathematicians will gain more relevance in HPC simulation compared to computer scientists who will help to widen the HPC community. The most important change over the coming years may, however, be the impact that data will have on HPC. Even if big data is not going to entirely replace HPC over the coming years, the focus on data and its exploration and interpretation will start to outgrow the current focus on simulation and computation. This will open new fields and new communities for HPC and should be considered a chance for both, HPC and the scientific community in general.

References

1. Courtland, R.: Transistors could stop shrinking in 2021. In: IEEE Spectrum <http://spectrum.ieee.org/semiconductors/devices/transistors-could-stop-shrinking-in-2021> (2016). Cited 24 Jan 2018
2. Flynn, M.J.: Some Computer Organizations and Their Effectiveness. IEEE Trans. Comput. C-21 (9), 948–960 (1972)
3. Kobayashi, H.: A Case Study of Urgent Computing on SX-ACE: Design and Development of a Real-Time Tsunami Inundation Analysis System for Disaster Prevention and Mitigation. In Resch, M.M., Bez, W., Focht, E., Patel, N., Kobayashi, H. (eds.) Sustained Simulation Performance 2016, pp 131–138. Springer (2016)
4. Marra, V.: On Solvers: Multigrid Methods. In: Comsol Blog <https://www.comsol.com/blogs/on-solvers-multigrid-methods/> (2013). Cited 24 Jan 2018
5. Moore, G.E.: Cramming more components onto integrated circuits. Electronics **38**(8), 114–117 (1965)
6. Resch, M.M.: Trends in Architectures and Methods for High Performance Computing Simulation. In Topping, B.H.V., Iványi, P. (eds.) Parallel Distributed and Grid Computing for Engineering, pp 37–48. Saxe-Coburg Publications, Stirlingshire, Scotland (2009)
7. Resch, M.M.: High Performance Computing Architectures: Trends, Opportunities and Challenges. In Iványi, P., Topping B.H.V. (eds.) Techniques for Parallel, Distributed and Cloud Computing in Engineering, pp 1–9. Saxe-Coburg Publications (2015)
8. Resch, M., Rantza, D., Stoy, R.: Meta-computing Experience in a Transatlantic Wide Area Application Test bed. Future Generation Computer Systems (15) 5–6, 807–816 (1999)
9. Vogler, P.: Data compression strategies for exascale CFD simulations. In: Exaflow Project <http://exaflow-project.eu/index.php/news/31-data-compression-strategies-for-exascale-cfd-simulations> (2017). Cited 24 Jan 2018

Modern Parallel Architectures to Speed Up Numerical Simulation



Mikhail Lavrentiev, Konstantin Lysakov, Alexey Romanenko,
and Mikhail Shadrin

1 Introduction

Today, such words as high performance computing (HPC) have become quite common. The lists of top 500 (see <http://top500.org>) of most productive cluster systems and the lists of Green-500 (cf. <http://green500.org>) of most energy-efficient computer systems are being published on a regular basis. Systems in the top 9 places in the lists apply as a co-processor, a GPU NVIDIA Tesla P100. The development of such computing capabilities has now turned into reality the possibility of solving applied and theoretical problems of such scale and complexity which used to be unattainable in the not so distant past. All leading universities in the world, medium to large research centres, and commercial companies invest into the development of their own HPC systems, or purchase these type of computing powers on the market.

At the same time, very few problems require all the capabilities of supercomputers from the top 500. Most applied and research challenges can be effectively solved with an application of modern personal computers with the use of proper software (algorithms) and hardware (architecture). The latter means graphics processing units (GPUs) and field programmable gate arrays (FPGAs).

The GPU is a processor with the SIMD architecture, which, according to [4], can be used to perform in parallel same operations with different data sets. Although it was initially applied in graphics only, with its more expanded use in other

M. Lavrentiev (✉) · A. Romanenko
Novosibirsk State University, Pirogov st., 1, Novosibirsk 630090, Russian Federation
e-mail: mmlavr@nsu.ru; romanenko.alexey@gmail.com

K. Lysakov · M. Shadrin
Institute of Automation and Electrometry, Ac. Koptyug pr., 1, Novosibirsk 630090, Russian Federation
e-mail: lysakov@sl.iae.nsk.su; miksha@sl.iae.nsk.su

areas, most challenges of computer modelling with data sets being processed in parallel can be well supported by this architecture. Compared to processors of general purpose, the main part of the GPU consists of relatively simple computing elements [2]. This type of architecture also enables the increase in the conductivity of memory bus, which is quite important to boost a computing power. To date, the peak performance of a single modern GPU card can be as high as 15 TFLOPS [21].

One of the key advantages in using hybrid computing systems is its energy efficiency. Let us compare 2 systems from the top 500: Cray CX50 with Xeon E5-2690v3 processor and GPU NVIDIA Tesla P100 [20], rated as #3, and Cray CX40, rated as #10, which only uses central processors [23]. Normalizing peak performance reveals that a hybrid system consumes 4 times less energy. According to the conference GTC-2017 [7] discussion, the Volta GPU shows as many as 50 GFLOPS per watt in single precision [19].

A brief description of FPGA is provided below. Gates array programmed by the user has been applied for multiple purposes since the 1990s. Initially, different PLD/CPLD with energy independent memory were used the most. PLD was successfully applied to develop small autonomous devices and controllers.

The introduction of FPGA (field-programmable gate array) has made it possible to have many more logical elements while, at the same time, to be able to configure the same microchip unlimited number of times. FPGA enables to code complex mathematical algorithms for data processing. It also allows to perform prototyping of computing devices for the further production of serial custom-made ASIC microcircuits.

However, until recently, the application of FPGA had two main obstacles, involving the necessity for manual coding of algorithms up to the register transfer level (RTL). This is time consuming and requires highly qualified developers.

Until recently, typical tools for development automation (e.g. Simulink, MatLab, SystemVerilog, and SystemC) have provided poor performance and unnecessary use of hardware resources compared to the RTL model.

The modern high-level synthesis (HLS) technology [9] uses the C-type language to describe digital schemes. This is the new level in computation devices development, presenting the entire cycle from architecture description to results verification by means of modelling tools. HLS is nothing but an automated design process, which interprets the algorithmic description of model and allows to create digital devices, meeting the prescribed conditions. The HLC technology easily allows to change the pipeline parameters, adopting to timing or utilization requirements. It also allows to verify the code prior to its transformation to a scheme for a particular FPGA microchip.

Thus, together the HLS technology and extended resources of modern FPGA microchips make it possible to utilize FPGA to answer new challenges in various application areas. Modern devices are able to support hundreds of thousands of parallel processes; onboard memory is compared to hundreds of MB allowing to construct wide and deep computation pipelines. Moreover, FPGA is software reconfigurable, i.e., the connections among computation primitives and onboard memory are defined by the user. Such system easily adopts the computation

architecture to a particular algorithm and is able to modify the given algorithm for a given hardware. One can configure FPGA microchip unlimited number of times; hence, the same hardware device could be used to solve different problems. FPGA-based devices could be used as a part of personal computer or as a standalone hardware. Therefore, autonomous data processing energy efficient devices of small sizes could be easily developed.

In the present paper we stress the performance gain results from several areas, achieved by the use of GPU and FPGA architectures.

2 Examples of Using GPU to Improve Code Execution Performance in Geophysical Problems

Extended computation power of the modern graphic processors draws the attention of developers worldwide. Today, there are several hundreds of open source packages and libraries, adopted for GPU use [10]. Depending on the problem type, performance gain from tens to hundreds of times is being reported. The authors achieve valuable code acceleration (up to tens and even hundred times), while using one GPU compared to one socket of the central processor. All the examples are practical problems in the area of Earth Sciences, namely seismic data processing and tsunami wave simulation.

2.1 Decomposition of Seismic Records by Wave Packages

A part of seismic studies on the earth surface is the measurement of elastic waves, reflected from the geological interfaces, boundaries between the layers in the Earth interior. The recorded wave field is used to construct the seismic sections of the Earth crust. The latter is then used for mineral resources exploration. The data, concerning reflecting waves, are multi-dimensional, very large (up to terabytes), and irregular. Therefore, an important problem is the efficient data representation. It is natural to use the data decomposition with respect to a certain basis, which fits better the sequel data processing. The so-called wave packages may serve as such a basis, being convenient ‘bricks’ to build the initial seismic waves. Mathematically proven that such basis is optimal is given in [1]. The method of decomposition with respect to wave packages is also useful to solve such problems in geophysics as: data compression, noise depression, interpolation, regularization (data recalculation from an arbitrary mesh to a regular one), and others (see details in [8, 16, 17], e.g.).

Algorithms for direct and inverse transformation with respect to wave packages are described in [3]. Without going into details, we stress here that the unequally spaced fast Fourier transform (USFFT) is used. This is done through data interpolation for a regular mesh and the use of the FFT library function.

In our case, the data interpolation (from unequally spaced to regular and back) is the most time consuming. However, interpolation at each point is independent, the algorithm is the same for all points. Thus, all the power of modern GPU could be used.

By using the NVIDIA CUDA technology, the GPU-based fast algorithms for direct and inverse data decomposition with respect to 3D wave packages have been developed. A number of optimizations were performed, accounting for the GPU architecture and decomposition algorithms structure. Performance gain was **45** times at a single GPU. Exploiting peculiarities of the seismic data structure, code was adopted to use multiple GPUs. Practically linear scalability has been achieved, namely **3.93** times code acceleration at 4 GPUs and **7.70** times acceleration at 8 GPUs. Thus, performance gain is 350 times at 8 GPUs compared to a single CPU.

Details for the code translation and optimization are given in [18].

2.2 The Convolution Problems with Green's Function

As mentioned earlier, seismic uses the reflecting waves to reconstruct, layer by layer, the Earth interior structure. Many of these reconstruction methods (the so-called methods for the inverse seismic problems) are based on the Green's function, presenting the wave field from the point source.

The recently developed theory of passing–spreading operators provides the exact analytical description of the Green's function even for rather complicated media. The theory is based on the analytical solution to a direct problem for inhomogeneous media with arbitrary smooth boundaries in a form of superposition of wave signals of many times reflected and refracted waves. Each separate signal is described as a composition of propagation operators of a convolution type at smooth boundaries. In such a way it is possible to reconstruct a structure of the Green's function, which is similar to a recorder wave field.

However, this approach has a rather high computational cost and requires extended memory resources. The propagation and reflection operators have N^2 dimension, N being the number of triangles, describing the boundary between layers. The typical value of parameter N in real problems is compared to 10^5 . Moreover, theory is constructed in the frequency domain. So, all the vector-matrix operations should be done for each value of frequency.

The matrix-vector type algorithm was expected to be well parallelizable. Indeed, the use of Intel MKL library results in linear scalability with respect to a number of cores in use. As for the GPU, cuBLAS library procedures have been used. Even for not the newest NVIDIA Tesla C2070, the computation time was only 10 minutes, or **162** times faster the original code version. Details are given in [24].

2.3 Modelling Tsunami Wave Propagation

The Great Tohoku Earthquake of 2011 (Japan) shows that the population and industry of the coastal regions are not protected from disasters of a seismic nature. Fast and reliable evaluation of tsunami wave parameters is needed prior to the moment when the wave approaches the shore to reduce dramatic effects of tsunami wave. In case of Japan it is only about 20 minutes. Worldwide tsunami wave propagation is approximated according to the shallow-water theory (both linear and non-linear). These models reflect rather accurately the basic wave parameters (propagation time period from the source to the recording device and wave amplitudes) even for a fairly rough numerical bathymetry under assumption that the initial displacement in the source is known. There are several software packages for modelling wave propagation throughout the ocean and the wave run-up heights. The most well-known software packages are MOST and TUNAMI.

MOST (method of splitting tsunami), proposed in [22], allows to make a forecast of the flood region in real-time mode using tsunameters’ data. This package is mostly used in the USA for charting inundation maps [5]. The online version of MOST, comMIT, is also available.

The first attempts to speed up the so-called slitting method were performed by the authors in 2008. We achieved 60x speedup on CELL BE processor compared to a single core of CPU. Later, in 2009 we moved the algorithm to GPU with 170x speedup [11, 12]. Testing the algorithm on real data leads us to understanding that storing data in single precision data types is not suited for transoceanic modelling. The surface of the ocean became unstable due to insufficient accuracy of float data type. The ratio of the wave height to the ocean depth is on the edge of the float point precision. Thus we developed the newer version of the MOST software package and applied our knowledge and experience to adopt it for the GPU use. We have achieved a stable ocean surface and our results have become the same as the NOAA PMEL modelling results. Computation accuracy is less than 10^{-3} cm.

The achieved integrated performance results are summarized in Table 1.

In terms of wave propagation modelling the achieved performance means that computational time for the entire Pacific water area (4’ mash computational grid) decreases from 7 hours for the original program to just 3 minutes for Tesla K40 GPU, see [13].

Table 1 Performance comparison at different platforms. Time required for one iteration.

No	Platform	Time (before optimization)	Time (after optimization)
1	Original code	3000 ms	3000 ms
2	AMD (6 cores)	1800 ms	420 ms
3	Intel (8 cores)	300 ms	180 ms
4	CELL BE (PS3)	5000 ms	60 ms
5	GPU Tesla C1060	530 ms	19 ms
6	GPU Tesla K40	–	19 ms (double precision)

Table 2 Performance achieved at FPGA platforms

No	Characteristic	xc6vsx315-1 (SLEDv7)	xc7vx690-2 (VC709)
1	Number of Processor Elements	2	8
2	Frequency (MHz)	200	250 ms
3	Time of one iteration	19 ms	3.8 ms

Even better results were achieved at FPGA platform, see [6, 14]. The numbers are given in Table 2.

Obtained speedup allows us to develop more accurate and complex algorithms, e.g., we investigate an approach where nested finer grids are used for modelling closer to shore line and inside bays and gulfs.

3 Examples of Solving Problems Using FPGA-Based Hardware and Software Systems

FPGA-based solutions are currently used in broadcast automation, HPC, resource-consuming modelling, neural networks construction, autonomous devices for data processing in the field conditions, and other. This is due to a high range of FPGA with respect to logic capacity, price, peripheral devices, and existence of special series for space and military purposes (extended work limits for temperature, radiation, etc.).

Modern FPGA provides a good basis to construct wide enough pipelines for data processing. The so-called register silicon memory is compared to hundreds of MB; therefore, data could be effectively processed at the mode of acquisition. Below several examples in the area of image processing are presented.

3.1 Searching for Small Objects on a Series of Images

The task is to detect an object of 10 pixels or less from the sequence of images of the Earth surface, obtained from the geostationary satellite. First of all, background compensation algorithms are used to filter false objects, arising either because of a landscape features or due to a poor weather conditions. Image processing for the problem is divided into two within one frame and one inter-frame loads.

Data processing includes the following stages: differentiation (in fact, the difference between two sequential frames), integer snap frames, background compensation, and subpixel snap of image fragments. Block schemes of the proposed pipelines are given in Figures 1 and 2.

Fig. 1 Pipeline for the integer snap frames

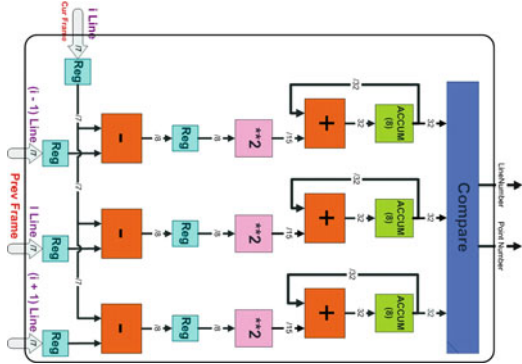


Fig. 2 Stage of subpixel snap of image fragments

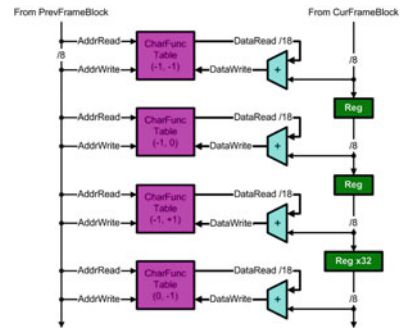


Table 3 Performance for the 1536 × 6200 frame processing

No	Platform	Time for one frame processing, sec.	Performance (million operations per second)
1	ADSP 21060	50,2	120
2	2x NM6403	28,8	210
3	Xilinx VirtexExc600e	1	6024

Implementing the integer snap algorithms, the 1 pixel horizontal frame moves were observed. So, it is required to have 32 × 32 fragment from the current frame and 34 × 32 fragment from the previous frame. Background compensation algorithm also requires shifts at all directions. Therefore, the 34 × 34 fragment from the previous frame is needed. Development of a joint pipeline for two algorithms above makes it possible to reduce two times the input data flow and, therefore, to achieve better overall performance.

The software application, which includes but not limited to the pipelines above, has been tested at the hardware complex ADP6203PCI (“Instrumental Systems”), based on Xilinx FPGA of Spartan series. The achieved performance, presented in Table 3, was compared to implementations of the same algorithms at the signal processor ADSP 21060 and two vector processors NM6403.

Fig. 3 Searching for $X \times Y$ object at $A \times B$ image.



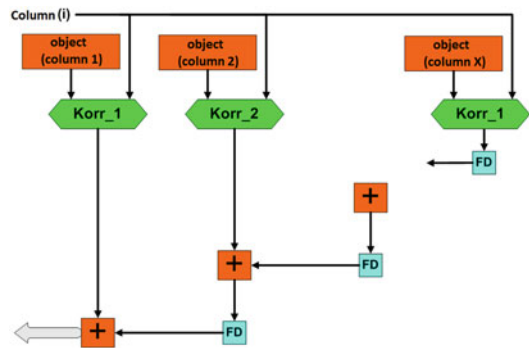
Fig. 4 Searching for $X \times Y$ object at $A \times Y$ image



Fig. 5 Pipeline to compute the mean square deviation



Fig. 6 Pipeline to search of an object in a row



3.2 Searching Object on the Image

There are $(A - X) \times (B - Y)$ possible positions of the $X \times Y$ object within the $A \times B$ image, see Figure 3.

In case the classic square mean difference deviation is used to identify an object, the total number of arithmetic operations for such search is $2 \times A \times B \times X \times Y$. In particular, looking for a 32×32 object at 2000×2000 image requires about 4×10^9 operations. Modern general purpose processor can handle it approximately in 1 sec.

However, the direct algorithm generates the auxiliary data flow of nearly 3, 8 GB/sec ($3, 8 \times 10^6$ possible positions). This fact limits performance, which is about 10 sec when using Core2Duo 2.4 GHz computer with DDR2-800 memory. The corresponding pipeline was developed for FPGA. Because of high parallelism, performance of search of $X \times Y$ mask in $A \times Y$ line is at the temp of data input as it takes one computer clock to compare the entire column, see Figure 4.

The developed block scheme of a pipeline to compute the mean square deviation is given in Figure 5.

Scheme of a pipeline, developed for the search of an object in a row, is given in Figure 6.

Just the columns serve as an input data for this pipeline, and then all are stored in the microchip inner memory.

A combination of above algorithms executes the object search with the following performance: to look for a 32×32 object in 2000×32 row, it takes just 2000 computer clocks (plus a certain pipeline lag). If we are looking for the same size object on the entire 2000×2000 image, we need to run the pipeline 1968 times. Hence, one needs about $3,8 \times 10^6$ computer clocks to complete that task, or, in case of 200 MHz processor frequency, it takes 20 msec. In the other words, it is possible to process the flow of 1,5 GB/sec, resulting, for example, at video recording with $2000 \times 2000 \times 8b$ resolution at 50 frame rate per second, see [15].

3.3 Motif Search in DNA Sequence

A lot of information about the DNA sequences is now available. Among the typical studies is the so-called motif search in these sequences. Mathematically the problem is nothing but the search of subsequence in the entire sequence (row). However, genetic code admits different nucleotides at the same position in the source sequence. So, according to IUPAC, it is necessary to use 15-symbols code for 4 symbols (A, T, G, C) set of nucleotides, see Figure 7.

FPGA architecture admits the lookup table (LUT) logic function with 6 inputs to compare a symbol in 4-digit alphabet with a symbol (character) of a 15-digit alphabet. So, within a pipeline, it is possible to compare 8-character words from the different alphabets above in one computer clock, see Figure 8.

Using FPGA facilities for parallel processing, it is possible to simultaneously compare all possible positions of an 8-character word in a 64-character row (57 positions) with the given 8-character motif, see Figure 9. Hence, we determine the presence of a given motif (8 characters) in a 64-symbol row in one computer clock.

This approach accelerates the code execution (compared to typical Core2Duo processor) up to 20000 times with FPGA XC5VLX330T, see Table 4.

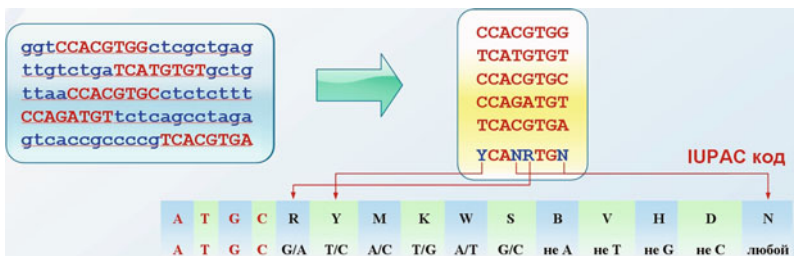


Fig. 7 15 symbols UIPAC code and 4 symbols nucleotides sequence

Fig. 8 Pipeline to compare words from different alphabets

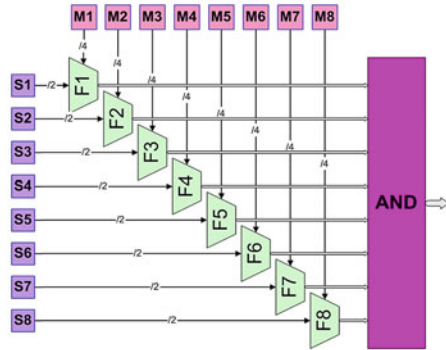


Fig. 9 One computer clock search of an 8-digit motif in a 64-digit row

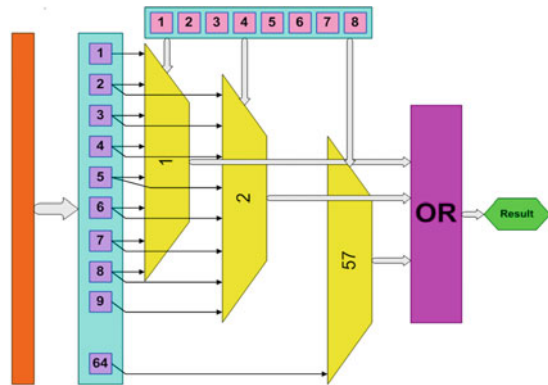


Table 4 Performance for the motif search problem

No	FPGA type	1.1.4. Processing speed (Motifs per 1 clock)	Processing time of 2, 6 × 10 ⁹ motifs (minutes)	Quantity of comparisons / clock
1	XC5VLX50T	8	22	3 650
2	XC5VLX110T	36	5	16 500
3	XC5VLX330T	170	1	77 500

4 Conclusion

As was shown, for a number of practical problems the HPC power is available by using GPU and FPGA. Many tasks from geophysics and bioinformatics are very well parallelizable and hence the corresponding software could be perfectly scalable. The authors achieve 350 times performance gain at 8 GPUs compared to a single CPU. In case of tsunami wave simulation the required CPU time has been reduced from hours to minutes or even seconds which will make it possible to avoid casualties in the future. By using the proper FPGA microchip the motif search in 65000 DNA sequence takes only 2 minutes. It is 20000 times faster compared to PC.

References

1. Candes, E. J., Donoho, D. L.: New Tight Frames of Curvelets and Optimal Representations of Objects with Piecewise-C2 Singularities. *Comm. Pure Appl. Math.* **57**, 219–266 (2002).
2. What's the Difference Between a CPU and a GPU? <https://blogs.nvidia.com/blog/2009/12/16/whats-the-difference-between-a-cpu-and-a-gpu/>. Cited 14 Aug 2017.
3. Duchkov, A. A., Andersson, F. A., Hoop, M. V.: Discrete Almost-Symmetric Wave Packets and Multiscale Geometrical Representation of (Seismic) Waves. *IEEE Transactions on Geoscience and Remote Sensing*. **48**, No. 9, 3408–3423 (2010).
4. Flynn's taxonomy https://en.wikipedia.org/wiki/Flynn%27s_taxonomy. Cited 14 Aug 2017.
5. Forecast Inundation Models <http://nctr.pmel.noaa.gov/sim.html>. Cited 23 Aug 2017.
6. Goryunov, E., Romanenko, A., Lavrentiev, M., Lysakov, K.: Modern simulation tools for real time numerical simulation of ocean-related processes. OCEANS 2015 - MTS/IEEE Washington 7404385 (2016)
7. GPU Technology conference <http://www.gputechconf.com/>. Cited 14 Aug 2017.
8. Hennenfent, G., Herrmann, F.: Seismic Denoising with Non-Uniformly Sampled Curvelets. *Computing in Science and Engineering*. **8** (3), 16–25 (2006).
9. High-level synthesis. https://en.wikipedia.org/wiki/High-level_synthesis. Cited 14 Aug 2017.
10. Hundreds of applications accelerated. <http://www.nvidia.com/object/gpu-applications.html>. Cited 14 Aug 2017.
11. Lavrentiev, M.M., Romanenko, A.A.: Modern Hardware Solutions to Speed Up Tsunami Simulation Codes. *Geophysical research abstracts*, **12**, EGU2010-3835 (2010)
12. Lavrentiev, M., Romanenko, A.: Modern Hardware to Simulate Tsunami Wave Propagation. *Proc. Automation, Control, and Information Technology (ACIT 2010)*, 151–157 (2010)
13. Lavrentiev, M., Romanenko, A.: Tsunami Wave Parameters Calculation before the Wave Approaches Coastal Line. *Proceedings of the Twenty-fourth (2014) International Ocean and Polar Engineering Conference, Busan, Korea, June 15-20, 2014*, **3**, 96–102 (2014)
14. Lavrentiev, M., Romanenko, A., Lysakov, K.: Modern Computer Architecture to Speed-Up Calculation of Tsunami Wave Propagation. *Proceedings of the Eleventh (2014) Pacific/Asia Offshore Mechanics Symposium, Shanghai, China, October 12–16*, 186–191 (2014)
15. Lysakov, K., Shadrin M.: FPGA Based Hardware Accelerator for High Performance Data-Stream Processing. *Pattern Recognition and Image Analysis*, **23**, No. 1, 26–34 (2013)
16. Naghizadeh, M., Sacchi, M. D.: Beyond Alias Hierarchical Scale Curvelet Interpolation of Regularly and Irregularly Sampled Seismic Data. *Geophysics*. **75**, 189–202 (2010).
17. Neelamani, R., Baumstein, A. I., Gillard, D. G., Hadidi, M. T., Soroka, W. L.: Coherent and Random Noise Attenuation Using the Curvelet Transform. *The Leading Edge*. **27**, No. 2, 240–246 (2008).
18. Nikitin, V.V., Romanenko, A.A., Duchkov, A.A., Andersson, F.: Parallel implementation of 3D-wave package decomposition on GPU and its application in geophysics. *Vestnik of NSU. IT series*. **11**, No. 1, 93–104 (2013) (in Russian).
19. NVIDIA Tesla V100 GPU architecture. <http://images.nvidia.com/content/volta-architecture/pdf/Volta-Architecture-Whitepaper-v1.0.pdf>. Cited 14 Aug 2017.
20. Piz Daint computer system. <https://www.top500.org/system/177824>. Cited 14 Aug 2017.
21. NVIDIA Tesla V100 GPU architecture <http://images.nvidia.com/content/volta-architecture/pdf/Volta-Architecture-Whitepaper-v1.0.pdf>. Cited 14 Aug 2017.
22. Titov, V.V., Synolakis, C.E.: Numerical modeling of tidal wave runup. *Journal of Waterway, Port, Coastal and Ocean Engineering*. **124**, No 4, 157–171 (1998).
23. Trinity computer system. <https://www.top500.org/system/178610>. Cited 14 Aug 2017.
24. Zyatkov, N., Ayzenberg, A., Aizenberg, A.M., Romanenko, A.: Highly-optimized TWSM Algorithm for Modeling Cascade Diffraction in Terms of Propagation-absorption Matrices. *Extended Abstracts, 75-th Conference and Exhibition, European Association of Geoscientists & Engineers, London, England, 10–13 June 2013, Th-P02–11* (2013)

Parallel Algorithms for Low Rank Tensor Arithmetic



Lars Grasedyck and Christian Löbber

1 Introduction

A problem is said to be parameter-dependent if there exist problem parameters $p_1, \dots, p_d, d \in \mathbb{N}$, such that the solution $u(p_1, \dots, p_d)$ of the problem depends on the choice of the parameters. With no loss of generality we may assume $p_\mu \in P_\mu$ with $P_\mu \subset \mathbb{N}, \mu = 1, \dots, d$. Furthermore we consider finite parameter domains P_μ and also the solution itself shall be a finite vector: $u(p_1, \dots, p_d) \in \mathbb{R}^n$. In short, we can solve the problem for any parameter combination $(p_1, \dots, p_d) \in P_1 \times \dots \times P_d$, which results in a solution vector $u(p_1, \dots, p_d) \in \mathbb{R}^n$ depending on the parameters we chose.

All vector entries for each possible parameter combination can be written as

$$u(p_1, \dots, p_d, i) \in \mathbb{R}, \quad p_\mu \in P_\mu, \quad i = 1, \dots, n,$$

which we regard as a $(d + 1)$ -dimensional tensor $u \in \mathbb{R}^{P_1 \times \dots \times P_d \times \{1, \dots, n\}}$ with d dimensions for the d parameters plus one *spatial dimension*.

For ease of presentation, we will for now assume the P_μ to be all of the same size m , which results in a $(d + 1)$ -dimensional solution tensor $u \in \mathbb{R}^{m \times \dots \times m \times n}$.

Thinking of sophisticated problems, e.g., discretized PDEs, already the computation of one solution $u(p_1, \dots, p_d)$, for one single parameter combination, may pose a challenging task. Since the number of parameter combinations is m^d , which grows exponentially with the number d of parameters, it becomes impractical to compute the whole solution tensor u in a naive way (by just running through all parameter

L. Grasedyck (✉) · C. Löbber

Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, 52056 Aachen, Germany

e-mail: lg@igpm.rwth-aachen.de; loebbert@igpm.rwth-aachen.de

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41, https://doi.org/10.1007/978-3-030-02487-1_16

271

combinations) already for moderate numbers m and d . We therefore assume the solution tensor u to be of small *tensor rank*, which basically means that u can be represented (or approximated) by a sum of separable tensors (*tensors of rank 1*):

$$u(p_1, \dots, p_d, i) = \sum_{k=1}^r u_{k,1}(p_1) \cdots u_{k,d}(p_d) u_{k,d+1}(i). \quad (1)$$

A compact form of (1) is $u = \sum_{k=1}^r u_{k,1} \otimes \cdots \otimes u_{k,d} \otimes u_{k,d+1}$.

In practice, it turns out that for many parameter-dependent problems the solution tensor allows for good approximations by low rank tensors.

Considering a parameter-dependent problem which can be expressed by a linear system

$$A(p_1, \dots, p_d) \cdot u(p_1, \dots, p_d) = b(p_1, \dots, p_d), \quad (2)$$

our approach is to represent (or approximate) the matrix A and the right-hand side b as tensors in the hierarchical Tucker format (short HT format). Then we can use (parallel) algorithms to perform arithmetic like matrix-vector products, scalar products, sums, etc. for tensors in the HT format (cf. Section 3). For this we also refer to [3]. With that it is possible to apply iterative methods like *conjugate gradients* or *multigrid* directly in the HT format which we did in [5].

In this article we present first results on using hybrid parallelization: In addition to the parallelization between the distributed nodes of the HT tensor, we use shared memory parallelization to accelerate the computations on each node.

In Section 2 we give an insight into the HT format, which is the low rank tensor format we are using for all our tensor approximations.

In Section 3 we give an overview of the parallel algorithms we developed in order to perform arithmetic operations for tensors in the HT format. These operations include the application of a matrix to a tensor, both stored in the HT format, possibly distributed over several compute nodes.

In Section 4 we finally present some runtime tests, including first tests for the hybrid parallelization.

2 The Hierarchical Tucker format

In this section we explain the hierarchical Tucker format (HT format) which we use to represent tensors of low rank. Notice that we now denote by d the number of *all* tensor dimensions, whereas in Section 1 the letter d stands for the number of parameters of the underlying problem, the solution of which has an additional (space) dimension, which yields tensor dimension $d + 1$.

Let $\mathcal{I}_1, \dots, \mathcal{I}_d \subset \mathbb{N}$ be finite index sets, then a vector $A \in \mathbb{R}^{\mathcal{I}_1 \times \cdots \times \mathcal{I}_d}$ over the product index set $\mathcal{I}_1 \times \cdots \times \mathcal{I}_d$ is called tensor of dimension d .

For $d = 1$ these are just vectors in $\mathbb{R}^{\mathcal{I}_1}$. For $d = 2$ we have matrices in $\mathbb{R}^{\mathcal{I}_1 \times \mathcal{I}_2}$. For $d = 3$ one may think of matrices stapled on each other along one of the three dimensions.

In practice one often needs an order on the product index set $\mathcal{I}_1 \times \dots \times \mathcal{I}_d$, which can, e.g., be the lexicographical order.

The rank of a tensor can be defined as the smallest possible number r in (1), i.e., the smallest number r such that we find a representation

$$A(i_1, \dots, i_d) = \sum_{k=1}^r A_1(i_1, k) \cdots A_d(i_d, k), \quad \text{for } (i_1, \dots, i_d) \in \mathcal{I}_1 \times \dots \times \mathcal{I}_d. \quad (3)$$

For matrices $M \in \mathbb{R}^{\mathcal{I}_1 \times \mathcal{I}_2}$ this rank definition coincides with the well-known matrix rank: If M is of rank r , we find a representation

$$M(i_1, i_2) = \sum_{k=1}^r C(i_1, k) \cdot D(i_2, k), \quad \text{i.e. } M = C \cdot D^\top, \quad (4)$$

with¹ $C \in \mathbb{R}^{\mathcal{I}_1 \times r}$, $D \in \mathbb{R}^{\mathcal{I}_2 \times r}$, and there is no such representation with a smaller number than r . The factors C and D of (4) can, e.g., be computed by a reduced singular value decomposition (rSVD) $M = U \Sigma V^\top$, where *reduced* means that we leave out all zero singular values as well as the corresponding columns of U and V , i.e., $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$, $\sigma_1 \geq \dots \geq \sigma_r > 0$ and $U \in \mathbb{R}^{\mathcal{I}_1 \times r}$, $V \in \mathbb{R}^{\mathcal{I}_2 \times r}$ have orthonormal columns.

From (4) one sees that $\text{range}(C) = \text{range}(M)$ and $\text{range}(D) = \text{range}(M^\top)$, where $\text{range}(M) = \{Mx : x \in \mathbb{R}^{\mathcal{I}_2}\} \subset \mathbb{R}^{\mathcal{I}_1}$. Note that M^\top is just another representation (we will call it *matricization*) of the tensor $M \in \mathbb{R}^{\mathcal{I}_1 \times \mathcal{I}_2}$: Normally we choose \mathcal{I}_1 to act as row index set, whereas in M^\top we have \mathcal{I}_2 as row index set. Further notice that if the rank r is small ($r \ll \#\mathcal{I}_1, \#\mathcal{I}_2$), storing the factors $C \in \mathbb{R}^{\mathcal{I}_1 \times r}$ and $D \in \mathbb{R}^{\mathcal{I}_2 \times r}$ instead of $M \in \mathbb{R}^{\mathcal{I}_1 \times \mathcal{I}_2}$ yields a massive storage reduction. In the HT format these observations are used to store a tensor $A \in \mathbb{R}^{\mathcal{I}_1 \times \dots \times \mathcal{I}_d}$ of small rank: We split the set $D := \{1, \dots, d\}$ of all tensor dimensions in two non-empty disjoint subsets $t \subset D$ and $[t] := D \setminus t$ with $t, [t] \neq \emptyset$. The corresponding product index sets are denoted by \mathcal{I}_t and $\mathcal{I}_{[t]}$:

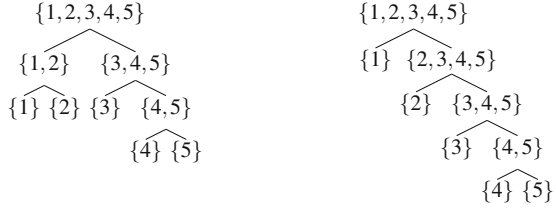
$$\mathcal{I}_t := \times_{\mu \in t} \mathcal{I}_\mu, \quad \mathcal{I}_{[t]} := \times_{\mu \in [t]} \mathcal{I}_\mu.$$

By choosing \mathcal{I}_t as row index set, we can represent the tensor A as a matrix $\mathcal{M}_t(A)$, which we call the *matricization* of A with respect to the set t :

$$\mathcal{M}_t(A)(\mathbf{i}_t, \mathbf{i}_{[t]}) = A(i_1, \dots, i_d), \quad \mathbf{i}_t \in \mathcal{I}_t, \quad \mathbf{i}_{[t]} \in \mathcal{I}_{[t]},$$

¹We abbreviate $\mathbb{R}^{\mathcal{I}_1 \times \{1, \dots, r\}}$ by $\mathbb{R}^{\mathcal{I}_1 \times r}$.

Fig. 1 Two different trees defining different HT format structures for tensors of dimension $d = 5$.



where we use the abbreviation $\mathbf{i}_t = (i_\mu)_{\mu \in t}$ for multi-indices in \mathcal{I}_t . By a rSVD $\mathcal{M}_t(A) = U_t \Sigma V_t^\top$ we get $U_t \in \mathbb{R}^{\mathcal{I}_t \times r_t}$ with $\text{range}(U_t) = \text{range}(\mathcal{M}_t(A))$, where r_t is the rank of $\mathcal{M}_t(A)$.

For the matricization of A with respect to $[t]$ we have $\mathcal{M}_{[t]}(A) = \mathcal{M}_t(A)^\top$ (now $\mathcal{I}_{[t]}$ is the row index set) with rSVD $\mathcal{M}_{[t]}(A) = V_t \Sigma U_t^\top$, i.e., $U_{[t]} := V_t \in \mathbb{R}^{\mathcal{I}_{[t]} \times r_t}$ with $\text{range}(U_{[t]}) = \text{range}(\mathcal{M}_{[t]}(A))$. By multiplying Σ either to U_t or to $U_{[t]} = V_t$, we see that it is sufficient to only store U_t and $U_{[t]}$, which can already save vast amounts of storage if the rank r_t is small ($r_t \ll \#\mathcal{I}_t, \#\mathcal{I}_{[t]}$).

Nevertheless, the index sets \mathcal{I}_t and $\mathcal{I}_{[t]}$ may still be large, possibly even too large to store U_t and $U_{[t]}$. The idea of the HT format is to repeat the above splitting of the tensor dimensions and the according factorization recursively: For each $t \subset D$ which we obtain as part of a split set of tensor dimensions and which is not a singleton we choose $s_1, s_2 \neq \emptyset$ such that

$$t = s_1 \cup s_2, \quad s_1 \cap s_2 = \emptyset, \tag{5}$$

and construct matrices $U_{s_1} \in \mathbb{R}^{\mathcal{I}_{s_1} \times r_{s_1}}, U_{s_2} \in \mathbb{R}^{\mathcal{I}_{s_2} \times r_{s_2}}$ with $\text{range}(U_{s_1}) = \text{range}(\mathcal{M}_{s_1}(A))$ and $\text{range}(U_{s_2}) = \text{range}(\mathcal{M}_{s_2}(A))$, e.g., by a rSVD. We continue like this for s_1 and s_2 , which leads to a binary tree T with singletons as leaves and $\text{root}(T) = \{1, \dots, d\}$ (cf. Figure 1). We demand that each non-singleton $t \in T$ has *exactly two* sons s_1, s_2 , in order to be able to use the rSVD or some other suitable matrix factorization.²

It is not hard to verify that the following *nestedness property* always holds true:

$$U_t(-, k) \in \text{span} \{ U_{s_1}(-, k_1) \otimes U_{s_2}(-, k_2) : 1 \leq k_1 \leq r_{s_1}, 1 \leq k_2 \leq r_{s_2} \}, \tag{6}$$

²A rSVD exists for any matrix which is not the zero matrix. This is, in general, not the case for tensors of higher dimension $d > 2$. Nevertheless, if \mathcal{I}_t is large, a rSVD of $\mathcal{M}_t(A)$ may be no more computable. This is, however, not a handicap for us when we have available HT representations of a matrix A and a right-hand side B and want to solve $AX = B$ by some iterative method inside the HT format. We can then choose a starting vector X_0 in the HT format (e.g., $X_0 := B$) and we will never have to transfer a tensor into the HT format.

If we need to approximate large tensors in the HT format, we may use other approximation techniques as, e.g., the cross approximation for HT tensors [2].

where $U_t(-, k)$ stands for the k -th column of U_t and $\text{sons}(t) = \{s_1, s_2\}$. By this each column of U_t is a linear combination of all Kronecker products of the sons' columns, which can be written as

$$U_t(\mathbf{i}_t, k) = \sum_{k_1=1}^{r_{s_1}} \sum_{k_2=1}^{r_{s_2}} B_t(k, k_1, k_2) \cdot U_{s_1}(\mathbf{i}_{s_1}, k_1) \cdot U_{s_2}(\mathbf{i}_{s_2}, k_2) \tag{7}$$

for all $\mathbf{i}_t = (i_\mu)_{\mu \in t} \in \mathcal{I}_t$ and the according subindices $\mathbf{i}_{s_1} \in \mathcal{I}_{s_1}$ and $\mathbf{i}_{s_2} \in \mathcal{I}_{s_2}$.

The 3-dimensional coefficients B_t are called *transfer tensors*: With B_t we can reconstruct U_t from U_{s_1} and U_{s_2} , where $\text{sons}(t) = \{s_1, s_2\}$.

For $\text{root}(T) = D$ the matricization $\mathcal{M}_D(A)$ means taking $\mathcal{I}_D = \mathcal{I}_1 \times \dots \times \mathcal{I}_d$ as row index set, i.e., $\mathcal{M}_D(A)$ is one column vector containing all entries of A . To be consistent with (7) we can define $U_D := \mathcal{M}_D(A)$ and get

$$A(\mathbf{i}_D) = U_D(\mathbf{i}_D) = \sum_{k_1=1}^{r_1} \sum_{k_2=1}^{r_2} B_D(k_1, k_2) U_t(\mathbf{i}_t, k_1) \cdot U_{[t]}(\mathbf{i}_{[t]}, k_2) \tag{8}$$

for all $\mathbf{i}_D = (i_1, \dots, i_d) \in \mathcal{I}_1 \times \dots \times \mathcal{I}_d$. Note that the root transfer tensor B_D is a matrix, since U_D only consists of one column.

From (8) and (7) it can be seen that it is sufficient to store the U_t for all leaves $t = \{\mu\}$, $\mu = 1, \dots, d$, of the tree T and to store the transfer tensors $B_t \in \mathbb{R}^{r_t \times r_{s_1} \times r_{s_2}}$ for each non-leaf node $t \in T$, $\text{sons}(T) = \{s_1, s_2\}$. Then each tensor entry $A(i_1, \dots, i_d)$ can be computed recursively by (8) and (7).

Definition 1 (Hierarchical Tucker Format, Hierarchical Tucker Rank)

Let $A \in \mathbb{R}^{\mathcal{I}_1 \times \dots \times \mathcal{I}_d}$ be a tensor of dimension d and T a binary tree with $\text{root}(T) = \{1, \dots, d\}$, such that each $t \in T$ which is not a singleton has exactly two successors $s_1, s_2 \neq \emptyset$ fulfilling (5). By $\mathcal{L}(T)$ we denote the set of all leaves in T .

Then $((B_t)_{t \in T \setminus \mathcal{L}(T)}, (U_t)_{t \in \mathcal{L}(T)})$ fulfilling (8) and (7) is called representation of A in the *hierarchical Tucker format (HT format)*.

The *hierarchical Tucker ranks (HT ranks)* of A with respect to the tree T are the minimal numbers $(r_t)_{t \in T}$ such that there exists a representation $((B_t)_{t \in T \setminus \mathcal{L}(T)}, (U_t)_{t \in \mathcal{L}(T)})$ of A in the HT format.

Remark 1 For the root node $D = \{1, \dots, d\}$ we always have $r_D = 1$.

Remark 2 The HT ranks were defined as the smallest possible numbers $(r_t)_{t \in T}$ in (7) and (8). For any HT representation (not necessarily minimal) we call the numbers $(r_t)_{t \in T}$ the *HT representation ranks*. The HT representation ranks are therefore always bounded below by the HT ranks.³

³We define $(r_t)_{t \in T} \leq (s_t)_{t \in T} :\Leftrightarrow r_t \leq s_t$ for all $t \in T$.

Remark 3 By Definition 1 the columns of $U_t, t \in T$, need to be neither orthogonal nor normalized nor linear independent.

Remark 4 Instead of $U_{\{\mu\}}$ and $r_{\{\mu\}}$ we use the shorter notations U_μ and r_μ for the leaves $\{\mu\} \in \mathcal{L}(T)$.

2.1 Storage Complexity of the HT Format

Let $A \in \mathbb{R}^{\mathcal{I}_1 \times \dots \times \mathcal{I}_d}$ be a tensor with HT representation $((B_t)_{t \in T \setminus \mathcal{L}(T)}, (U_t)_{t \in \mathcal{L}(T)})$ for a suitable tree T . For simplicity let all index sets $\mathcal{I}_\mu, \mu = 1, \dots, d$, be of the same size $n := \#\mathcal{I}_\mu$ and let $r := \max\{r_t : t \in T\}$ be the maximum of the HT ranks.

Then the tensor A consists of n^d entries, whereas in the HT format we would store not more than

$$(\#\mathcal{L}(T)) \cdot n \cdot r + \#(\text{inner nodes}) \cdot r^3 + r^2 = dnr + (d - 2)r^3 + r^2 \quad (9)$$

entries, which is a number *linear* in the tensor dimension d .

2.2 The Choice of the Tree

In principle the choice of the underlying tree T is a non-trivial issue [1]: It can be the case that a tensor has smaller HT ranks with respect to one tree T_1 than for another tree T_2 . This question is, however, not part of this article.

We always choose a tree which splits each non-singleton $t \in T$ into sons s_1, s_2 , either of the same cardinality (if possible) or with $\#s_2 = \#s_1 + 1$ (cf. left tree in Figure 1). This tree construction minimizes the number of tree levels, which will be advantageous for our parallel algorithms (see Section 3) which scale linear with the number of levels of the underlying tree T .

2.3 Other Tensor Formats

The *canonical* tensor rank of a tensor $A \in \mathbb{R}^{\mathcal{I}_1 \times \dots \times \mathcal{I}_d}$ is defined as the minimal number r for which we can find a decomposition like (3). It is not difficult to show that A possessing canonical rank r always yields $r_t \leq r, t \in T$, for the HT ranks of A with respect to any suitable tree T . Note, however, that we have an r^3 dependence in (9) which we do not have for the storage of A_1, \dots, A_d from (3). One huge advantage of the HT format is that for prescribed $(r_t)_{t \in T}$ the set of all tensors with HT ranks bounded by $(r_t)_{t \in T}$ is closed. This is not the case for the set of all tensors with canonical rank bounded by a prescribed number r (for tensor dimensions $d > 2$).

There are a number of other tensor formats for which we refer to [6]. For $d = 2$ all tensor rank models coincide with the matrix rank.

In the literature we found several approaches on the parallelization of algorithms for different tensor formats, e.g., a parallel ALS algorithm for the HT format in [3], parallel tensor completion for the CP format in [7], parallel tensor compression for the Tucker format in [9] and the parallel computation of contractions for distributed tensor networks in [8].

3 Parallel Arithmetic in the HT Format

For huge tensors it may be unavoidable to have the tensor data stored distributed over distinct compute nodes. We developed algorithms which perform arithmetic operations on tensors in the HT format, where the tensors may be stored in a distributed way. In [3] we encountered similar work for an ALS algorithm in the HT format. Let us suppose that the tensor data for each $t \in T$ of the underlying tree is stored on its own compute node and, for simplicity, that the tensor dimension is a power of two: $d = 2^\ell$ with $\ell = 1, 2, 3, \dots$. Then our algorithms run in parallel on all nodes belonging to the same level of the tree, which is why we want to minimize the number of tree levels by choosing a balanced tree. This means that the sons s_1 and s_2 of $t \in T$ are of equal or nearly equal size (cf. left tree in Figure 1). For a balanced tree the number of tree levels equals $\log_2(d) + 1$, where d is the tensor dimension. For fully distributed tensors we can thus expect a parallel runtime of our algorithms which grows logarithmically with the tensor dimension d (cf. Section 4).

So far we implemented algorithms which compute sums, inner products, and Hadamard products⁴ of HT tensors. Also the matrix-vector multiplication can directly be carried out in the HT format, where both, the matrix and the vector, are stored in the HT format. Furthermore we have available a parallel algorithm for the *truncation* of an HT tensor down to lower HT ranks (cf. [4]). The truncation is essential for the computation of sums, Hadamard products and the matrix-vector multiplication since these operations typically increase the HT ranks. Finally also the evaluation of single tensor entries has been implemented to run in parallel on all nodes of the same tree level.

In this section we illustrate the evaluation of tensor entries and the matrix-vector multiplication in the HT format. For further details on (parallel) HT arithmetic we refer to [6] or [3, 5].

⁴The Hadamard product $x \circ y$ of two vectors $x, y \in \mathbb{R}^{\mathcal{I}}$ is the vector of the entry-wise products: $(x \circ y)(i) = x(i) \cdot y(i)$ for all $i \in \mathcal{I}$.

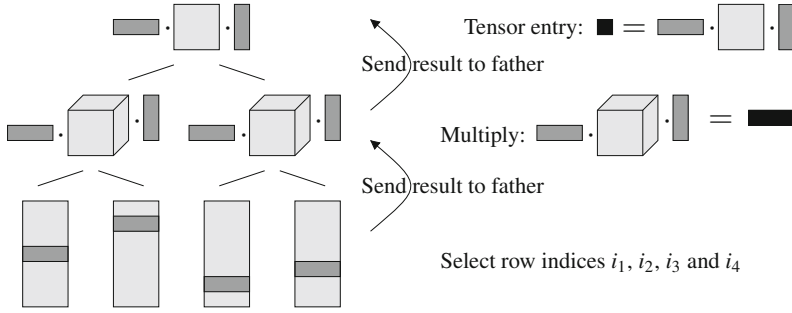


Fig. 2 Parallel evaluation of an HT tensor $A \in \mathbb{R}^{\mathcal{I}_1 \times \mathcal{I}_2 \times \mathcal{I}_3 \times \mathcal{I}_4}$ at index (i_1, i_2, i_3, i_4) .

3.1 The Evaluation of Tensor Entries

From (7) and (8) we see how the evaluation of an HT tensor $A \in \mathbb{R}^{\mathcal{I}_1 \times \dots \times \mathcal{I}_d}$ at $(i_1, \dots, i_d) \in \mathcal{I}_1 \times \dots \times \mathcal{I}_d$ works: On each leaf node $\{\mu\}$ we just choose the row $U_\mu(i_\mu, -)$ corresponding to the index (i_1, \dots, i_d) . This can of course be carried out in parallel for all leaves. Each leaf node then sends the selected row to its father, where the rows of both sons are multiplied to the transfer tensor B_t , according to (7) and (8). On the root node we obtain the tensor entry for the requested entry (cf. Figure 2). Apparently we only need communication between father and son nodes, i.e., the algorithm can run in parallel on nodes of the same tree level (remember: $d = 2^\ell$).

3.2 Matrix-Vector Multiplication in the HT Format

Since a d -dimensional tensor X is defined as a vector over a product index set of the form $\mathcal{I}_1 \times \dots \times \mathcal{I}_d$, we can define the matrix-vector multiplication $A \cdot X$ for a matrix $A \in \mathbb{R}^{(\mathcal{I}_1 \times \dots \times \mathcal{I}_d) \times (\mathcal{I}_1 \times \dots \times \mathcal{I}_d)}$. The resulting vector $A \cdot X$ is then an element of $\mathbb{R}^{\mathcal{I}_1 \times \dots \times \mathcal{I}_d}$. Rearranging the index set of A to $(\mathcal{I}_1 \times \mathcal{I}_1) \times \dots \times (\mathcal{I}_d \times \mathcal{I}_d)$, we can regard A itself as a d -dimensional tensor. Suppose that we have an HT representation of X for some underlying tree T . If we find an HT representation (or approximation) for the matrix A , based on the same tree T , we can compute the matrix-vector multiplication $A \cdot X$ directly in the HT format, which we sketched out in Figure 3. The computation of $A \cdot X$ in the HT format can be carried out in parallel on each node of the underlying tree T . The transfer tensors of the result $A \cdot X$ are just the Kronecker products of the respective transfer tensors of A and X . For the root node this is the well-known Kronecker product of matrices and for the inner nodes this is the obvious generalization to 3-dimensional tensors. Note that the columns in the leaves of A stand for matrices (stored as columns). Each leaf of $A \cdot X$ consists

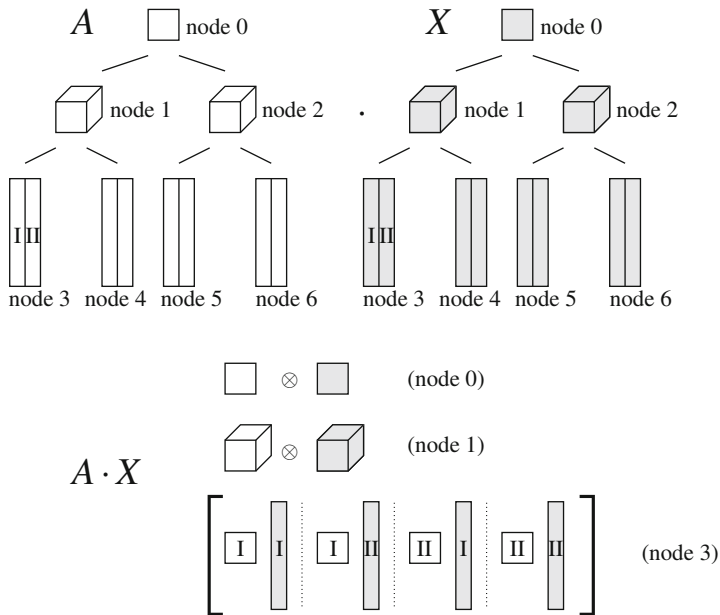


Fig. 3 Matrix-vector product for a matrix A (white) with a tensor X (grey), both stored in the HT format, based on the same underlying tree T . Below we show the root node, one inner node and one leaf node of the resulting HT tensor $A \cdot X$. Here we suppose the compute nodes of A to be the same as the corresponding compute nodes of X . If this was not the case we would have to communicate between the respective nodes.

of all columns which can be obtained as matrix-vector products of any column in the respective leaf of A with any column in the respective leaf of X . Altogether this results in an HT representation of $A \cdot X$ with HT representation ranks $r_t = r_t^{(1)} \cdot r_t^{(2)}$, $t \in T$, provided that A and X are represented as HT tensors of ranks $r_t^{(1)}$ and $r_t^{(2)}$.

Although the sole computation of $A \cdot X$ may run in parallel on all nodes of the underlying tree, this operation will typically be followed by a truncation down to lower HT ranks, which can run in parallel on all nodes of the same tree level. The parallel runtime of several matrix-vector products will therefore be determined by the parallel runtime of the truncations.

In many applications we encounter parameter-dependent matrices $A(p_1, \dots, p_d)$, which are of *affine type*, which means

$$A(p_1, \dots, p_d) = A_0 + \psi_1(p_1) \cdot A_1 + \dots + \psi_d(p_d) \cdot A_d, \quad (10)$$

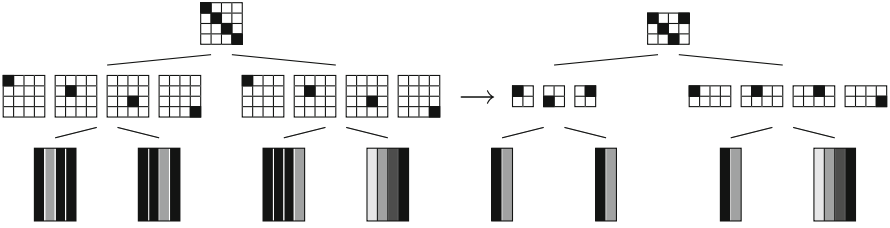


Fig. 4 HT representation of the matrix $A(p_1, p_2, p_3) = A_0 + \psi_1(p_1)A_1 + \psi_2(p_2)A_2 + \psi_3(p_3)A_3$, i.e., (10) for $d = 3$: The first three leaves correspond to the three parameter dimensions, the last leaf corresponds to the matrix dimension (row and column indices are combined to one index). For the sake of clarity, the layers of the 3-dimensional transfer tensors are displayed next to each other. Notice that each addend in (10) corresponds to a tensor of rank 1 and can therefore be represented in the HT format with all HT ranks equal to 1 (cf. paragraph 2.3), i.e., each transfer tensor can be chosen as the 1×1 matrix $(1) \in \mathbb{R}^{1 \times 1}$. In the HT format we can add tensors by collecting in each leaf the columns of all addends and for the transfer tensors we get block diagonal tensors with the transfer tensors of all addends as blocks. This results in the left HT representation for A : black parts stand for ones, white parts stand for zeros. The grey column in each of the first three leaves equals $(\psi_\mu(p_\mu))_{p_\mu \in P_\mu}$, $\mu = 1, 2, 3$. The columns of the last leaf are the matrices A_0, A_1, A_2, A_3 , written as columns. The left HT representation can obviously be reduced by taking the black column $(1, \dots, 1)^T$ only once per leaf, which results in the HT representation of A on the right.

where the ψ_μ , $\mu = 1, \dots, d$, are arbitrary functions and the matrices A_μ , $\mu = 0, \dots, d$, are parameter-independent. For this affine parameter dependence, the matrix A can directly be represented as a $(d + 1)$ -dimensional tensor in the HT format (see Figure 4) with HT ranks

$$r_t = \begin{cases} 1 & \text{if } t = \text{root}(T) = \{0, \dots, d\}, \\ d + 1 & \text{if } d \in t, \\ r_{s_1} + r_{s_2} - 1 & \text{if } d \notin t, \text{ sons}(t) = \{s_1, s_2\}, \\ 2 & \text{if } t = \{\mu\}, \mu \neq d, \end{cases}$$

with $t \neq \text{root}(T)$ in the last three cases.

4 Parallel Runtime Tests

We tested the parallel runtime of our algorithms for fully distributed tensors of dimensions $d = 4, 8, 16, 32, 64$, i.e., we used $2d - 1 = 7, 15, 31, 63, 127$ compute nodes. Our algorithms use MPI for the communication between the nodes.

Figure 5 shows the parallel runtimes for the inner product of two HT tensors and for the truncation of an HT tensor down to lower rank, which are nearly identical. The HT tensors were chosen by random with $\#\mathcal{J}_\mu = 10\,000$ for all $\mu = 1, \dots, d$ and with HT ranks $r_t = 100$ for all $t \in T \setminus \text{root}(T)$. One clearly sees the logarithmical dependence of the parallel runtime on the tensor dimension d .

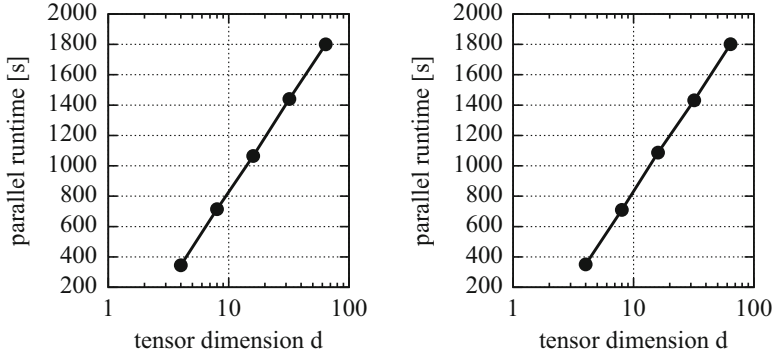
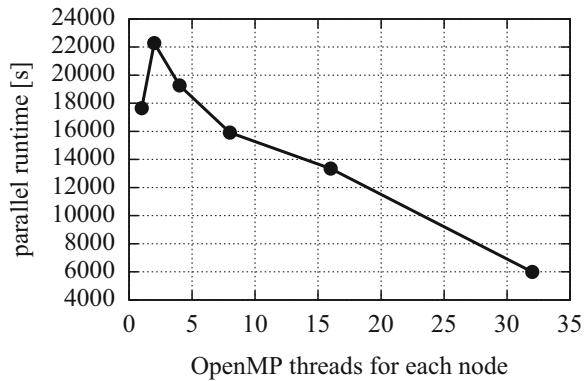


Fig. 5 Parallel runtime of the inner product (left) and the truncation (right) for HT tensors of dimension $d = 4, 8, 16, 32, 64$ with each tensor dimension of size 10 000 and HT ranks 100.

Fig. 6 Parallel runtime for the truncation of an HT tensor of dimension $d = 8$ with each tensor dimension of size 100 000 and HT ranks 200 using hybrid parallelization with MPI and OpenMP.



In addition to the MPI parallelization between the different HT nodes, we started using OpenMP for shared memory parallelization on each node. First results can be seen in Figure 6, which shows the parallel runtime for the truncation of an HT tensor of dimension $d = 8$ with $\#\mathcal{S}_\mu = 100\,000$ for all $\mu = 1, \dots, d$ and HT ranks $r_t = 200$ for all $t \in T \setminus \text{root}(T)$.

For an example of using our parallel algorithms for iterative methods in order to solve a linear equation $AX = B$ which stems from a diffusion problem with parameter-dependent diffusion coefficients, we refer to [5]. For that example the matrix A is of affine type (10), such that it can directly be represented in the HT format (cf. Section 3).

5 Conclusions

In [5] we concluded that our parallel algorithms for HT tensor arithmetic seem to be useful to solve parameter-dependent problems of affine type (10) by means of

iterative methods. In principle one could apply this approach to parameter-dependent linear problems with arbitrary matrix A , provided that we can approximate A in the HT format.

On the other hand, we can use our algorithms to compute statistical quantities of interest for a solution tensor X , as, e.g., mean values (with respect to any subset of the tensor dimensions) or the expected value and the variance with respect to a probability distribution of the parameters, given in the HT format. For these computations we typically need the inner product and for the variance as well the Hadamard product.

Furthermore, we may apply our algorithms to compute the relative residual norm $\|AX - B\|_2 / \|B\|_2$ for a solution tensor X , which might result from some non-deterministic sampling algorithm.

For tensors of HT rank r the inner product and the truncation are of complexity $\mathcal{O}(r^4)$, which can be rather large already for moderate ranks r . Therefore it would be desirable to have additional parallelization on each node, which we introduced in this article: By using OpenMP on each compute node, we reduced the parallel computing time of the truncation by a factor of $\approx 1/3$ for HT ranks of 200.

Acknowledgement The authors gratefully acknowledge the support by the DFG priority programme 1648 (SPPEXA) under grant GR-3179/4-2.

References

1. Ballani, J., Grasedyck, L.: Tree Adaptive Approximation in the Hierarchical Tensor Format. *SIAM Journal on Scientific Computing* **36**(4), A1415–A1431 (2014)
2. Ballani, J., Grasedyck, L., Kluge, M.: Black box approximation of tensors in hierarchical Tucker format. *Linear Algebra Appl.* **438**(2), 639–657 (2013)
3. Etter, S.: Parallel ALS Algorithm for Solving Linear Systems in the Hierarchical Tucker Representation. *SIAM Journal on Scientific Computing* **38**(4), A2585–A2609 (2016)
4. Grasedyck, L.: Hierarchical Singular Value Decomposition of Tensors. *SIAM J. Matrix Anal. Appl.* **31**, 2029–2054 (2010)
5. Grasedyck, L., Löbbert, C.: Distributed Hierarchical SVD in the Hierarchical Tucker Format. *arXiv* (2017). URL <http://arxiv.org/abs/1708.03340>
6. Hackbusch, W.: Tensor spaces and numerical tensor calculus, *Springer series in computational mathematics*, vol. 42. Springer, Heidelberg (2012)
7. Karlsson, L., Kressner, D., Uschmajew, A.: Parallel algorithms for tensor completion in the CP format. *Parallel Computing* **57**(Supplement C), 222–234 (2016)
8. Solomonik, E., Matthews, D., Hammond, J.R., Stanton, J.F., Demmel, J.: A massively parallel tensor contraction framework for coupled-cluster computations. *Journal of Parallel and Distributed Computing* **74**(12), 3176–3190 (2014). *Domain-Specific Languages and High-Level Frameworks for High-Performance Computing*
9. Woody Austin Grey Ballard, T.G.K.: Parallel Tensor Compression for Large-Scale Scientific Data. 2016 IEEE International Parallel and Distributed Processing Symposium (IPDPS) pp. 912–922 (2016)

The Resiliency of Multilevel Methods on Next-Generation Computing Platforms: Probabilistic Model and Its Analysis



Mark Ainsworth and Christian Glusa

1 Introduction

Exascale computing is anticipated to have a huge impact on computational simulation. However, as the number of components in a system becomes larger, the likelihood of one or more components failing or function abnormally during an application run increases. The problem is exacerbated by the decreasing physical size of basic components such as transistors, and the accompanying increased possibility of quantum tunneling corrupting logic states [6, 7].

Sandia National Laboratories is a multimission laboratory managed and operated by the National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the US Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND2018-0720 C.

This research was performed at the Brown University as part of Christian Glusa's dissertation [11].

M. Ainsworth (✉)

Division of Applied Mathematics, Brown University, 182 George St, Providence, RI 02912, USA
e-mail: mark_ainsworth@brown.edu

C. Glusa

Division of Applied Mathematics, Brown University, 182 George St, Providence, RI 02912, USA
Center for Computing Research, Sandia National Laboratories, Albuquerque, NM 87185, USA
e-mail: caglusa@sandia.gov

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_17

283

Current day petascale systems already exhibit a diverse range of faults that may occur during computation. These faults can arise from failures in the physical components of the system, or intermittent software faults that appear only in certain application states. One source of faults is cosmic radiation with charged particles, which can lead to memory bit-flips or incorrect behavior of logic units. Future HPC systems are expected to be built from even larger numbers of components than current systems, and the rate of faults in the system will increase accordingly. It is generally accepted that future large-scale systems must operate within a 20-MW power envelope. This will require the usage of lower voltage logic thresholds. Moreover, cost constraints will result in greater utilization of consumer grade components, with accompanying reduced reliability [7].

Roughly speaking, faults can be classified as follows [3]: *hard* or *stop-fail* faults are faults which would otherwise lead to an immediate program termination, unless treated on the system level. *Soft* faults are those leading to program or data corruption, and which might only result in an erroneous program termination after some delay.

Reported fault rates seem to vary significantly from system to system. On current machines, hard faults have been reported as often as every 4 to 8 hours on the Blue Waters system [7], and (detected) L1-cache soft errors as often as every 5 hours on a large BlueGene/L system [8]. The next-generation supercomputers could have a mean time to failure of about 30 minutes [21].

Many of the existing algorithms in use today were derived and analyzed without taking account of the effect of these kinds of faults. We believe that the dawning of the exascale era poses new, and exciting, challenges to the numerical analyst in understanding and analyzing the behavior of numerical algorithms on a fault-prone architecture. Our view is that on future exascale systems, the possible impact of faults on the performance of a numerical algorithm must be taken fully into account in the analysis of the method.

In order to alleviate the impact of faults and ensure resilience in a fault-prone environment, several techniques have been proposed and implemented in various parts of the hardware-software stack. Checkpointing on the system and the application level as well as replication of critical program sections are widely used [5, 7, 15]. These techniques can be coupled with statistical analysis, fault models, and hardware health data [7]. On the application level, algorithm-based fault tolerance (ABFT) describes techniques that duplicate application data to create redundancy [16]. ABFT has been explored in the context of sparse linear algebra [19, 20], and specifically for matrix-vector products in stationary iterative solvers [8–10, 17, 22]. All methods have in common that a balance needs to be struck between protecting against corruption of results and keeping the overhead reasonable.

The multigrid method is the workhorse for distributed solution of linear systems but little is known about its resiliency properties and convergence behavior in a fault-prone environment [12, 17]. The current chapter presents a summary of our recent work addressing this problem [1, 2].

The outline of the remainder of this chapter is as follows: We give a short introduction to multilevel methods in Section 2. In Section 3, we introduce a model

for faults and show simulations of the convergence behavior of a fault-prone two-level method for a finite element method. Finally, in Section 4, we summarize the analytic bounds on the convergence rate and illustrate their behavior with further simulations. We refer the interested reader for further details and proof to the articles [1, 2].

2 Multilevel Methods

Let $\Omega \subset \mathbb{R}^d$ be a polygonal domain and set $V := H_0^1(\Omega)$. Starting from an initial triangulation \mathcal{T}_0 of Ω into simplices, we obtain \mathcal{T}_l through uniform refinement of \mathcal{T}_{l-1} . We define the finite element spaces $V_l := \{v \in H_0^1(\Omega) \cap C(\bar{\Omega}) \text{ such that } v|_K \in \mathbb{P}_1(K), \forall K \in \mathcal{T}_l\}$, and set $n_l := \dim V_l$. For $f \in H^{-1}(\Omega)$, consider the well-posed problem:

$$\text{Find } u \in V \text{ such that } a(u, v) = L(v), \quad \forall v \in V,$$

where $a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v$ and $L(v) = \int_{\Omega} f v$. The discretized problem is

$$\text{Find } u \in V_l \text{ such that } a(u, v) = L(v), \quad \forall v \in V_l.$$

Let $\phi_l^{(i)}$ for $i = 1, \dots, n_l$ be the global shape function basis of V_l , and ϕ_l the vector of global shape functions. Then, the stiffness matrix and the load vector are defined as $A_l := a(\phi_l, \phi_l)$ and $b_l := L(\phi_l)$, so that the problem becomes

$$\text{Find } u = \phi_l \cdot x_l \in V_l \text{ such that } A_l x_l = b_l. \quad (1)$$

Since $V_{l-1} \subset V_l$, there exists a restriction matrix r_{l+1}^l satisfying $\phi_l = r_{l+1}^l \phi_{l+1}$ along with the corresponding prolongation matrix $p_l^{l+1} = (r_{l+1}^l)^T$. In particular, this means that the stiffness matrix on level l can be expressed in terms of the matrix at level $l+1$:

$$A_l = a(\phi_l, \phi_l) = r_{l+1}^l a(\phi_{l+1}, \phi_{l+1}) p_l^{l+1} = r_{l+1}^l A_{l+1} p_l^{l+1}.$$

We shall omit the sub- and superscripts on r and p whenever it is clear which operator is meant. We shall consider solving the system (1) using the multigrid method [4, 13, 14, 18, 23]. The coarse-grid correction is given by $x_l \leftarrow x_l + p A_{l-1}^{-1} r (b_l - A_l x_l)$, and has iteration matrix $C_l := I - p A_{l-1}^{-1} r A_l$, while the damped Jacobi smoother is given by $S_l = I - \theta D_l^{-1} A_l$, where D_l is the diagonal of A_l and θ the relaxation parameter. The multilevel method for the solution of $A_L x_L = b_L$ is given in Algorithm 1. Here, ν_1 and ν_2 are the number of pre- and post-smoothing steps, γ is the number of coarse-grid corrections, and θ is the smoothing parameter.

Function \mathcal{M}_l (right-hand side b_l , initial guess x_l)

```

if  $l = 0$  then return  $A_0^{-1}x_0$            (Exact solve on coarsest grid)
else
  for  $i \leftarrow 1$  to  $v_1$  do
     $x_l \leftarrow x_l + \theta D_l^{-1} (b_l - A_l x_l)$            (Pre-smoothing)
   $d_{l-1} \leftarrow r (b_l - A_l x_l)$            (Restriction to coarser grid)
   $e_{l-1}^{(0)} \leftarrow 0$ 
  for  $j \leftarrow 1$  to  $\gamma$  do
     $e_{l-1}^{(j)} \leftarrow \mathcal{M}_{l-1} (d_{l-1}, e_{l-1}^{(j-1)})$        (Solve on coarser grid)
   $x_l \leftarrow x_l + p e_{l-1}^{(\gamma)}$            (Prolongation to finer grid)
  for  $i \leftarrow 1$  to  $v_2$  do
     $x_l \leftarrow x_l + \theta D_l^{-1} (b_l - A_l x_l)$            (Post-smoothing)
  return  $x_l$ 

```

Algorithm 1: Multilevel method \mathcal{M}_l

3 Fault Model

The first issue is to decide on how the effect of a fault should be incorporated into the analysis of the algorithm. The simplest and most convenient course of action if a component is subject to corruption, or fails to return a value, is to overwrite the value by zero. We therefore propose to model the effect of a fault on a vector using a random diagonal matrix \mathcal{X} , of the form:

$$\mathcal{X} = \begin{pmatrix} \chi_1 & & \\ & \ddots & \\ & & \chi_n \end{pmatrix}, \quad \chi_i = \begin{cases} 1 & \text{with probability } 1 - q, \\ 0 & \text{with probability } q. \end{cases} \quad (2)$$

In particular, if a vector $x \in \mathbb{R}^n$ is subject to faults, then the corrupted version of x is given by $\mathcal{X}x$. If all χ_i are independent, we will call the random matrix a matrix of *component-wise* faults. More generally, we shall make the following assumption on the set \mathcal{S} of all the involved faults matrices \mathcal{X} :

- (A) There exist constants v , $C_e \geq 0$, and for each $\mathcal{X} \in \mathcal{S}$ there exists $e_{\mathcal{X}} \geq 0$ such that for all $\mathcal{X} \in \mathcal{S}$:
- \mathcal{X} is a random diagonal matrix.
 - $\|\text{Var}[\mathcal{X}]\|_2 = \max_{i,j} |\text{Cov}[\mathcal{X}_{ii}, \mathcal{X}_{jj}]| \leq v$.
 - $\mathbb{E}[\mathcal{X}] = e_{\mathcal{X}} I$.
 - $|e_{\mathcal{X}} - 1| \leq C_e v$.

We will think of v as being small. This means that each of the fault matrices \mathcal{X} is close to the identity matrix with high probability. Obviously, the model for component-wise faults introduced above satisfies these assumptions.

In the remainder of this work, we write random matrices in bold letters. If a symbol appears twice, the two occurrences represent the same random matrix and are therefore dependent. If the power of a random matrix appears, we mean the product of identically distributed independent factors.

In summary, we shall model the application of a fault-prone Jacobi smoother as:

$$x_l \leftarrow x_l + \mathcal{X}_l^{(\text{pre/post})} \theta D_l^{-1} (b_l - A_l x_l),$$

which has the same form as a standard Jacobi smoother in which the iteration matrix has been replaced by a random iteration matrix:

$$S_l^{(\text{pre/post})} = I - \mathcal{X}_l^{(\text{pre/post})} \theta D_l^{-1} A_l.$$

Here and in what follows, $\mathcal{X}_l^{(*)}$ are generic fault matrices. Suppose that only the calculation of the update can be faulty, and that the previous iterate is preserved. This could be achieved by writing the local components of the current iterate to nonvolatile memory or saving it on an adjacent node. The matrices $\mathcal{X}_l^{(\text{pre/post})}$ and D_l^{-1} commute, so that without loss of generality, we can assume that there is just one fault matrix, because any faults in the calculation of the residual can be included in $\mathcal{X}_l^{(\text{pre/post})}$ as well. Moreover, while the application of D_l^{-1} and A_l to a vector is fault-prone, we assume that the entries of D_l^{-1} and A_l itself are not subject to corruption, since permanent changes to them would effectively make it impossible to converge to the correct solution. The matrix entries are generally computed once and for all, and can be stored in nonvolatile memory which is protected against corruption. The low writing speed of NVRAM is not an issue since the matrices are written at most once.

The fault-prone two-level method has iteration matrix:

$$E_{TG,l}(v_1, v_2) = \left(S_l^{(\text{post})} \right)^{v_2} C_l \left(S_l^{(\text{pre})} \right)^{v_1},$$

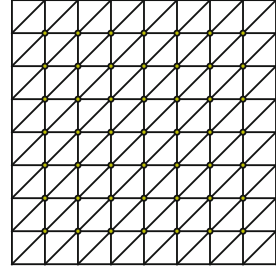
where

$$C_l = I - \mathcal{X}_l^{(p)} p A_{l-1}^{-1} \mathcal{X}_l^{(r)} r \mathcal{X}_l^{(A)} A_l.$$

Similar arguments as for the smoother can be used to justify the model of faults for the coarse-grid correction. The fault-prone multilevel algorithm is given in Algorithm 2.

In order to illustrate the effect of the faults on the convergence of the algorithm, we apply the two-level version of Algorithm 2 with one step of pre- and post-smoothing using a damped Jacobi smoother with optimal smoothing parameter $\theta = \frac{2}{3}$ for a piecewise linear discretization of the Poisson problem on a square domain.

Fig. 1 Mesh for the square domain.



Function \mathcal{M}_l (right-hand side b_l , initial guess x_l)

```

if  $l = 0$  then return  $A_0^{-1}x_0$            (Exact solve on coarsest grid)
else
  for  $i \leftarrow 1$  to  $v_1$  do
     $x_l \leftarrow x_l + \mathcal{X}_l^{(\text{pre})} \theta D_l^{-1} (b_l - A_l x_l)$            (Pre-smoothing)
   $d_{l-1} \leftarrow \mathcal{X}_{l-1}^{(r)} r \mathcal{X}_l^{(A)} (b_l - A_l x_l)$            (Restriction to coarser grid)
   $e_{l-1}^{(0)} \leftarrow 0$ 
  for  $j \leftarrow 1$  to  $\gamma$  do
     $e_{l-1}^{(j)} \leftarrow \mathcal{M}_{l-1} (d_{l-1}, e_{l-1}^{(j-1)})$            (Solve on coarser grid)
   $x_l \leftarrow x_l + \mathcal{X}_l^{(p)} p e_{l-1}^{(\gamma)}$            (Prolongation to finer grid)
  for  $i \leftarrow 1$  to  $v_2$  do
     $x_l \leftarrow x_l + \mathcal{X}_l^{(\text{post})} \theta D_l^{-1} (b_l - A_l x_l)$            (Post-smoothing)
  return  $x_l$ 

```

Algorithm 2: Fault-prone multilevel method \mathcal{M}_l

The domain is partitioned by a uniform triangulation (Figure 1), and we inject component-wise faults as given in Equation (2). We plot the evolution of the residual norm over 30 iterations for varying number of degrees of freedom n_L and different probabilities of faults q in Figure 2 on page 289. We can see that as q increases, the curves start to fan out, with a slope depending on the number of degrees of freedom n_L .

4 Summary of Results on Convergence

In [1, 2], a framework for the analysis of fault-prone stationary iterations was proposed. We summarize the obtained convergence results whose proofs can be found in [1, 2].

Theorem 1 ([1]) *Let $\Omega \subset \mathbb{R}^d$ with $\partial\Omega \in C^2$ or Ω convex and let A_l be the stiffness matrices associated to the finite element discretization of a second-order elliptic PDE on a hierarchy of quasi-uniform meshes, and let*

$$E_{TG,L}(v_1, v_2) = \left(S_L^{(\text{post})} \right)^{v_2} C_L \left(S_L^{(\text{pre})} \right)^{v_1}$$

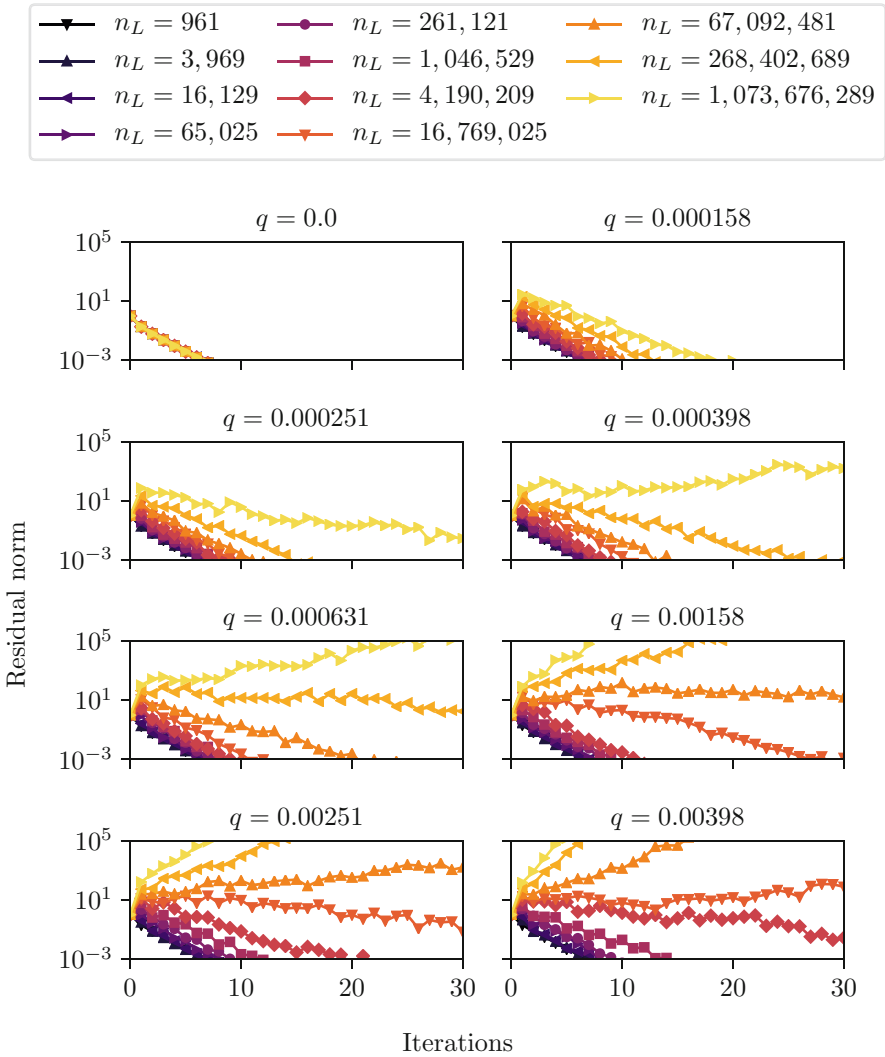


Fig. 2 Evolution of the norm of the residual of the two-level method for the 2d Poisson problem on square domain and component-wise faults in prolongation, restriction, residual, and smoother.

be the iteration matrix of the two-level method with component-wise faults of rate q in prolongation, restriction, residual, and smoother:

$$C_L = I - \mathcal{X}_L^{(p)} p A_{L-1}^{-1} \mathcal{X}_{L-1}^{(r)} r \mathcal{X}_L^{(A)} A_L,$$

$$S_L^{(pre/post)} = I - \mathcal{X}_L^{(pre/post)} D_L^{-1} A_L.$$

Assume that the usual conditions for multigrid convergence hold. Then, the rate of convergence of the fault-prone two-level method is bounded as:

$$\varrho(\mathbf{E}_{TG,L}(v_1, v_2)) \leq \|E_{TG,L}(v_1, v_2)\|_A + C \begin{cases} qn_L^{\frac{4-d}{2d}} & d < 4, \\ q(\log n_L)^{\frac{1}{2}} & d = 4, \\ q & d > 4, \end{cases}$$

where $E_{TG,L}$, C_L , and S_L are the unperturbed two-level iteration matrix, coarse-grid correction, and Jacobi smoother and $\|\cdot\|_A$ is the energy norm. C is independent of L and q .

In Figure 3 (top) on page 291, we plot the estimated rate of convergence of the two-level method for the 2d Poisson problem introduced above. We use 1000 iterations to estimate $\varrho(\mathbf{E}_{TG,L}(1, 1))$ for component-wise faults with varying probability q and varying problem size n_L . Moreover, we plot the behavior predicted by Theorem 1 and the level of $\varrho(\mathbf{E}_{TG,L}(1, 1)) = 1$. We can see that their slope matches.

Experimentally, it can be observed that the result also holds for the case of an L-shaped domain and for block-wise faults, provided that the size of the blocks is fixed, even though the conditions of Theorem 1 are not satisfied.

The above results indicate that two-level methods without protection of some components cannot be used in a fault-prone environment. In order to preserve convergence independent of the number of degrees of freedom, we will have to protect one of the fault-prone operations. The cheapest operations are the restriction and the prolongation. The next result shows that the two-level method converges, if the prolongation is protected.

Theorem 2 ([1]) *Let*

$$E_{TG,L}(v_1, v_2) = \left(S_L^{(post)}\right)^{v_2} C_L \left(S_L^{(pre)}\right)^{v_1}$$

with smoother and coarse-grid correction given by:

$$\begin{aligned} S_L^{(pre/post)} &= I - \mathcal{X}_L^{(pre/post)} D_L^{-1} A_L, \\ C_L &= I - p A_{L-1}^{-1} \mathcal{X}_{L-1}^{(r)} r \mathcal{X}_L^{(A)} A_L. \end{aligned}$$

Provided that the usual conditions for multigrid convergence and Assumption (A) with

$$\mathcal{S} = \left\{ \mathcal{X}_{L-1}^{(r)}, \mathcal{X}_L^{(A)}, \mathcal{X}_L^{(pre)}, \mathcal{X}_L^{(post)} \right\}$$

hold for some $v \geq 0$, we find for any level L that the fault-prone two-level method converges with a rate bounded as:

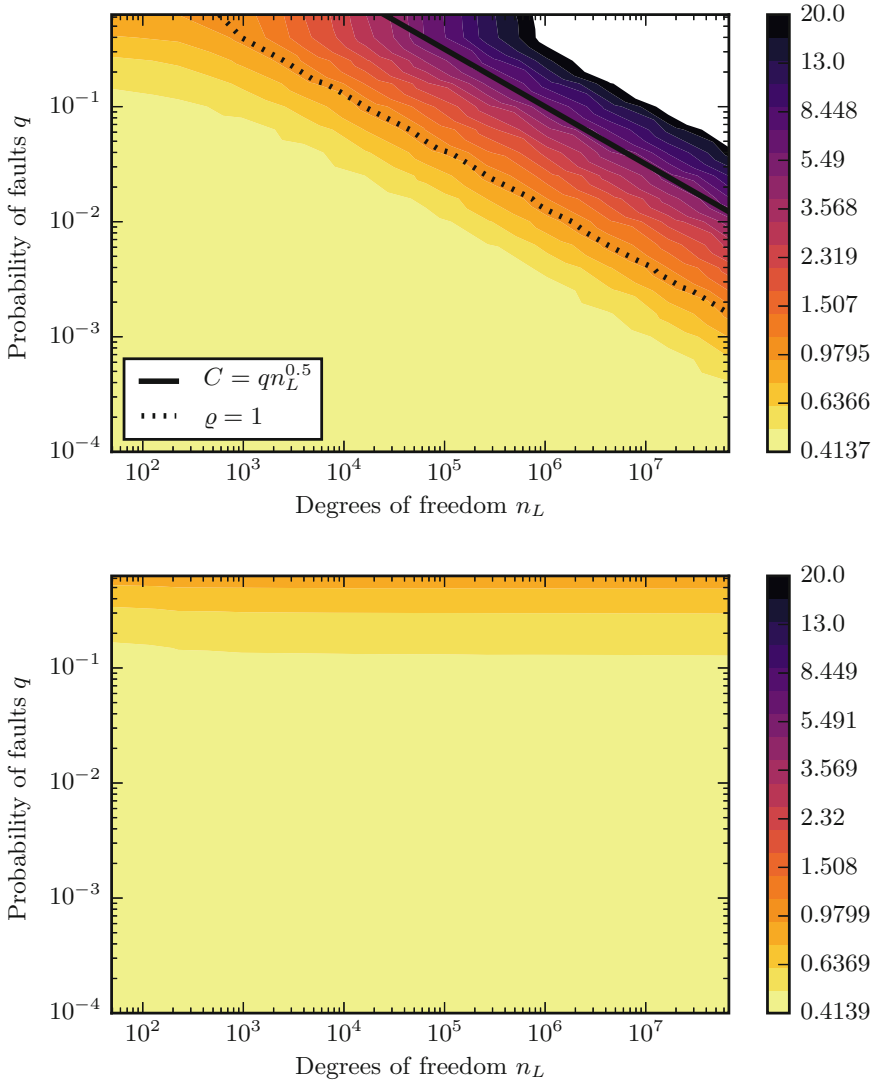


Fig. 3 Asymptotic convergence rate $\varrho(\mathbf{E}_{TG,L}(1, 1))$ of the fault-prone two-level method for the 2d Poisson problem on square domain with component-wise faults in prolongation, restriction, residual, and smoother (top) and protected prolongation (bottom).

$$\varrho(\mathbf{E}_{TG,L}(v_1, v_2)) \leq \|\mathbf{E}_{TG,L}(v_2, v_1)\|_2 + Cv.$$

and C is independent of v and L .

We note that the result holds for more general types of faults including block-wise faults. In Figure 3 (bottom) on page 291, we plot the rate of convergence of

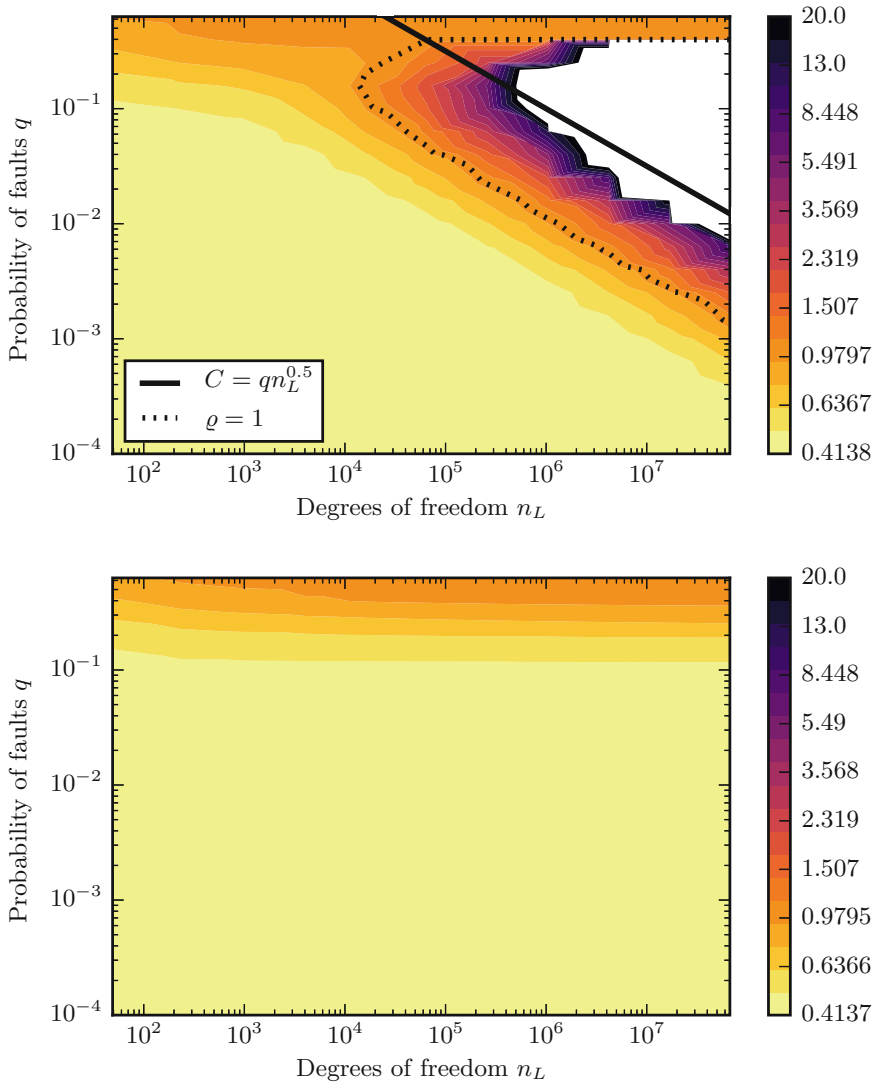


Fig. 4 Asymptotic convergence rate $\rho(\mathbf{E}_L(1, 1, 2))$ of the fault-prone multilevel method for the 2d Poisson problem on square domain with component-wise faults in prolongation, restriction, residual, and smoother (top) and protected prolongation (bottom).

the two-grid method for the already discussed example, this time with protected prolongation. We can see that the rate is essentially independent of the size of the problem and even is smaller than one for large values of q . The protection can be achieved by standard techniques such as replication. In order to retain performance, the protected prolongation could be overlapped with the application of the post-smoother.

The following theorem shows that the result carries over to the multilevel case:

Theorem 3 ([2]) *Provided that the usual conditions for multigrid convergence and Assumption (A) with*

$$S = \bigcup_{l=1}^L \{ \mathbf{x}_{l-1}^{(r)}, \mathbf{x}_l^{(A)}, \mathbf{x}_l^{(pre)}, \mathbf{x}_l^{(post)} \}$$

hold, the number of smoothing steps is sufficient and that v sufficiently small, the perturbed multilevel method converges with a rate bounded by:

$$\varrho(\mathbf{E}_L(v_1, v_2, \gamma)) \leq \begin{cases} \frac{\gamma}{\gamma-1} \xi + Cv, & \gamma \geq 2, \\ \frac{2}{1+\sqrt{1-4C_*\xi}} \xi + Cv, & \gamma = 2, \end{cases}$$

where

$$\xi = \max_{l \leq L} \|E_{TG,l}(v_2, v_1)\|_2,$$

and C_* and C depend on v_1, v_2 and the convergence rate of the two-level method, but are independent of L and v .

We also plot the rate of convergence of fault-prone multilevel algorithms with one coarse-grid correction for component-wise faults and protected prolongation in Figure 4 on page 292, and observe the predicted behavior.

In the current work, we proposed a probabilistic model for the effect of faults involving random diagonal matrices. We gave a summary of the theoretical analysis of the model for the rate of convergence of fault-prone multigrid methods which show that the standard multigrid method is not resilient. Finally, we presented a modification of the standard multigrid algorithm that is fault resilient.

References

1. Ainsworth, M., Glusa, C.: Is the Multigrid Method Fault Tolerant? The Two-Grid Case. *SIAM Journal on Scientific Computing* **39**(2), C116–C143 (2017). <https://doi.org/10.1137/16M1100691>
2. Ainsworth, M., Glusa, C.: Is the multigrid method fault tolerant? The Multilevel Case. *SIAM Journal on Scientific Computing* **39**(6), C393–C416 (2017)
3. Avižienis, A., Laprie, J.C., Randell, B., Landwehr, C.: Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing* **1**(1), 11–33 (2004). <https://doi.org/10.1109/tdsc.2004.2>
4. Bramble, J.H.: *Multigrid methods*, vol. 294. CRC Press (1993)
5. Cappello, F.: Fault tolerance in petascale/exascale systems: Current knowledge, challenges and research opportunities. *International Journal of High Performance Computing Applications* **23**(3), 212–226 (2009)

6. Cappello, F., Geist, A., Gropp, B., Kale, L., Kramer, B., Snir, M.: Toward exascale resilience. *International Journal of High Performance Computing Applications* **23**(4), 374–388 (2009). <https://doi.org/10.1177/1094342009347767>
7. Cappello, F., Geist, A., Gropp, W., Kale, S., Kramer, B., Snir, M.: Toward exascale resilience: 2014 update. *Supercomputing frontiers and innovations* **1**(1), 5–28 (2014). <https://doi.org/10.14529/jsfi140101>
8. Casas, M., de Supinski, B.R., Bronevetsky, G., Schulz, M.: Fault Resilience of the Algebraic Multi-grid Solver. In: *Proceedings of the 26th ACM International Conference on Supercomputing, ICS '12*, pp. 91–100. ACM, New York, NY, USA (2012). <https://doi.org/10.1145/2304576.2304590>
9. Cui, T., Xu, J., Zhang, C.S.: An error-resilient redundant subspace correction method. *Computing and Visualization in Science* **18**(2), 65–77 (2017). <https://doi.org/10.1007/s00791-016-0270-6>.
10. Elliott, J., Mueller, F., Stoyanov, M., Webster, C.G.: Quantifying the impact of single bit flips on floating point arithmetic. Tech. Rep. ORNL/TM-2013/282, Oak Ridge National Laboratory (2013)
11. Glusa, C.: Multigrid and domain decomposition methods in fault-prone environments. Ph.D. thesis, Brown University (2017).
12. Göddeke, D., Altenbernd, M., Ribbrock, D.: Fault-tolerant finite-element multigrid algorithms with hierarchically compressed asynchronous checkpointing. *Parallel Computing* **49**, 117–135 (2015)
13. Hackbusch, W.: *Multi-grid methods and applications*, vol. 4. Springer-Verlag Berlin (1985). <https://doi.org/10.1007/978-3-662-02427-0>
14. Hackbusch, W.: *Iterative solution of large sparse systems of equations, Applied Mathematical Sciences*, vol. 95. Springer-Verlag, New York (1994). <https://doi.org/10.1007/978-1-4612-4288-8>
15. Herault, T., Robert, Y.: *Fault-Tolerance Techniques for High-Performance Computing*. Springer International Publishing (2015). <https://doi.org/10.1007/978-3-319-20943-2>
16. Huang, K.H., Abraham, J.: Algorithm-based fault tolerance for matrix operations. *Computers, IEEE Transactions on* **100**(6), 518–528 (1984)
17. Huber, M., Gmeiner, B., Rüde, U., Wohlmuth, B.: Resilience for massively parallel multigrid solvers. *SIAM Journal on Scientific Computing* **38**(5), S217–S239 (2016)
18. McCormick, S.F., Briggs, W.L., Henson, V.E.: *A multigrid tutorial*. SIAM, Philadelphia (2000)
19. Shantharam, M., Srinivasmurthy, S., Raghavan, P.: Characterizing the Impact of Soft Errors on Iterative Methods in Scientific Computing. In: *Proceedings of the International Conference on Supercomputing, ICS '11*, pp. 152–161. ACM, New York, NY, USA (2011). <https://doi.org/10.1145/1995896.1995922>
20. Sloan, J., Kumar, R., Bronevetsky, G.: Algorithmic approaches to low overhead fault detection for sparse linear algebra. In: *Dependable Systems and Networks (DSN), 2012 42nd Annual IEEE/IFIP International Conference on*, pp. 1–12. IEEE, Boston, MA, USA (2012)
21. Snir, M., Wisniewski, R.W., Abraham, J.A., Adve, S.V., Bagchi, S., Balaji, P., Belak, J., Bose, P., Cappello, F., Carlson, B., et al.: Addressing failures in exascale computing. *International Journal of High Performance Computing Applications* **28**(2), 129–173 (2014)
22. Stoyanov, M., Webster, C.: Numerical Analysis of Fixed Point Algorithms in the Presence of Hardware Faults. *SIAM Journal on Scientific Computing* **37**(5), C532–C553 (2015). <https://doi.org/10.1137/140991406>
23. Trottenberg, U., Oosterlee, C.W., Schüller, A.: *Multigrid*. Academic Press Inc., San Diego, CA (2001). With contributions by A. Brandt, P. Oswald and K. Stüben

Visualization of Data: Methods, Software, and Applications



Gintautas Dzemyda, Olga Kurasova, Viktor Medvedev,
and Giedrė Dzemydaitė

1 Introduction

Data science and data analytics become a key for solving complex problems. They enable new data-driven scientific discoveries. Data science combines various aspects of computer science, statistics, applied mathematics, and visualization that allow to analyze massive amounts of data and to extract knowledge.

Research proves that the human brain processes visualizations better. Graphical representations of multidimensional data are widely used in research and applications of many disciplines. Human participation plays an essential role in most decisions when analyzing data. The huge storage capacity and computational power of computers cannot replace the human flexibility, perceptual abilities, creativity, and general knowledge. Real data in technologies and sciences are often high-dimensional, i.e., data are characterized by many features which can get numerical values. For human perception, the data must be represented in some structured form (direct visualization methods [16]) or in a low-dimensional space, usually 2D (projection and dimensionality reduction methods [16]). The goal of projection methods is to represent the multidimensional data in a low-dimensional space so that certain properties of the structure of the data were preserved as faithfully as possible.

G. Dzemyda (✉) · O. Kurasova · V. Medvedev
Vilnius University, Institute of Data Science and Digital Technologies, Akademijos St. 4, 08663
Vilnius, Lithuania
e-mail: gintautas.dzemyda@mii.vu.lt; olga.kurasova@mii.vu.lt; viktor.medvedev@mii.vu.lt

G. Dzemydaitė
Vilnius University, Faculty of Economics and Business Administration, Saulėtekio Ave 9, 10221
Vilnius, Lithuania
e-mail: giedre.dzemydaite@ef.vu.lt

Here, the starting data may be interpreted as points in the multidimensional space. After the dimensionality reduction, we get a corresponding set of projected points on a plane. Such an approach to data visualization is applied in this paper.

In this paper, the visualization methods based both on dimensionality reduction and on artificial neural networks are applied to the visual efficiency analysis of regional economic development to evaluate how regional resources are reflected in the economic results. Both regions and indicators (features) characterizing them are analyzed visually.

2 Methods of Data Visualization

Data visualization is an extensive and essential area of data science. Recently data scientists observe large amounts of data, which is hard to process at once. To analyze data, different data mining algorithms have been developed. In this paper, we focus on visualization and dimensionality reduction algorithms which reduce data dimensionality from original high-dimensional space to target dimension (2D in visualization case). The main idea is to represent a large set of some measured features with a reduced set of more informative ones and to present data visually.

The data from the real world can usually be described by an array of features x_1, x_2, \dots, x_n . A combination of values of all features characterizes a particular data item (object) $X_i = (x_{i1}, x_{i2}, \dots, x_{in})$, $i \in \{1, \dots, m\}$, from the whole set X_1, X_2, \dots, X_m , where n is the number of features, m is the number of analyzed objects. If X_1, X_2, \dots, X_m are described by more than one feature, the data are called multidimensional data. X_i , $i = 1, \dots, m$ are often interpreted as points in the multidimensional space.

Several approaches have been proposed for representing multidimensional data in a lower-dimensional space. Comprehensive reviews of the dimensionality reduction-based visualization methods are presented in [15, 16]. These reviews cover most of visualization approaches: direct methods, dimensionality reduction methods, artificial neural networks-based methods, and manifold learning methods.

Principal component analysis (PCA) [26] is a well-known method for dimensionality reduction. It can be used to display the data as a linear projection in a subspace of the original data space so that it preserves the variance of the data best. However, the interpretation of principal components can be difficult at times.

An alternative approach to PCA for dimensionality reduction is multidimensional scaling (MDS) [6]. It is a classical approach that maps the original high-dimensional space to a lower-dimensional one by using the information on the proximities between the objects in the original space so that the proximities between the corresponding data points are preserved. Despite the fact that the MDS problem was addressed by a few decades ago, this problem remains relevant nowadays [7, 16, 21]. To solve the MDS problems, the methods based on function majorization are

applied [22]. The MDS problem is formulated as an optimization problem where the number of variables depends on the number of points analyzed. The so-called stress function (projection error) is minimized. In the case of a large number of variables, the optimization problem becomes complex. There are developed approaches to solve the MDS problem using parallel computing technologies [38]. Usually, the Euclidean distance is used as the metric proximity of multidimensional points. Non-metric proximities are used, when, e.g., the proximity is obtained by some expert opinion. Proximity measure of projected points influences the complexity of the optimization problem. For example, if city-block distances are used, the optimization problem becomes even more complicated, because the objective function is not everywhere differentiable [16]. There exists a multitude of variants of MDS with different stress functions and their optimization algorithms [6]. Commonly, MDS stress function is minimized using the SMACOF algorithm, based on the iterative majorization.

Isometric feature mapping (ISOMAP) can also be assigned to the group of multidimensional scaling. ISOMAP is designed for dimensionality reduction as well as for visualization of multidimensional data [35]. Using ISOMAP, an assumption that the points of the initial space are located on a lower-dimensional manifold is done. Therefore, the geodesic distances are used as a measure of proximity between the points analyzed. The assumption is applied in locally linear embedding (LLE) [16] and Laplacian eigenmaps [3], too.

Several artificial neural network-based methods for visualizing the multidimensional data have been proposed, including SAMANN [30, 31] and SOM [27]. The specific back-propagation-like learning rule SAMANN allows a feed-forward artificial neural network to learn Sammon's mapping, that is one of MDS algorithms, in an unsupervised way. After the training, the neural network can project previously unseen points, using the obtained generalized mapping rule.

Self-organizing map (SOM) is another artificial neural network suitable for data visualization [27, 36]. A distinctive characteristic of this type of neural networks is that they can be used for both clustering and visualization of multidimensional data. SOM is a set of neurons connected to one another via a rectangular or hexagonal topology. Each neuron is defined by the place on SOM and by the so-called codebook vector. After the SOM learning, the analyzed data points X_1, X_2, \dots, X_m are presented to SOM and winning neurons are found for each data point. In such a way, the data points are distributed on the SOM table. Using SOM, we can draw a table with cells corresponding to the neurons. The cells corresponding to the neurons-winners are filled with the order numbers or names of data points. Some cells may remain empty. One can make a decision visually on the distribution of the points in the n -dimensional space in accordance with their distribution among the cells of the table.

The dimensionality reduction methods can be applied for the additional mapping of the codebook vectors of the winning neurons on the plane. The ways of combining SOM and MDS have been proposed and investigated in [14, 28, 29].

3 Software for Data Visualization

Data visualization is an important part of the processes of knowledge discovery in medicine, economics, telecommunication, and various scientific fields. For several decades, the attention was focused not only on new data mining methods, but also on software implementing these methods [12, 16, 34]. However, most of the widely used software solutions were designed as standalone desktop applications. They include methods for data preprocessing, classification, clustering, regression, and dimensionality reduction [34].

Software systems, in which data visualization methods are implemented, were developed to facilitate solving the data mining problems. Therefore, they have become trendy among researchers. Recently, software applications have been developed under the SOA (service-oriented architecture) paradigm. Thus, some new data mining systems are based on web services. Attempts are made to develop scalable, extensible, interoperable, modular, and easy-to-use data mining systems. Some popular data mining systems have been reoriented to web services. PCA, MDS, and SOM are implemented only in the open source data mining tools: WEKA [24], Orange [11], KNIME [5], RapidMiner [25], and R. Commercial statistics software, such as Statistica (StatSoft), SAS/STAT, IBM SPSS Modeler, as well as MATLAB include visualization methods, too.

To create an approach for intelligent visualization of multidimensional data with the intention to avoid drawbacks of the existing data mining tools, a new cloud-based web application, called DAMIS (DATA Mining System), is under development. The open source solution DAMIS (<http://www.damis.lt>) [32] implements data mining solution as a service for the end user and has a graphical user-friendly interface that allows researchers to carry out a data analysis, to visualize multidimensional data, to investigate data projections and data item similarities, as well as to identify the influence of individual features and their relationships by various data mining algorithms, taking advantage of cloud computing. The relations between the implemented data mining algorithms and the graphical user interface are supported by web services (SOAP—simple object access protocol). The algorithms for multidimensional data preprocessing, clustering, classification, and dimensionality reduction-based visualization have been implemented. To analyze the data by the implemented data mining algorithms, the user initializes and manages an experiment by constructing scientific workflows, i.e., the order in which the data mining algorithms are executed. The user can modify the created workflow by adding or removing nodes and reuse it for other data. The user can select high-performance computing resource from the proposed alternatives. All the performed experiments including the workflows and data analysis results are saved in the cloud. Thus the management of the accomplished experiments can be accessible by the user on demand.

The following dimensionality reduction-based methods for multidimensional data visualization are implemented in DAMIS: principal component analysis (PCA),

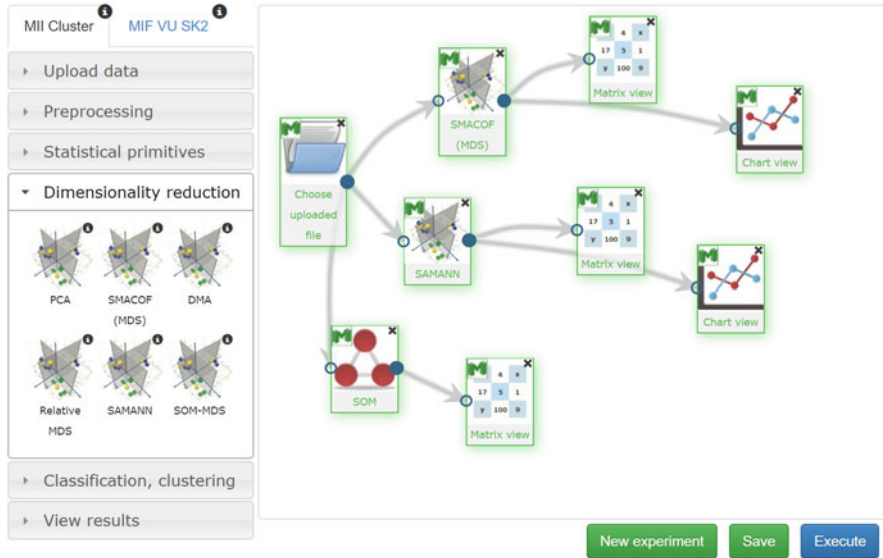


Fig. 1 DAMIS scientific workflow for data visualization

multidimensional scaling (MDS), relative MDS, SAMANN, and combination of SOM and MDS (SOM-MDS).

The DAMIS user can benefit from:

- Use of the individual online data repository for data storage and easy management;
- Selection of high-performance computing resources;
- Use of the latest version of the data mining algorithms;
- Executing scientific workflows of data mining experiments on the selected cloud-based infrastructure and management of the accomplished experiments.

An example of the scientific workflow for a particular data visualization problem is given in Figure 1. The workflow consists of the connected nodes. Each node corresponds to either data preprocessing or visualization algorithm. The nodes for data file uploading and viewing the results are also used when constructing the workflow.

4 Visual Analysis of Regional Economic Development

Resources and their availability could be understood as preconditions for value creation in regions, but the economic results even in regions with a similar level of resources tend to vary, and this poses some inefficiencies [8, 17]. Intensive

investment in science and technology does not necessarily bring the high efficiency of the innovation systems and cannot guarantee success in innovation [23].

This research aims to evaluate economic efficiency by measuring how regional resources are reflected in the economic results. The analysis focuses on new member states of European Union that joined the EU in 2004. Regions of the Central and Eastern European Union and the Baltic States are analyzed. These territories have a comparatively common experience of market and infrastructure development and could be evaluated as economically comparable units. As the EU Cohesion policy funds are distributed according to NUTS2 regional level, this regional level is selected for the analysis. Selected NUTS2 territorial units are of 800 thousand to 3 million population size. Because of these criteria, more densely populated areas form separate region, as Prague or Bratislava, that consist of capitals and are more economically developed than less urban territories. Overall, 40 regions are involved in the analysis of 8 countries: Estonia (EE), Latvia (LV), Lithuania (LT), Poland (PL), Czech Republic (CZ), Slovakia (SK), Slovenia (SL), and Hungary (HU). Data are from the Eurostat database [19]. The EU nomenclature of territorial units for statistics ensures harmonized standards in the collection and transmission of regional data, guarantees that published regional statistics are based on comparable data, and enables the analysis and comparison of the socioeconomic situation of the regions [19].

The efficiency analysis of regional decision-making units is commonly related to the regional production function, measuring the level of resources and outputs achieved [9, 33]. Recent researches of efficiency are based on the Farrell's idea [20] that the economic efficiency consists of a technical efficiency and an allocative efficiency. The technical efficiency is the ability of a decision-making unit to achieve the maximum output with the available economic resources. The resource allocation efficiency reveals the ability to choose the best resource ratio according to the market prices.

Recent studies have focused on measurement of regional efficiency from different perspectives to get insights which regional resources are not insufficiently used for the creation of economic value. The efficiency of transport infrastructure and human capital was measured in [18, 33]. The efficiency of regional innovation systems, evaluating regional resources and their feasible economic output, is considered in, e.g., [2, 4, 17, 37]. The empirical analysis of efficiency is mainly carried out by mathematical programming techniques based on a data envelopment analysis [10].

The regional economic development is affected by various processes that could be described by multidimensional data. By applying multidimensional data visualization methods, the aim is to acquire information on the processes in the regional economy, the interaction, and similarity of indicators. Regions X_1, X_2, \dots, X_m are characterized by common economic indicators (features) x_1, x_2, \dots, x_n . For the analysis, a matrix of regional indicators $X = \{x_{ij}, i = 1, \dots, m, j = 1, \dots, n\}$ is set, where m is the number of regions ($m = 40$), n is the number of indicators ($n = 11$), and x_{ij} is the value of the j -th indicator for the i -th region. All the

Table 1 Regional economic indicators

x_1 (GDP)	Gross domestic product, PPS (purchasing power standard) per inhabitant
x_2 (HR_SC_TH)	Persons employed in science and technology, per cent of total population
x_3 (HR_TER)	Persons with tertiary education, per cent of total population
x_4 (R&D_EXP)	Cumulative intramural research and development expenditure (during previous 5 years), PPS per inhabitant
x_5 (PATENT)	Patents (during previous 5 years), per capita
x_6 (HTC_EMP)	Employment in high-technology manufacturing and knowledge-intensive high-technology services, per cent of total employment
x_7 (LMTC_EMP)	Employment in low and medium technology manufacturing, per cent of total employment
x_8 (AGR_EMP)	Employment in agriculture, forestry, and fishing; mining and quarrying, per cent of total employment
x_9 (POP_DENS)	Population density, inhabitants per km^2
x_{10} (ROAD_DENS)	Railway network density, total railway lines per thousand km^2
x_{11} (TOUR_NGHT)	Nights spent at tourist accommodation establishments, per thousand inhabitants

indicators involved in the analysis are significantly correlated with the GDP per inhabitant and reveal the level of regional resources or economic outputs (Table 1).

Since we have a data matrix $X = \{x_{ij}, i = 1, \dots, m, j = 1, \dots, n\}$ for the analysis, we can visualize both the set of regions (depending on indicators) and the set of indicators (depending on regions). The dimensionality of data on regions is n , and the dimensionality of data on indicators is m .

Multidimensional scaling (MDS) and principal component analysis (PCA) were applied to get the visual insight into the similarities of regions. 11-dimensional data were visualized. Visual distribution of points corresponding to the regions is given in Figures 2 and 3. The regions of the same country are marked by the same label. The multidimensional data are normalized by z -score before the visualization is applied. We do not present legends and units for both axes in the figures, because we are interested in observing the interlocation of points on a plane only. The matrix X before and after normalization is available online in MIDAS archive [1].

The values obtained by any indicator x_k can depend on the values of other indicators $x_j, j = 1, \dots, n, k \neq j$, i.e., the indicators are correlated. There exist groups of indicators that characterize different properties of the region. The problem is to discover knowledge about the similarities not only of separate indicators but also about their groups. Here, the visual analysis of indicators (features) would be valuable. The number of regions m is much larger than the number of indicators n ($n \ll m$), so the visualization of indicators becomes complicated by a direct use of matrix X . The method proposed in [13] gives a theoretically grounded possibility for a new view to the analysis of correlations, in particular, to the visualization of data stored in correlation matrices.

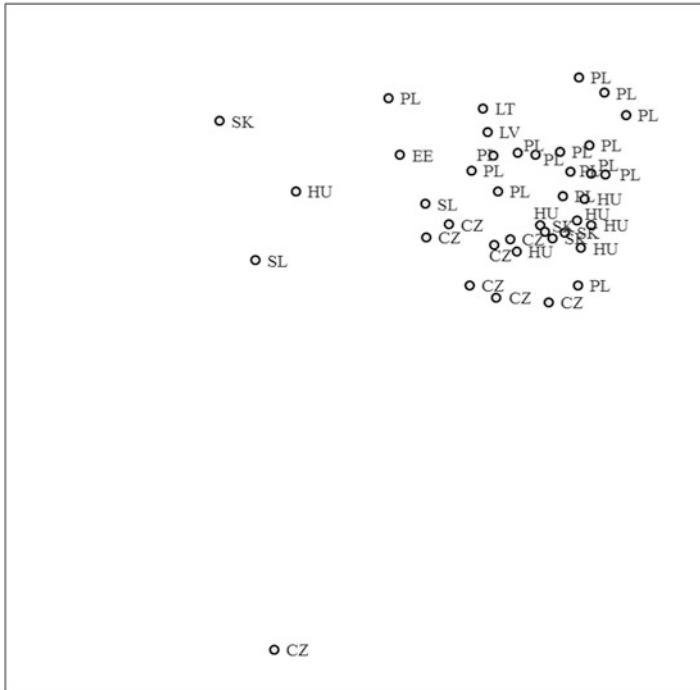


Fig. 2 Visualization of regional economic development data using MDS

The correlation matrix $R = \{r_{ij}, i, j = 1, \dots, n\}$ of indicators x_1, x_2, \dots, x_n can be calculated by analyzing the regions that compose the set $X = \{X_1, X_2, \dots, X_m\}$. Here r_{ij} is the correlation coefficient of x_i and x_j .

Let S^n be a subset of an n -dimensional Euclidean space R^n containing such points $V = (v_1, v_2, \dots, v_n), V \in S^n$, where $\|V\| = \sum_{k=1}^n v_k^2 = 1$, i.e., S^n is a unit sphere. The idea of analysis of a correlation is based on determining a set of points $V_1, V_2, \dots, V_n \in S^n$, corresponding to indicators x_1, x_2, \dots, x_n so that $\cos(V_i, V_j) = r_{ij}$, if all $r_{ij} \geq 0$, and $\cos(V_i, V_j) = r_{ij}^2$, if the correlation matrix contains both positive and negative values. It means that V_i and V_j will be closer, if the absolute value of r_{ij} is larger. Basing on the matrix of cosines $K = \{\cos(V_i, V_j), i, j = 1, \dots, n\}$, it is possible to create a set of points $V_s = (v_{s1}, v_{s2}, \dots, v_{sn}) \in S^n, s = 1, \dots, n$, as follows: $v_{sk} = \sqrt{\lambda_k} e_{ks}, k = 1, \dots, n$. Here λ_k is the k th eigenvalue of matrix $K, E_k = (e_{k1}, e_{k2}, \dots, e_{kn})$ is a normalized eigenvector (the length of eigenvector is equal to one) that corresponds to the eigenvalue λ_k . In a result, we get a matrix V , whose n rows are n -dimensional points $V_s = (v_{s1}, v_{s2}, \dots, v_{sn}), s = 1, \dots, n$. Each row corresponds to the particular indicator. If a visualization method is applied to such matrix, we see visually the set of considered indicators.

The matrices R and V are available online in MIDAS archive [1].

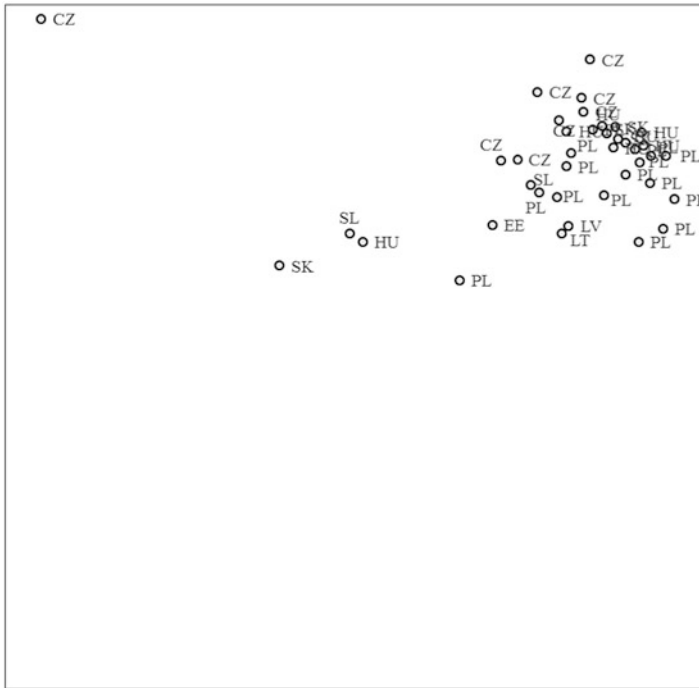


Fig. 3 Visualization of regional economic development data using PCA

The data in matrix V was analyzed using MDS (Figure 4) and SAMANN (Figure 5). Visualization of indicators using SOM of various grid dimensions (4×4 and 5×5) is presented in Tables 2 and 3.

5 Discussion and Conclusions

In this paper, the methods and software for visualization of multidimensional data are reviewed. Web-based DAMIS solution for data analysis is applied to the visual efficiency analysis of regional economic development to evaluate how regional resources are reflected in the economic results. The projection methods (PCA, MDS) and artificial neural networks (SOM, SAMANN) were used. Both regions and indicators characterizing the regions were analyzed visually.

Analysis of regions allows comparing the regions among themselves. Most regions of the same country have a tendency to be similar. However, we observe economical differences between various regions of the same country, too. In Figures 2 and 3, we see one definite outlier from CZ. It is Prague city. It is not the only outlier from the main group of regions of the similar economical state.



Fig. 4 Visualization of indicators using MDS

Outliers from HU, SL, SK, and PL are big cities, too. Note that following MDS and PCA, the neighboring and historically related small countries LV and LT are very similar from the economical point of view.

Indicator analysis shows that the core indicator is gross domestic product (GDP). It is located in the center of visual presentations of the set of indicators by MDS, SAMANN, and SOM. The second essential indicator is cumulative intramural research and development expenditure (R&D_EXP). These two indicators are very related and appear in the same cell of SOM (5×5). They are close in SOM (4×4) and in MDS and SAMANN visualizations (see Tables 2, 3 and Figures 4, 5). Remaining indicators are more independent among themselves. From Figures 4, 5 and Tables 2, 3 we can evaluate visually both the main tendencies in pairwise similarities of indicators and tendencies in their grouping.

From the analysis, some inefficiencies are seen. Firstly, HR_TER is apart from GDP and other important indicators, e.g., HTC_EMP, R&D_EXP. This reveals that higher level of tertiary education does not necessarily reflect in economic results of regions. Secondly, AGR_EMP and LMTC_EMP are also apart from GDP and other regional resources. This supposes that less economically developed regions do not have sufficient resources needed for higher value creation and their path to higher GDP could be hardly based on high-technology development or knowledge-intensive activities.

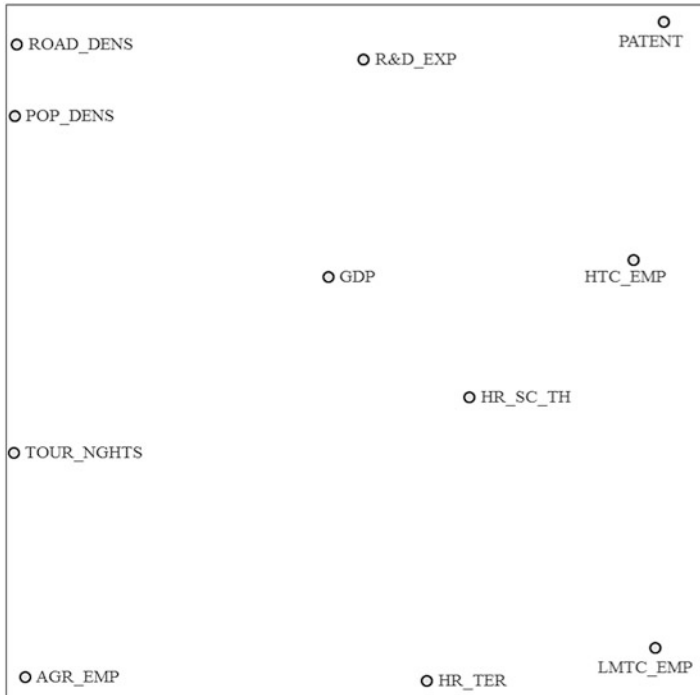


Fig. 5 Visualization of indicators using SAMANN

Table 2 Visualization of indicators using SOM (grid dimension 4 × 4)

AGR_EMP	TOUR_NGHTS		POP_DENS, ROAD_DENS
LMTC_EMP		GDP	R&D_EXP, HTC_EMP
HR_TER	HR_SC_TH		PATENT

Table 3 Visualization of indicators using SOM (grid dimension 5 × 5)

AGR_EMP		ROAD_DENS		POP_DENS
LMTC_EMP		GDP, R&D_EXP		TOUR_NGHTS
HR_TER	HR_SC_TH	HTC_EMP		PATENT

The paper has disclosed a new field for graphical representations of multidimensional data, where the human participation plays an essential role in decisions. Moreover, the visualization has served here as a mean for the sophisticated analysis of data. Its application can uncover non-trivial knowledge from the real-world data.

References

1. MIDAS: The National Open Access Research Data Archive. https://doi.org/10.18279/MIDAS_RegionalData.xlsx.30026
2. Bai, J.: On regional innovation efficiency: evidence from panel data of China's different provinces. *Regional Studies* **47**(5), 773–788 (2013)
3. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: *Advances in Neural Information Processing systems*, pp. 585–591 (2002)
4. Bengoa, M., Martínez-San Román, V., Pérez, P.: Do R&D activities matter for productivity? A regional spatial approach assessing the role of human and social capital. *Economic Modelling* **60**, 448–461 (2017)
5. Berthold, M.R., Cebon, N., Dill, F., Gabriel, T.R., Kötter, T., Meinel, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. In: *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer (2007). https://doi.org/10.1007/978-3-540-78246-9_38
6. Borg, I., Groenen, P.: *Modern Multidimensional Scaling: Theory and Applications*. Springer (2005). <https://doi.org/10.1007/0-387-28981-X>
7. Borg, I., Groenen, P.J., Mair, P.: *Applied Multidimensional Scaling*. Springer Science & Business Media (2012)
8. Cai, Y., Hanley, A.: Innovation rankings: good, bad or revealing? *Applied Economics Letters* **21**(5), 325–328 (2014)
9. Daouia, A., Florens, J.P., Simar, L.: Regularization of nonparametric frontier estimators. *Journal of Econometrics* **168**(2), 285–299 (2012)
10. Daraio, C., Simar, L.: Introducing environmental variables in nonparametric frontier models: a probabilistic approach. *Journal of Productivity Analysis* **24**(1), 93–121 (2005)
11. Demšar, J., Curk, T., Erjavec, A., Gorup, C., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., Zupan, B.: Orange: Data mining toolbox in Python. *Journal of Machine Learning Research* **14**, 2349–2353 (2013)
12. Dubitzky, W. (ed.): *Data Mining Techniques in Grid Computing Environments*. John Wiley and Sons, Ltd (2009). <https://doi.org/10.1002/9780470699904.ch1>
13. Dzemyda, G.: Visualization of a set of parameters characterized by their correlation matrix. *Computational Statistics & Data Analysis* **36**(1), 15–30 (2001)
14. Dzemyda, G., Kurasova, O.: Heuristic approach for minimizing the projection error in the integrated mapping. *European Journal of Operational Research* **171**(3), 859–878 (2006). <https://doi.org/10.1016/j.ejor.2004.09.011>
15. Dzemyda, G., Kurasova, O., Medvedev, V.: Dimension reduction and data visualization using neural networks. In: I. Maglogiannis, K. Karpouzis, M. Wallace, J. Soldatos (eds.) *Emerging Artificial Intelligence Applications in Computer Engineering*, *Frontiers in Artificial Intelligence and Applications*, vol. 160, pp. 25–49. IOS Press (2007)
16. Dzemyda, G., Kurasova, O., Žilinskas, J.: *Multidimensional Data Visualization: Methods and Applications*, *Springer Optimization and its Applications*, vol. 75. Springer (2013). <https://doi.org/10.1007/978-1-4419-0236-8>
17. Dzemydaitė, G., Dzemyda, I., Galinienė, B.: The efficiency of regional innovation systems in new member states of the European Union: a nonparametric DEA approach. *Economics and Business* **28**(1), 83–89 (2016)
18. Dzemydaitė, G., Galinienė, B.: Evaluation of regional efficiency disparities by efficient frontier analysis. *Ekonomika* **92**(4), 21 (2013)
19. Eurostat-European Commission and others: *Regions in the European Union. Nomenclature of territorial units for statistics*. Tech. rep., NUTS 2010/EU-27. Luxembourg: Publications Office of the European Union (2011)
20. Farrell, M.J.: The measurement of productive efficiency. *Journal of the Royal Statistical Society. Series A (General)* **120**(3), 253–290 (1957)

21. Groenen, P., Borg, I.: Past, present, and future of multidimensional scaling. *Visualization and Verbalization of Data* pp. 95–117 (2014)
22. Groenen, P.J., van de Velden, M.: Multidimensional scaling by majorization: A review. *Journal of Statistical Software* **73**(8), 1–26 (2016)
23. Guan, J., Chen, K.: Modeling the relative efficiency of national innovation systems. *Research Policy* **41**(1), 102–115 (2012)
24. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. *SIGKDD Explorations Newsletter* **11**(1), 10–18 (2009). <https://doi.org/10.1145/1656274.1656278>
25. Hofmann, M., Klinkenberg, R.: *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. Chapman & Hall/CRC (2013)
26. Jolliffe, I.: *Principal Component Analysis*. Springer, Berlin (1986). <https://doi.org/10.1007/b98835>
27. Kohonen, T.: Overture. In: *Self-Organizing Neural Networks: Recent Advances and Applications*, pp. 1–12. Springer-Verlag, New York, NY, USA (2002)
28. Kurasova, O., Molytė, A.: Integration of the self-organizing map and neural gas with multidimensional scaling. *Information Technology and Control* **40**(1), 12–20 (2011)
29. Kurasova, O., Molytė, A.: Quality of quantization and visualization of vectors obtained by neural gas and self-organizing map. *Informatica* **22**(1), 115–134 (2011)
30. Mao, J., Jain, A.K.: Artificial neural networks for feature extraction and multivariate data projection. *IEEE Transactions on Neural Networks* **6**(2), 296–317 (1995). <https://doi.org/10.1109/72.363467>
31. Medvedev, V., Dzemyda, G., Kurasova, O., Marcinkevičius, V.: Efficient data projection for visual analysis of large data sets using neural networks. *Informatica* **22**(4), 507–520 (2011)
32. Medvedev, V., Kurasova, O., Bernatavičienė, J., Treigys, P., Marcinkevičius, V., Dzemyda, G.: A new web-based solution for modelling data mining processes. *Simulation Modelling Practice and Theory* (2017)
33. Schaffer, A., Simar, L., Rauland, J.: Decomposing regional efficiency. *Journal of Regional Science* **51**(5), 931–947 (2011)
34. Talia, D., Trunfio, P.: *Service-oriented Distributed Knowledge Discovery*. Chapman and Hall/CRC (2012). <https://doi.org/10.1201/b12990-4>
35. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500), 2319–2323 (2000)
36. Venskus, J., Treigys, P., Bernatavičienė, J., Medvedev, V., Vozňák, M., Kurmis, M., Bulbenkiene, V.: Integration of a self-organizing map and a virtual pheromone for real-time abnormal movement detection in marine traffic. *Informatica* **28**(2), 359–374 (2017)
37. Vila, L.E., Cabrer, B., Pavía, J.M.: On the relationship between knowledge creation and economic performance. *Technological and Economic Development of Economy* **21**(4), 539–556 (2015)
38. Žilinskas, J.: Parallel branch and bound for multidimensional scaling with city-block distances. *Journal of Global Optimization* **54**(2), 261–274 (2012)

HPC Technologies from Scientific Computing to Big Data Applications



L. M. Patnaik and Srinidhi Hiriyannaiah

1 Introduction

Scientific research in disciplines such as astronomy, genomics, neuroscience and social sciences faces a major bottleneck in the areas of processing and analytical capabilities due to increase in the size of the data over the years. The basic research starts with a small amount of data and code running on a single-node workstation, then extended to a distributed framework for further improvement. Scientific applications typically involve many small task applications and are connected via dataflow patterns. Some of the examples of scientific applications are sequence alignment tool BLAST [1] and high energy physics histogram analysis [2]. The data workflow systems used for building such scientific applications include HTCondor [5], MPI [4] and Hadoop [3]. Each of these approaches differs in supporting scalability, parallelism and job tracking and is limited by lack of fault tolerance, rigid programming model and flexibility.

The new buzz word both in science and industry is Big Data. Data intensive science plays a key role in the emerging Big data technologies indicating a new form of technology for different human activities in the world. Such activities include social media, digital services, e-commerce, logistics, transportation, etc. Technologies like cloud computing and ubiquitous network provide necessary platforms for organizing such processes in data collection, storing, processing and visualization. There is a need for big data technologies to align with scientific

L. M. Patnaik

National Institute of Advanced Studies (NIAS), Bengaluru, India

e-mail: lalitblr@gmail.com

S. Hiriyannaiah (✉)

Ramaiah Institute of Technology, Department of CSE, Bengaluru, India

e-mail: srinidhi.hiriyannaiah@gmail.com

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_19

309

discovery methods for model improvement and prediction. The use of advanced statistical and data mining methods helps in finding patterns and discover the value in data.

HPC systems are essential for solving scientific problems using thousands of processors and high throughput networks. The use of HPC systems for solving big data problems is a recent trend across HPC centres. Big data refers to complex, diverse and massive data that contain structured, semi-structured and unstructured data. Such data sets are difficult to store, process and analyse with traditional database technologies. New technologies and advanced analytics are needed for data management, distribution and processing. More recent trend is to extend the HPC systems from computationally intensive scientific domain to data-driven domain or Big Data. The problems related to Big data are proven to be solvable by HPC systems because of high processing power, low-latency and non-blocking communications. Some of the analytical problems related to cyber security, social networks, medical and health-care typically run on HPC systems. There are numerous approaches developed to support HPC systems for such data intensive problems such as MapReduce, MPI, OpenMP and OpenCL.

The significant contribution of HPC is towards solving scientific application and this now expanding towards big data analytics. There is an increasing demand for resources required for big data applications with HPC. Hence, the existing or alternating solutions that extend the capabilities of HPC systems for big data applications are needed. This paper focuses on the aspects of emergence of certain technologies in HPC from scientific computing to big data applications.

2 Evolution of Scientific Computing and Big Data

Scientific computing applications involve computational modelling, advanced science and engineering techniques where time and cost play key roles. Some of these computational models in the field of astronomy involve diverse information such as stellar dynamics, black hole behaviour and dark matter, which bring new insights and appreciate the use of experimental data. In the field of climate analysis, computational models that capture the effect of green house gases, deforestation, illustrate the effects of human behaviour with climate change. Thus, scientific computing or computational science is a multidisciplinary field involving design, prototyping, optimizing and reducing time and cost of different mathematical models for solving complex science and engineering problems. For example, with the help of advanced simulation techniques Cummins is able to build faster and less expensive diesel engines [18], Good Year is able to design safe tyres in less time and Boeing is able to build fuel-efficient aircrafts [22].

In scientific computing, mathematical models and numerical solutions are used to solve social, scientific and engineering problems. The models often require computing resources in large amount for performing large-scale experiments to reduce the time required and computational complexity. These needs are made available by HPC infrastructure in the form of clusters or grids [6]. With the

availability of computer grids, network of machines with high power to perform large experiments are provided for scientific computing applications. Computing grids offer services such as dynamic discovery of services, network of resources for meeting the application requirements. Grid computing has been used in many scientific computing applications [7–9].

The recent trends are Big Data, machine learning and predictive analytics which are seen as the next shift in the paradigm of computing allowing research in computational science or scientific computing with instruments. Machine learning applications are used in the field such as healthcare in identifying the spread of diseases, high-energy physics and molecular biology [16]. The successive generations of computing and large-scale scientific instruments bring in advanced new capabilities of engineering with technical challenges and economical trade-offs. In a broad sense, data generation is increasing in most of the scientific domains causing them to be data intensive and requiring more computational abilities. High performance computing and Big Data are intrinsically tied to each other to meet the forecasting demands of the scientific computing applications.

The factors that are responsible for increasing data in big data and HPC applications are as follows:

- HPC systems are able to run data intensive applications that involve considerable modelling and simulation problems at large scale.
- Rapid increase in larger, scientific instruments and sensor networks leading to Internet of Things (IoT).
- Evolving multidisciplinary engineering domains and transformation of them into data-driven science such as biology, archaeology and linguistics.
- Increasing data volume from the results of stochastic models, parametric models and other iterative problem-solving methods in various fields of engineering.
- Need to perform near real-time analytics in various commercial applications.
- Newer platforms and programming models for analytics such as MapReduce/Hadoop, semantic analysis and graph analytics.

3 Software Stack for HPC and Big Data

The software stacks for Big data and HPC are different due to the class of the applications that belong to these areas. Software should essentially comprise of software for power management, operating systems, cluster monitoring, scheduling and performance monitoring. The difference between the HPC and Big data software stack is summarized as shown in Figure 1.

- **Programming languages:** C, C++ are the commonly used languages for the development of HPC applications. Java, Scala, R, Python are used for Big data applications.
- **File systems:** Hadoop is the fundamental file system used in most of the big data applications. Hadoop supports a variety of formats of data with databases such as

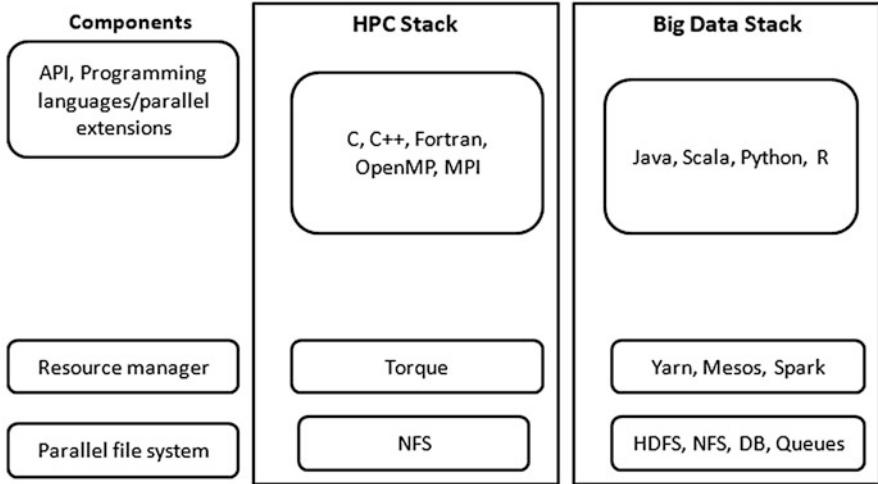


Fig. 1 HPC and Big Data stack

Hive and HBase. The significance of Hadoop is distributed file system (HDFS). It splits the file into various blocks and distributes them across the file system to different network nodes. HPC applications rely on the legacy network file system (NFS) to store and access the data across different locations in the cluster.

- **Operating system:** In big data systems, frameworks are built around JVM and thus support portability across different OS and platforms. In the case of HPC systems, as they rely on C/C++, recompilation of the applications is needed initially to run on different operating systems.
- **Hardware:** The main intention of the design of Hadoop was to use a conventional hardware for scalability with less cost. It uses the conventional Ethernet sockets for communication and distribution of data among clusters. HPC software makes efficient use of remote directory memory access (RDMA) technology such as Infiniband interconnect.
- **Debugging:** HPC systems support various debugging and profiling tools such as Scalasca and Tau. In the case of big data and Hadoop, resource managers such as YARN, Mesos are the only solutions for debugging the jobs in the cluster.

4 Solutions for HPC with Big Data

4.1 Cloud Computing

Grid computing with pre-packaged environment is used for running scientific applications. The use of grid computing for scientific computing applications is limited to tools, API, hosting operating systems, services that are required at specific

times. In practice, options available in grid computing are not elastic to cover the needs of scientific applications. The applications that run on grids are implemented as a set of workflows and parallel processes. Scientific applications have to be reorganized into such workflows for execution in grid environments which is more expensive. Cloud computing is one of the solutions that address the challenges of grid computing [12].

Cloud computing is pay-as-you-go service model offering many services from hardware to the application using virtualization. The main advantage of use of cloud computing is scalability, computing infrastructure can be scaled up or down based on the application requirements. A distributed approach in cloud computing provides ease of access to large infrastructures to carry out various experiments. Pay-as-you go model of cloud computing allows renting the infrastructure as per the usage and required resources without any capacity planning initially. The computing stack varies from the hardware to the application level in cloud computing. At the hardware level, appliances are provided by means of virtualization. At the application level, software itself is provided as a service to the users. HPC and big data applications can explore wide spectrum of options available with cloud computing and other services [11].

In order to move from the traditional science grids to cloud computing, different solutions are available in the market through multiple cloud providers. Amazon web services and VMware solutions provide infrastructure level resources through virtualization and thus offer resources on demand as a service. Microsoft Azure and Google App engine provide virtualization at the application level rather than at the hardware level to support specific requirements of the applications and thus utilize the large infrastructure of clusters of Google and Microsoft. Aneka [10] is also one of the cloud solutions offering better quality of service to the end users through virtualization, support of multiple programming models such as MapReduce [13]. The other services offered by such cloud providers include end-to end platform as a service as well to the users.

4.2 MPI

MPI [19] is a library for message passing to program distributed and shared memory systems with bindings to C, C++ with support of both collective and point-to-point communications. MPI consists of a collection of processes where the starter process or rank 0 process will be a separate process of the application. `MPI_init()` is used to start the MPI process initializing the communication among the distributed processes using `MPI_COMM_WORLD` as the initiator. `MPI_finalize()` is used to terminate the process. Many standards have been released from MPI 1.0 to MPI 3.0. Initially in MPI 1.0, the functionality of two-sided and collective communication was provided and later in MPI 2.0, the one-sided communication was supported. Random memory access (RMA) specification was later introduced in MPI 3.0 for addressing global address space models. The notion of RMA specification is

that memory segment can be remotely accessed by other processes using get/put operations. MPI supported initially a sequential I/O approach where a master process was responsible for both read and write processes. The master process reads the data and sends it to other processes for further operations. In case of write, all the processes send the data to the master for aggregating the data into files. Thus, there was a bottleneck on I/O limiting the performance and scalability of applications. In MPI 2.0 sequential I/O was eliminated and parallel processing was enabled to provide non-contiguous data layout for memory and files.

4.3 CUDA (*Compute Unified Device Architecture*)

General purpose computing with GPU or graphics processors has been studied for several years [19]. The different domains that have explored the use of GPUs and their applications include genomics, climate analysis, molecular engineering and other scientific domains. The use of GPUs for development of such applications is limited to the use of graphics processing language and its flexibility for non-graphic programmers. NVIDIAs CUDA provides ease of access for scalable parallel programming and scientific computing. CUDA platform provides massive general purpose multi-threaded architecture with up to 128 processor cores, programmable in C and capable of executing billions of floating-point operations per second. The ubiquitous nature of NVIDIA GPUs provides a compelling platform for HPC and Big Data applications. Many applications related to computation and searching are being rapidly developed using the CUDA programming model [20].

The three basic features that are offered by CUDA are hierarchy of thread groups, shared memories and barrier synchronization. In developing a parallel algorithm for a problem using CUDA, first the portion of the problem that can be solved independently needs to be identified and further these parts are executed across independent parallel threads [20]. For example, consider the following snippet for addition of two matrices a and b. In this example, each element of the resultant matrix is not dependent on the operations of the other elements. Thus, in this case the execution can be parallelized across different elements in the resultant matrix.

```

for ( i =0; i <m; i ++):
    for ( j =0; j <n; j ++):
        r e s u l t [ i ] [ j ] = a [ i ] [ j ] + b [ i ] [ j ];
    end
end

```

A basic organization of CUDA with threads, blocks and grids is shown in Figure 2 [20]. The programmer has to specify the number of threads per block and the number of blocks in the grid. CUDA supports 512 threads in one thread block. Each thread is given a unique id threadIdx numbered from 0,1..blockdim-1 within a thread block. Each block is given a unique id blockIdx in the grid. In order to execute the kernels on CUDA, the data that resides on the host need to be transferred to the

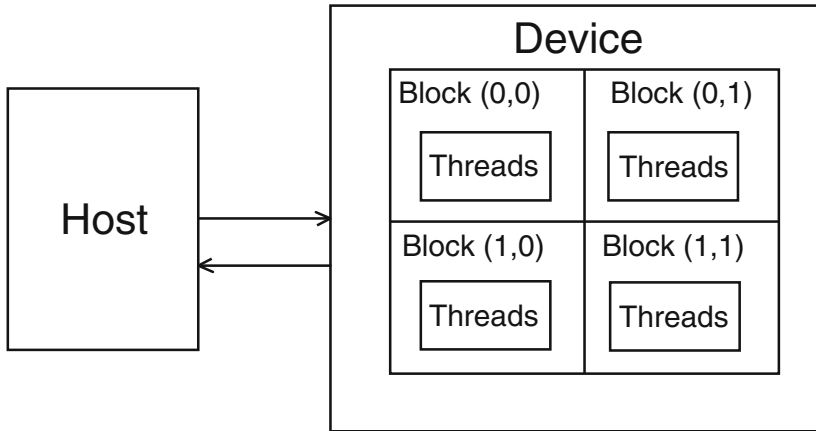


Fig. 2 CUDA thread organization

device or GPU first, next invoke the kernel with appropriate number of threads and then finally execute it on the device. To perform these operations, certain keywords are used that suggest that the kernels are executed on the device and not on the host. Some of the keywords used are `cudaMemcpy()`, `_global_`, `<<<>>>`.

4.4 OpenMP

OpenMP [21] is a program interface that provides multi-threaded programming model with shared memory for parallel programming. The interface provides directives with bindings for C and C++. These directives are used by the programmers to specify the piece of code that needs to be executed in parallel. The directives are in the form `#pragma omp` that specifies the compiler to switch to the openmp clause and syntax for parallel execution.

The following example demonstrates parallelizing of a loop with OpenMP. The directive `#pragma omp parallel for` specifies that the for loop has to be executed in parallel. The number of threads is divided based on the cores used for program execution. In case of this example, if for loop is executed on a 2-core processor with `n` value as 2, the number of threads will be 2 pertaining to `b[0]` and `b[1]` that represents first and second thread, respectively.

```
void simple(int n, float *a, float *b)
{
    int i;
    //pragma omp parallel for
    for (i=1; i<n; i++) /* i is private by default */
        b[i] = (a[i] + a[i-1]) / 2.0;
}
```

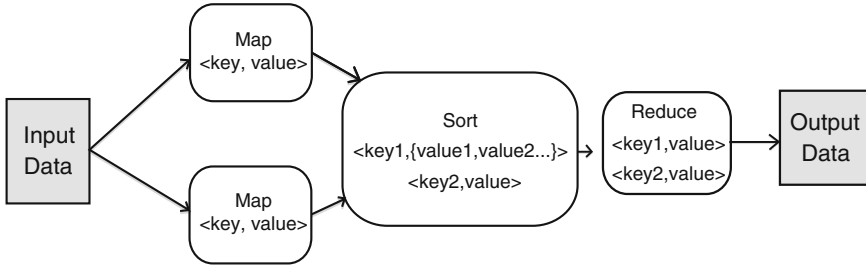


Fig. 3 MapReduce programming model

4.5 MapReduce

MapReduce is a programming model for large-scale processing of data stored on Hadoop-based systems [13, 14]. A MapReduce program consists of map and reduce tasks for processing. In both map and reduce tasks, key-value pairs are used for input and output data processing. In map method, the input data is read and transformed into intermediate key-value pairs. In the reduce method, the intermediate pairs are aggregated into key-value pairs by summing up the values based on the key. The main features of MapReduce programming include load balancing, failure tolerant and recovery from the failed tasks. In Hadoop, each query runs a MapReduce job by reading the data, running the map task to produce the intermediate key-value pairs and finally running the reduce task to output the aggregated key-value pairs.

The map function takes the input and splits into $\langle \text{key}, \text{value} \rangle$ pairs. A sort phase exists in between these map and reduce phases that aggregates several input key-value pairs from the map phase into intermediate $\langle \text{key}, \text{value} \rangle$ pairs. The reduce phase picks up the intermediate key-value pairs and produces the output $\langle \text{key}, \text{value} \rangle$ pairs that can be understood by the user [13]. The basic framework of MapReduce programming model is shown in Figure 3.

5 Big Data Applications

Big data analytics is one of the advanced analysis methods for massive data. The data-driven applications and business analytics are becoming popular nowadays in many areas such as environmental science, genomics, social networks, social computing and business intelligence, and smart health. In this section, the different and promising applications from typical big data domains are listed as follows. The specific areas of big data technology that scientific researchers are deploying in order to see significant increases in their ability to manage the scientific data deluge are discussed below.

- **Scientific applications:** Sensors deployed in a smart environment such as smart city, self-driving cars generate lot of data that needs to be analysed. In the field of physics and astronomical experiments, scientific studies are being carried out for designing, operating and analysing sensor networks and detectors. One of the applications includes developing earth observation system (EOS) for gathering information and approaches to analyse information about earth's physical, chemical and biological properties via remote sensing technologies, to improve social and economic well-being and its applications for weather forecasting, monitoring and responding to natural disasters, and climate change predictions, etc.
- **Genomics:** An area of scientific research that is benefiting from HPC and big data merging is genomics. Applications relate to big data analytics of various genomic data, studying biomedical applications such as evolution of pathogen affecting humans, cancer biology and chemical–genetic interactions in drug design [17].
- **Environmental Sciences:** A large amount of climate and ecosystem data is now available from satellite and ground-based sensors, while climate model simulations offer huge potential for understanding the behaviour of the Earth's ecosystem and for advancing the science of climate change [15].
- **Business analytics:** A large number of e-commerce platforms are interested in knowing the user information such as their likely products of purchase and offer personalized recommendations. In the field of recommender systems, non-negative matrix factorization methods are employed for recommendations [23]. Recommender systems play an important role for decision making and analytics. In [23], more details on the necessity of recommender systems for business analytics are discussed.

However, with such wide variety of applications and convergence of big data and scientific computing, certain challenges exist. Some of the challenges are outlined below.

- **Data acquisition and management:** In the pursuit of the development of applications related to Big data, the main challenge is the acquisition of the data that exists in different formats. Since the data exists in various formats either in structured, semi-structured or unstructured, it is a challenge to have a single platform for accessing and managing all the types of data. The same challenge also applies to scientific applications wherein large amount of data needs to be managed in an efficient way for further processing. Hadoop and other distributed systems with SQL support can help in overcoming this challenge.
- **Data access and processing:** Once the data is acquired and stored in different storage systems, then processing of data needs to be addressed carefully with certain methodologies. These methodologies can be modelling, simulation, prediction, forecasting and others. Thus, in order to process the data efficient data mining and machine learning methods can be applied.

- **Data understanding:** Another challenge that exists with big data applications is the understanding of the domain of the data. The data differs from one domain to another domain in different areas such as chemical, physics, petroleum, energy and requires specific methodologies to be incorporated for analysis.

6 Conclusions

HPC and Big data analytics are an emerging platform to address the need for large-scale data processing and decision making. The technologies related to Big data and HPC such as cloud computing, MPI, OpenMP, CUDA and MapReduce are evolving and changing the existing traditional databases with effective data organization and workloads for processing with machine learning techniques. The maturity of these technologies is accelerating several areas of scientific computing that involve data intensive applications. This paper has discussed an overview of evolution of HPC technologies starting from scientific computing to Big data analytics. The solutions help to construct a programming model that will handle both computational and data intensive applications while meeting user's expectations with regard to programmability, performance portability and fault tolerance, for developing applications related to HPC and Big data.

References

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3), 403–410.
2. Ekanayake, J., Pallickara, S., & Fox, G. (2008). Mapreduce for data intensive scientific analyses. In *eScience, 2008. eScience'08. IEEE Fourth International Conference on* (pp. 277–284). IEEE.
3. Apache. Apache Hadoop. <http://hadoop.apache.org/>.
4. Gropp, W., Lusk, E., Doss, N., & Skjellum, A. (1996). A high-performance, portable implementation of the MPI message passing interface standard. *Parallel computing*, 22(6), 789–828.
5. Litzkow, M. J., Livny, M., & Mutka, M. W. (1988, June). Condor-a hunter of idle workstations. In *Distributed Computing Systems, 1988., 8th International Conference on* (pp. 104–111). IEEE.
6. Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: the Condor experience. *Concurrency and computation: practice and experience*, 17(24), 323–356.
7. Foster, I., & Kesselman, C. (Eds.). (2003). *The Grid 2: Blueprint for a new computing infrastructure*. Elsevier.
8. Chetty, M., & Buyya, R. (2002). Weaving computational Grids: How analogous are they with electrical Grids?. *Computing in Science & Engineering*, 4(4), 61–71.
9. Chin, J., Harvey, M. J., Jha, S., & Coveney, P. V. (2005). Scientific grid computing: The first generation. *Computing in science & engineering*, 7(5), 24–32.
10. Vecchiola, C., Chu, X., & Buyya, R. (2009). Aneka: a software platform for .NET-based cloud computing. *High Speed and Large Scale Scientific Computing*, 18, 267–295.

11. Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., ... & Zaharia, M. (2009). Above the clouds: A Berkeley view of cloud computing (Vol. 17). Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley.
12. Buyya, R., Yeo, C. S., Venugopal, S., Broberg, J., & Brandic, I. (2009). Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility. *Future Generation computer systems*, 25(6), 599–616.
13. Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1), 107–113.
14. White, T. (2012). Hadoop: The definitive guide. “O’Reilly Media, Inc”.
15. <https://eosps.nasa.gov/>
16. John Walker, S. (2014). Big data: A revolution that will transform how we live, work, and think.
17. Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–1014.
18. <https://cumminsengines.com/industrial-diesel-engines>
19. USING, M. (1994). Portable parallel programming with the Message-Passing Interface.
20. Kirk, D. B., & Wen-Mei, W. H. (2016). Programming massively parallel processors: a hands-on approach. Morgan kaufmann.
21. Tinetti, F. G. (2010). Using OpenMP: Portable Shared Memory Parallel Programming. *Journal of Computer Science & Technology*, 10.
22. Cronin, M. J., & Seid, G. (1983). U.S. Patent No. 4,419,926. Washington, DC: U.S. Patent and Trademark Office.
23. Patnaik, L. M., & Hiriyannaiah, S. (2017). Business Analytics Using Recommendation Systems. In *International Conference on Computational Intelligence, Communications, and Business Analytics* (pp. 35–44). Springer, Singapore.

Part III
Models, Methods, and Applications Based
on Partial Differential Equations

Analysis and Simulation of Time-Domain Elliptical Cloaks by the Discontinuous Galerkin Method



Yunqing Huang, Chen Meng, and Jichun Li

1 Introduction

The study of invisibility cloaks with metamaterials was initiated by Leonhardt and Pendry *et al.* [14, 26] in 2006, since then there have been growing interests in design of cloak devices [27], abstract mathematical analysis of cloaking phenomena [1–4, 7, 8, 12, 13, 15, 24], and numerical simulations of cloaking phenomena by the FDTD method [9, 23], finite element methods [5, 18, 25, 29], and the spectral element method [31, 32].

Broadband cloaking [8, 15, 24] inspired us to pursue the time-domain cloaking simulation and analysis. In 2012, we [19] performed some mathematical analysis for the cylindrical cloak proposed by Pendry *et al.* [26] and developed a finite element time-domain (FETD) method to simulate the cylindrical cloak. Later on, we investigated the analysis and simulation for the elliptical cloak [21], rectangular cloak [6, 17], arbitrary polygonal cloaks [20, 30], the spherical cloak [16], the 2D carpet cloak [22], and total reflection and cloaking by zero index metamaterials [10, 28].

In this paper, we continue our study of the elliptical cloak [21]. More specifically, in Section 2, we establish a stability for the cloaking model. In Section 3, we develop

Y. Huang · C. Meng

Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education, School of Mathematics and Computational Science, Xiangtan University, Xiangtan, China
e-mail: huangyq@xtu.edu.cn; mccm28@163.com

J. Li (✉)

Department of Mathematical Sciences, University of Nevada Las Vegas, Las Vegas, NV 89154-4020, USA
e-mail: jichun.li@unlv.edu

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41, https://doi.org/10.1007/978-3-030-02487-1_20

323

a discontinuous Galerkin method for solving the model. In Section 4, we present four numerical examples showing that our method works for simulating the elliptical cloaking phenomena.

2 The Elliptic Cloak Model and its Stability Analysis

Consider an elliptical shell (cf. Figure 1) with semiaxes a and b in the y direction, and semiaxes ka and kb in the x direction, where k denotes the axis ratio. Note that $k > 1$ yields a horizontal elliptical cloak, while $k < 1$ leads to a vertical elliptical cloak, and $k = 1$ reduces to a circular cloak.

In the elliptical cloak region, where $r := \sqrt{x^2 + k^2y^2} \in (ka, kb)$, the relative permittivity and permeability can be expressed in the Cartesian coordinate system as [11]:

$$\epsilon = \mu = \begin{pmatrix} \epsilon_{xx} & \epsilon_{xy} & 0 \\ \epsilon_{xy} & \epsilon_{yy} & 0 \\ 0 & 0 & \epsilon_{zz} \end{pmatrix}, \tag{1}$$

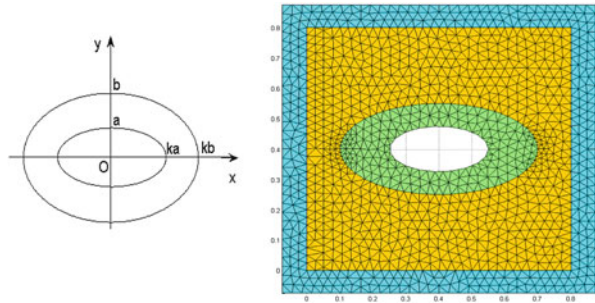
where we denote $R = \sqrt{x^2 + k^4y^2}$,

$$\begin{aligned} \epsilon_{xx} &= \frac{r}{r - ka} + \frac{k^2a^2R^2 - 2kar^3}{(r - ka)r^5}x^2, & \epsilon_{xy} &= \frac{k^2a^2R^2 - ka(1 + k^2)r^3}{(r - ka)r^5}xy, \\ \epsilon_{yy} &= \frac{r}{r - ka} + \frac{k^2a^2R^2 - 2k^3ar^3}{(r - ka)r^5}y^2, & \epsilon_{zz} &= \left(\frac{b}{b - a}\right)^2 \frac{r - ka}{r}. \end{aligned}$$

Furthermore, we denote

$$\lambda_1 = \frac{m - 1}{m + 1}, \quad \lambda_2 = \frac{1}{\lambda_1}, \quad \lambda_3 = \epsilon_{zz}, \quad m = \sqrt{1 + \frac{4r^5(r - ka)}{k^2a^2R^2(x^2 + y^2)}},$$

Fig. 1 (Left) An elliptical cloak in the Cartesian coordinate system; (right) an exemplary mesh (the most inner ellipse is not used in simulation).



$$p_1 = \frac{\epsilon_{xy}}{\sqrt{\epsilon_{xy}^2 + (\lambda_2 - \epsilon_{yy})^2}}, \quad p_2 = \frac{\lambda_2 - \epsilon_{yy}}{\sqrt{\epsilon_{xy}^2 + (\lambda_2 - \epsilon_{yy})^2}}.$$

Note that $\lambda_1, \lambda_3 < 1$ in the cloak region, which means that the medium in that region is nonphysical and is often mapped by dispersive media such as the lossless Drude model:

$$\lambda_i = \epsilon_{\infty i} - \frac{\omega_{pi}^2}{\omega^2}, \quad i = 1, 3, \tag{2}$$

where ω is the angular frequency, ω_{pi} is the plasma frequency, and $\epsilon_{\infty i}$ is the permittivity at infinite frequency. After some lengthy algebra, we derived the governing equations for simulating time-domain elliptical cloaking phenomena as follows [21]:

$$\mathbf{B}_t = -\nabla \times \mathbf{E}, \tag{3}$$

$$\mathbf{D}_t = \nabla \times \mathbf{H}, \tag{4}$$

$$\epsilon_0 \lambda_2 \left(\epsilon_{\infty 1} \mathbf{E}_{t^2} + \omega_{p1}^2 \mathbf{E} \right) = M_E \mathbf{D}_{t^2} + M_F \mathbf{D}, \tag{5}$$

$$\mathbf{B}_{t^2} = \mu_0 \left(\epsilon_{\infty 3} \mathbf{H}_{t^2} + \omega_{p3}^2 \mathbf{H} \right), \tag{6}$$

where we denote u_{tk} for the k -th derivative $\frac{\partial^k u}{\partial t^k}$, 2D vectors $\mathbf{D} = (D_x, D_y)'$ and $\mathbf{E} = (E_x, E_y)'$, and matrices M_E and M_F as follows:

$$M_E = \begin{pmatrix} p_1^2 \lambda_2 + \epsilon_{\infty 1} p_2^2 & p_1 p_2 (\epsilon_{\infty 1} - \lambda_2) \\ p_1 p_2 (\epsilon_{\infty 1} - \lambda_2) & p_2^2 \lambda_2 + \epsilon_{\infty 1} p_1^2 \end{pmatrix}, \quad M_F = \omega_{p1}^2 \begin{pmatrix} p_2^2 & p_1 p_2 \\ p_1 p_2 & p_1^2 \end{pmatrix}.$$

Here and below, we follow the convention to adopt the 2-D curl operators: For a scalar function H , we denote $\nabla \times H = (\frac{\partial H}{\partial y}, -\frac{\partial H}{\partial x})'$; while for the vector $\mathbf{E} = (E_x, E_y)'$, we denote $\nabla \times \mathbf{E} = \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y}$.

To complete our modeling problem, we assume that Equations (3)–(6) are subject to the perfectly conducting (PEC) boundary condition:

$$\hat{\tau} \cdot \mathbf{E} = 0, \quad \text{on } \partial\Omega, \tag{7}$$

and the following initial condition:

$$B(\mathbf{x}, 0) = B_0(\mathbf{x}), \quad \mathbf{D}(\mathbf{x}, 0) = \mathbf{D}_0(\mathbf{x}), \quad H(\mathbf{x}, 0) = H_0(\mathbf{x}), \quad \mathbf{E}(\mathbf{x}, 0) = \mathbf{E}_0(\mathbf{x}),$$

where $\hat{\tau}$ denotes the unit tangential vector along $\partial\Omega$, and $B_0(\mathbf{x}), \mathbf{D}_0(\mathbf{x}), H_0(\mathbf{x}),$ and $\mathbf{E}_0(\mathbf{x})$ are some given functions. The following lemma is proved in [21].

Lemma 1 *The matrix M_F is symmetric and nonnegative definite, and matrix M_E is symmetric and positive definite.*

For our cloaking model, we can prove the following stability.

Theorem 1 *Denote the energy:*

$$\begin{aligned} ENG(t) := & \left[(M_E \mathbf{D}_t, \mathbf{D}_t) + (M_F \mathbf{D}, \mathbf{D}) + \epsilon_0 \mu_0 \epsilon_{\infty 3} (\lambda_2 \epsilon_{\infty 1} M_E^{-1} \mathbf{E}_{t^2}, \mathbf{E}_{t^2}) \right. \\ & + \epsilon_0 \mu_0 \epsilon_{\infty 3} (\lambda_2 \omega_{p1}^2 M_E^{-1} \mathbf{E}_t, \mathbf{E}_t) + \|\nabla \times \mathbf{E}_t\|^2 + \epsilon_0 \epsilon_{\infty 1} (\lambda_2 M_E^{-1} \mathbf{E}_t, \mathbf{E}_t) \\ & \left. + \epsilon_0 (\lambda_2 \omega_{p1}^2 M_E^{-1} \mathbf{E}, \mathbf{E}) + \mu_0 \epsilon_{\infty 3} (H_t, H_t) + \mu_0 (\omega_{p3}^2 H, H) \right] (t). \end{aligned} \quad (8)$$

Here and below, we denote $\|\cdot\|$ for the L_2 norm over domain Ω . Then for any $t \geq 0$, we have

$$ENG(t) \leq C \cdot ENG(0).$$

Proof To make our proof easy to follow, we split the proof into three parts.

(i) Multiplying (5) by \mathbf{D}_t and integrating over Ω , we obtain

$$\frac{1}{2} \frac{d}{dt} [(M_E \mathbf{D}_t, \mathbf{D}_t) + (M_F \mathbf{D}, \mathbf{D})] = \epsilon_0 \epsilon_{\infty 1} (\lambda_2 \mathbf{E}_{t^2}, \mathbf{D}_t) + \epsilon_0 (\lambda_2 \omega_{p1}^2 \mathbf{E}, \mathbf{D}_t),$$

integrating which from $t = 0$ to t yields

$$\begin{aligned} \frac{1}{2} [(M_E \mathbf{D}_t, \mathbf{D}_t) + (M_F \mathbf{D}, \mathbf{D})] (t) &= \frac{1}{2} [(M_E \mathbf{D}_t, \mathbf{D}_t) + (M_F \mathbf{D}, \mathbf{D})] (0) \\ &+ \int_0^t \epsilon_0 \epsilon_{\infty 1} (\lambda_2 \mathbf{E}_{t^2}, \mathbf{D}_t) ds + \int_0^t \epsilon_0 (\lambda_2 \omega_{p1}^2 \mathbf{E}, \mathbf{D}_t) ds. \end{aligned} \quad (9)$$

(ii) Differentiating (5) with respect to t , left-multiplying M_E^{-1} (the inverse of matrix M_E is guaranteed by Lemma 1), and using (4), we have

$$\epsilon_0 \lambda_2 M_E^{-1} (\epsilon_{\infty 1} \mathbf{E}_{t^3} + \omega_{p1}^2 \mathbf{E}_t) = \mathbf{D}_{t^3} + M_E^{-1} M_F \mathbf{D}_t = \nabla \times H_{t^2} + M_E^{-1} M_F \mathbf{D}_t. \quad (10)$$

Multiplying (10) by $\mu_0 \epsilon_{\infty 3}$ and using (6) and (3), we obtain

$$\begin{aligned} \epsilon_0 \mu_0 \epsilon_{\infty 3} \lambda_2 M_E^{-1} (\epsilon_{\infty 1} \mathbf{E}_{t^3} + \omega_{p1}^2 \mathbf{E}_t) &= \nabla \times (B_{t^2} - \mu_0 \omega_{p3}^2 H) + M_E^{-1} M_F \mathbf{D}_t \\ &= -\nabla \times \nabla \times \mathbf{E}_t - \mu_0 \nabla \times (\omega_{p3}^2 H) + M_E^{-1} M_F \mathbf{D}_t. \end{aligned} \quad (11)$$

Multiplying (11) by \mathbf{E}_{t^2} , integrating the resultant over Ω , and using the identity

$$\nabla \times (\omega_{p3}^2 H) = \omega_{p3}^2 \nabla \times H + H \nabla \times (\omega_{p3}^2),$$

we have

$$\begin{aligned} & \frac{\epsilon_0 \mu_0 \epsilon_{\infty 3}}{2} \frac{d}{dt} \left[(\lambda_2 \epsilon_{\infty 1} M_E^{-1} \mathbf{E}_{t^2}, \mathbf{E}_{t^2}) + (\lambda_2 \omega_{p1}^2 M_E^{-1} \mathbf{E}_t, \mathbf{E}_t) \right] + \frac{1}{2} \frac{d}{dt} \|\nabla \times \mathbf{E}_t\|^2 \\ &= -\mu_0 (\omega_{p3}^2 \nabla \times \mathbf{H}, \mathbf{E}_{t^2}) + \mu_0 (\mathbf{H} \nabla \times (\omega_{p3}^2), \mathbf{E}_{t^2}) + (M_E^{-1} M_F \mathbf{D}_t, \mathbf{E}_{t^2}). \end{aligned} \quad (12)$$

Integrating (12) from $t = 0$ to t yields

$$\begin{aligned} & \frac{\epsilon_0 \mu_0 \epsilon_{\infty 3}}{2} \left[(\lambda_2 \epsilon_{\infty 1} M_E^{-1} \mathbf{E}_{t^2}, \mathbf{E}_{t^2}) + (\lambda_2 \omega_{p1}^2 M_E^{-1} \mathbf{E}_t, \mathbf{E}_t) \right] (t) + \frac{1}{2} \|\nabla \times \mathbf{E}_t\|^2 (t) \\ &= \frac{\epsilon_0 \mu_0 \epsilon_{\infty 3}}{2} \left[(\lambda_2 \epsilon_{\infty 1} M_E^{-1} \mathbf{E}_{t^2}, \mathbf{E}_{t^2}) + (\lambda_2 \omega_{p1}^2 M_E^{-1} \mathbf{E}_t, \mathbf{E}_t) \right] (0) + \frac{1}{2} \|\nabla \times \mathbf{E}_t\|^2 (0) \\ & \quad - \int_0^t \mu_0 (\omega_{p3}^2 \mathbf{D}_t, \mathbf{E}_{t^2}) ds + \int_0^t \mu_0 (\mathbf{H} \nabla \times \omega_{p3}^2, \mathbf{E}_{t^2}) ds + \int_0^t (M_E^{-1} M_F \mathbf{D}_t, \mathbf{E}_{t^2}) ds. \end{aligned} \quad (13)$$

(iii) Substituting (3) into (6), we have

$$\mu_0 \epsilon_{\infty 3} \mathbf{H}_{t^2} + \mu_0 \omega_{p3}^2 \mathbf{H} = \mathbf{B}_{t^2} = -\nabla \times \mathbf{E}_t. \quad (14)$$

Multiplying (14) by H_t , then integrating over Ω and using the PEC boundary condition (7), we obtain

$$\frac{1}{2} \frac{d}{dt} \left[\mu_0 \epsilon_{\infty 3} (H_t, H_t) + \mu_0 (\omega_{p3}^2 H, H) \right] = -(\nabla \times \mathbf{E}_t, H_t) = -(\mathbf{E}_t, \nabla \times H_t) \quad (15)$$

Left-multiplying (5) by M_E^{-1} and using (4), we have

$$\epsilon_0 \lambda_2 M_E^{-1} (\epsilon_{\infty 1} \mathbf{E}_{t^2} + \omega_{p1}^2 \mathbf{E}) = \mathbf{D}_{t^2} + M_E^{-1} M_F \mathbf{D} = \nabla \times H_t + M_E^{-1} M_F \mathbf{D}. \quad (16)$$

Multiplying (16) by \mathbf{E}_t and integrating over Ω , we obtain

$$\begin{aligned} & \frac{1}{2} \frac{d}{dt} \left[\epsilon_0 \epsilon_{\infty 1} (\lambda_2 M_E^{-1} \mathbf{E}_t, \mathbf{E}_t) + \epsilon_0 (\lambda_2 \omega_{p1}^2 M_E^{-1} \mathbf{E}, \mathbf{E}) \right] \\ &= (\nabla \times H_t, \mathbf{E}_t) + (M_E^{-1} M_F \mathbf{D}, \mathbf{E}_t). \end{aligned} \quad (17)$$

Adding (15) and (17), and integrating the resultant from 0 to t , we have

$$\begin{aligned} & \frac{1}{2} \left[\epsilon_0 \epsilon_{\infty 1} (\lambda_2 M_E^{-1} \mathbf{E}_t, \mathbf{E}_t) + \epsilon_0 (\lambda_2 \omega_{p1}^2 M_E^{-1} \mathbf{E}, \mathbf{E}) + \mu_0 \epsilon_{\infty 3} (H_t, H_t) + \mu_0 (\omega_{p3}^2 H, H) \right] (t) \\ &= \frac{1}{2} \left[\epsilon_0 \epsilon_{\infty 1} (\lambda_2 M_E^{-1} \mathbf{E}_t, \mathbf{E}_t) + \epsilon_0 (\lambda_2 \omega_{p1}^2 M_E^{-1} \mathbf{E}, \mathbf{E}) + \mu_0 \epsilon_{\infty 3} (H_t, H_t) + \mu_0 (\omega_{p3}^2 H, H) \right] (0) \\ & \quad + \int_0^t (M_E^{-1} M_F \mathbf{D}, \mathbf{E}_t) ds. \end{aligned} \quad (18)$$

Summing up (9), (13), and (18), we obtain

$$\begin{aligned}
 ENG(t) &= ENG(0) + \int_0^t \epsilon_0 \epsilon_{\infty 1} (\lambda_2 \mathbf{E}_{t2}, \mathbf{D}_t) ds + \int_0^t \epsilon_0 (\lambda_2 \omega_{p1}^2 \mathbf{E}, \mathbf{D}_t) ds \\
 &\quad - \int_0^t \mu_0 (\omega_{p3}^2 \mathbf{D}_t, \mathbf{E}_{t2}) ds + \int_0^t \mu_0 (H \nabla \times \omega_{p3}^2, \mathbf{E}_{t2}) ds \\
 &\quad + \int_0^t (M_E^{-1} M_F \mathbf{D}_t, \mathbf{E}_{t2}) ds + \int_0^t (M_E^{-1} M_F \mathbf{D}, \mathbf{E}_t) ds. \tag{19}
 \end{aligned}$$

Using the Cauchy-Schwarz inequality to all integral terms of (19), we easily see that they can be bounded by some corresponding terms in $ENG(t)$. Hence, the proof is completed by the Gronwall inequality. \square

3 The Discontinuous Galerkin Method

To discretize the cloaking model (3)–(6), we consider a shape-regular mesh T_h that partitions the domain Ω into disjoint triangular elements $\{T_i\}$, such that $\Omega = \bigcup_{i=1}^{N_T} T_i$. Furthermore, we denote $a_{ik} = T_i \cap T_k$ for an interior edge between two elements T_i and T_k , and \mathbf{n}_{ik} for the unit normal vector pointed from T_i to T_k . For any given element T_i , we denote v_i for the set of all neighboring elements of T_i .

In the DG method, the finite element space is given by discontinuous piecewise polynomials of degree k on each element, that is:

$$U_h = \{u_h \in L^2(\Omega) : u_h|_{T_i} \in P_k, \forall T_i \in T_h\}, \quad V_h = U_h \times U_h.$$

Moreover, we denote V_h^0 for the subspace of V_h satisfying the PEC boundary condition (7). To define a fully discrete scheme, we divide the time interval $(0, T)$ into N uniform subintervals by points $0 = t_0 < t_1 < \dots < t_N = T$, where $t_k = k\tau$, and $\tau = T/N$. For any function $u_h \in U_h$, we denote its average on any internal face a_{ik} as $\{\{u_h\}\}_{ik} = \frac{1}{2}(u_{hi} + u_{hk})$, where u_{hi} and u_{hk} denote the function values of u_h from the current element T_i and the neighboring element T_k , respectively.

Now, we can construct our fully discrete leap-frog type scheme: Given proper initial approximations of $H_h^0, B_h^0, \mathbf{D}_h^{-\frac{1}{2}}, \mathbf{D}_h^{-\frac{3}{2}}, \mathbf{E}_h^{-\frac{1}{2}}, \mathbf{E}_h^{-\frac{3}{2}}$, for any $n \geq 0$, find $H_h^{n+1}, B_h^{n+1} \in U_h, \mathbf{D}_h^{n+\frac{1}{2}}, \mathbf{E}_h^{n+\frac{1}{2}} \in V_h^0$ such that

$$\begin{aligned}
 &\int_{T_i} \frac{\mathbf{D}_h^{n+\frac{1}{2}} - \mathbf{D}_h^{n-\frac{1}{2}}}{\tau} \cdot \mathbf{v}_h - \int_{T_i} H_h^n \cdot \nabla \times \mathbf{v}_h - \sum_{T_k \in v_i} \int_{a_{ik}} \mathbf{v}_h \times \mathbf{n}_{ik} \cdot \{\{H_h^n\}\}_{ik} = 0, \tag{20} \\
 &\epsilon_0 \lambda_2 \left(\epsilon_{\infty 1} \int_{T_i} \frac{\mathbf{E}_h^{n+\frac{1}{2}} - 2\mathbf{E}_h^{n-\frac{1}{2}} + \mathbf{E}_h^{n-\frac{3}{2}}}{\tau^2} \cdot \mathbf{v}_h + \int_{T_i} \omega_{p1}^2 \mathbf{E}_h^{n+\frac{1}{2}} \cdot \mathbf{v}_h \right)
 \end{aligned}$$

$$= \int_{T_i} M_E \frac{\mathbf{D}_h^{n+\frac{1}{2}} - 2\mathbf{D}_h^{n-\frac{1}{2}} + \mathbf{D}_h^{n-\frac{3}{2}}}{\tau^2} \cdot \mathbf{v}_h + \int_{T_i} M_F \mathbf{D}_h^{n+\frac{1}{2}} \cdot \mathbf{v}_h, \quad (21)$$

$$\int_{T_i} \frac{B_h^{n+1} - B_h^n}{\tau} \cdot \psi_h + \int_{T_i} \mathbf{E}_h^{n+\frac{1}{2}} \cdot \nabla \times \psi_h + \sum_{T_k \in \nu_i} \int_{a_{ik}} \psi_h \cdot \mathbf{n}_{ik} \times \{\{\mathbf{E}_h^{n+\frac{1}{2}}\}\}_{ik} = 0, \quad (22)$$

$$\int_{T_i} \frac{B_h^{n+1} - 2B_h^n + B_h^{n-1}}{\tau^2} \psi_h$$

$$= \mu_0 \left(\epsilon_{\infty 3} \int_{T_i} \frac{H_h^{n+1} - 2H_h^n + H_h^{n-1}}{\tau^2} \psi_h + \int_{T_i} \omega_{p3}^2 H_h^{n+1} \psi_h \right), \quad (23)$$

hold true for any $\mathbf{v} \in \mathbf{V}_h^0$ and $\psi_h \in U_h$. Note that this scheme can be simply implemented as follows: At each time step, first solving (20) for $\mathbf{D}_h^{n+\frac{1}{2}}$, then solving (21) for $\mathbf{E}_h^{n+\frac{1}{2}}$, followed by solving B^{n+1} from (22), and finally solving for H_h^{n+1} from (23).

To simulate the wave propagation in unbounded domain, we have to truncate the unbounded domain to a bounded one by using the perfectly matched layer (PML) introduced by Berenger. Here, we use the unsplit PML:

$$\epsilon_0 \frac{\partial \mathbf{E}}{\partial t} + \sigma \mathbf{E} = \nabla \times H_z,$$

$$\mu_0 \frac{\partial H_z}{\partial t} + \sigma^* H_z = -\nabla \times \mathbf{E},$$

where the electric field $\mathbf{E} = (E_x, E_y)'$, and parameters σ and σ^* are the electric and magnetic conductivities, respectively. To avoid reflection from interfaces, we need the impedance-matching condition: $\frac{\sigma}{\epsilon_0} = \frac{\sigma^*}{\mu_0}$.

To couple the PML model with the cloaking model, we construct a similar DG scheme for solving the PML model as follows: Given proper approximations of $H_h^0, \mathbf{E}_h^{-\frac{1}{2}}$, for any $n \geq 0$, find $\mathbf{E}_h^{n+\frac{1}{2}} \in \mathbf{V}_h^0, H_h^{n+1} \in U_h$ such that

$$\epsilon_0 \int_{T_i} \frac{\mathbf{E}_h^{n+\frac{1}{2}} - \mathbf{E}_h^{n-\frac{1}{2}}}{\tau} \cdot \mathbf{v}_h - \int_{T_i} H_h^n \cdot \nabla \times \mathbf{v}_h - \sum_{T_k \in \nu_i} \int_{a_{ik}} \mathbf{v}_h \times \mathbf{n}_{ik} \cdot \{\{H_h^n\}\}_{ik}$$

$$+ \int_{T_i} \sigma \cdot \frac{\mathbf{E}_h^{n+\frac{1}{2}} + \mathbf{E}_h^{n-\frac{1}{2}}}{2} \times \mathbf{v}_h = 0,$$

$$\mu_0 \int_{T_i} \frac{H_h^{n+1} - H_h^n}{\tau} \cdot u_h + \int_{T_i} \mathbf{E}_h^{n+\frac{1}{2}} \cdot \nabla \times \psi_h + \sum_{T_k \in \nu_i} \int_{a_{ik}} \psi_h \cdot \mathbf{n}_{ik} \times \{\{\mathbf{E}_h^{n+\frac{1}{2}}\}\}_{ik}$$

$$+ \int_{T_i} \sigma^* \cdot \frac{H_h^{n+1} + H_h^n}{2} \cdot \psi_h = 0,$$

hold true for any $\mathbf{v}_h \in \mathbf{V}_h^0$ and $\psi_h \in U_h$.

4 Numerical Results

To validate our elliptic cloaking model, we implemented the finite element method (20)–(23) and carried out extensive numerical tests. In all our simulations, we fixed the physical domain $\Omega = [0, 0.8]m \times [0, 0.8]m$, which is partitioned by unstructured triangular meshes with mesh size $h = 6.25 \times 10^{-3}m$, and the time domain $[0, 8]ns$ discretized by a time step size $\tau = 2 \times 10^{-13}s$, i.e., the total number of time steps is 40000. For our cloaking simulation, the total number of triangular elements is about 33000.

On both left and right sides of the PML region, the damping function σ^* is a function of x given as follows:

$$\sigma^*(x) = \begin{cases} \sigma_{max} \left(\frac{x-0.8}{dd} \right)^6, & \text{if } x \geq 0.8 \\ \sigma_{max} \left(\frac{x}{dd} \right)^6, & \text{if } x \leq 0 \\ 0, & \text{otherwise,} \end{cases}$$

where $\sigma_{max} = -\log(err) * (6 + 1) * 0.8 * c_v / (2 * dd)$ with $err = 10^{-7}$. While in the y -direction, $\sigma^*(y)$ has the same form. In our tests, we used $dd = 12h$ in all directions.

Example 1 In this example, we choose the cloaking parameters $a = 0.075$, $b = 0.15$, $k = 2$, $\epsilon_{\infty 1} = 1$, and $\epsilon_{\infty 3} = 2$. The incident wave is generated by a plane wave $H_z = 0.1 \sin(\omega t)$ imposed at the line segment connected by points $(6.25 \times 10^{-3}, 0.01)m$ and $(6.25 \times 10^{-3}, 0.79)m$, where $\omega = 2\pi f$ with frequency $f = 2$ GHz. The computed electric fields E_y obtained at various time steps are presented in Figure 2, which shows that the plane wave pattern is resumed quite well after the wave passes the cloaking region, i.e., the invisibility cloaking phenomenon is achieved very well.

Example 2 This example has the same physical parameters as Example 1, except that the incident source wave $H_z = 0.1 \sin(\omega t)$ is imposed at point $(6.25 \times 10^{-3}, 0.4)m$. Snapshots of obtained E_y are presented in Figure 3, which clearly shows that the invisibility cloaking phenomenon is also achieved very well in this case.

Example 3 This example has the same parameters as Example 2 except that $a = 0.15$, $b = 0.3$, and $k = 1/2$, i.e., the ellipse has the major axis lying on the y -axis. We tested both line source wave and point source wave, which show invisibility cloaking very well. Some snapshots of the electric fields E_y obtained with the point wave source is given in Figure 4.

Example 4 This example has the same setup as Example 3 except that $k = 1$, i.e., we have a circular cloak with inner radius $a = 0.15$ and outer radius $b = 0.3$. Both line source wave and point source wave are tested and both show invisibility

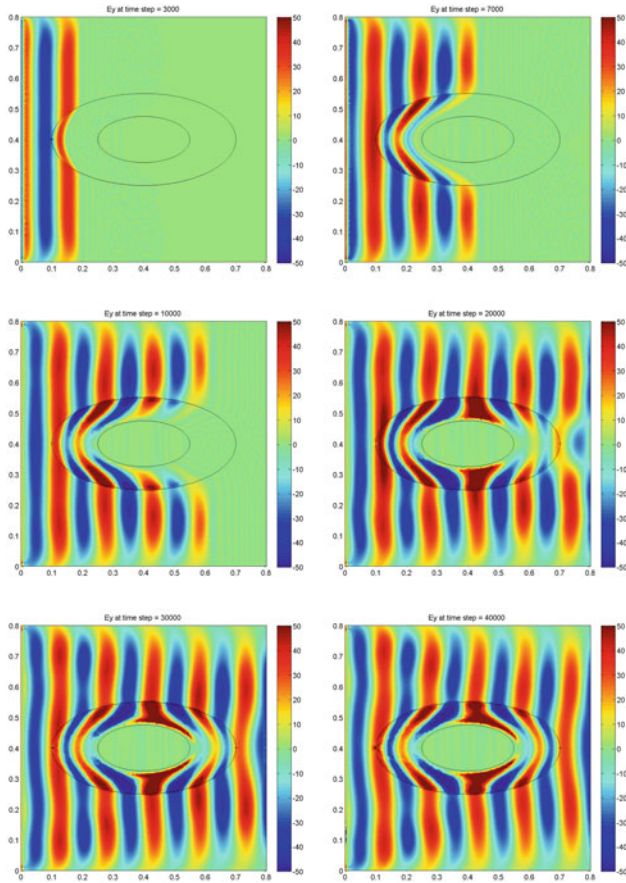


Fig. 2 Example 1. Snapshots of electric fields E_y obtained with a plane wave source for the ellipse with parameters $a = 0.075$, $b = 0.15$, $k = 2$: (top left) 3,000; (top right) 7,000; (middle left) 10,000; (middle right) 20,000; (bottom left) 30,000; (bottom right) 40,000.

cloaking very well. Some snapshots of the electric fields E_y obtained with the point wave source is presented in Figure 5.

Acknowledgements Work of the authors “Yunqing Huang and Chen Meng” was supported by the NSFC Key Project 91430213. Work of the author “Jichun Li” was supported by the NSF grant DMS-1416742 and NSFC project 11671340.

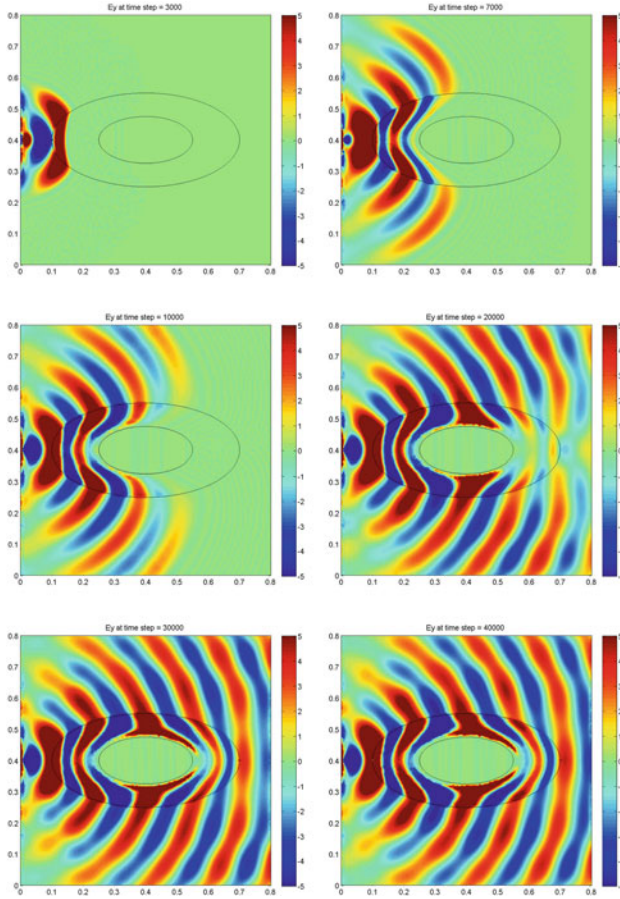


Fig. 3 Example 2. Snapshots of electric fields E_y obtained with a point wave source for the ellipse with parameters $a = 0.075$, $b = 0.15$, and $k = 2$: (top left) 3,000; (top right) 7,000; (middle left) 10,000; (middle right) 20,000; (bottom left) 30,000; (bottom right) 40,000.

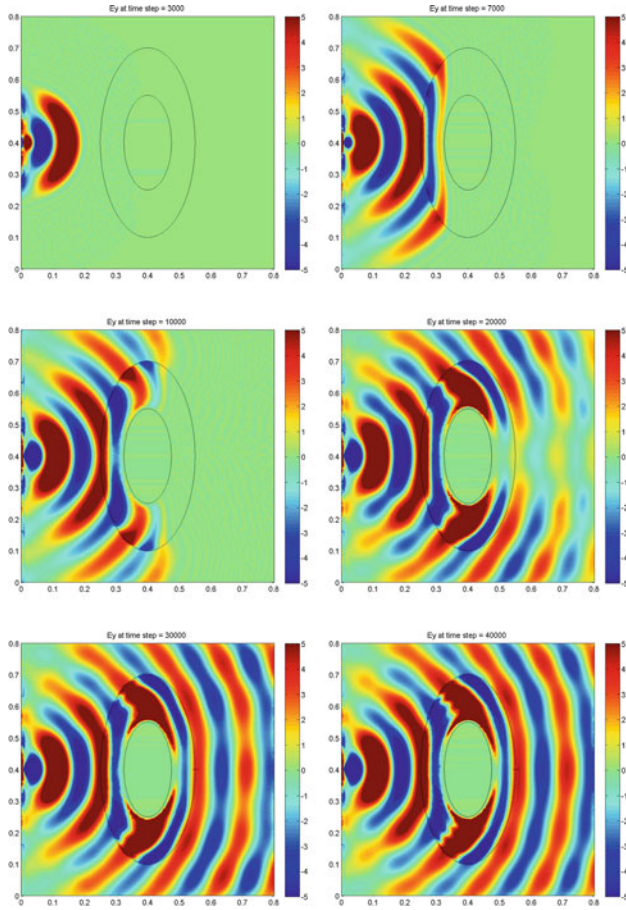


Fig. 4 Example 3. Snapshots of fields E_y obtained with a point wave source for the ellipse with parameters $a = 0.15$, $b = 0.3$, and $k = 1/2$: (top left) 3,000; (top right) 7,000; (middle left) 10,000; (middle right) 20,000; (bottom left) 30,000; (bottom right) 40,000.

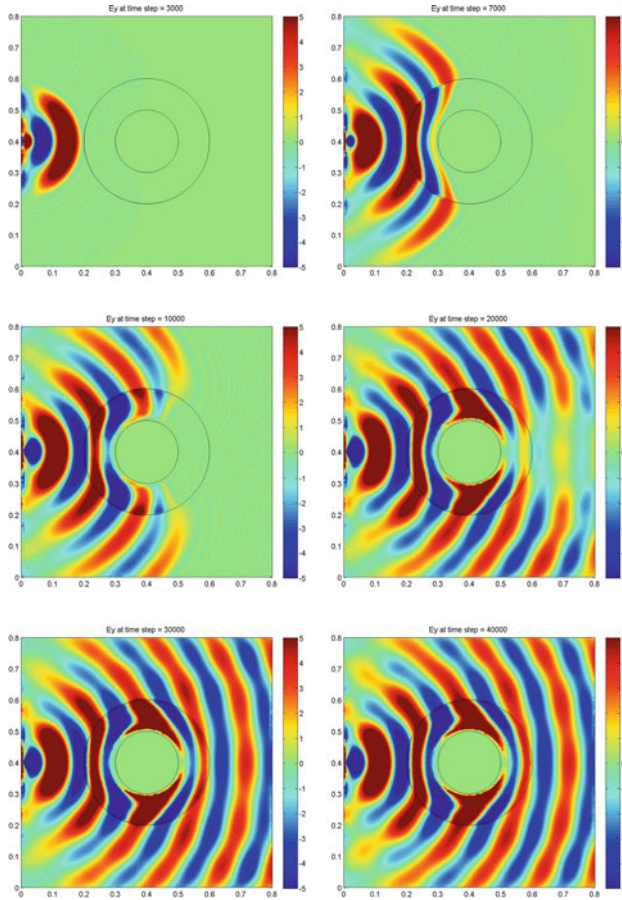


Fig. 5 Example 4. Snapshots of fields E_y obtained with a point wave source for a circle with $a = 0.15$ and $b = 0.3$: (top left) 3,000; (top right) 7,000; (middle left) 10,000; (middle right) 20,000; (bottom left) 30,000; (bottom right) 40,000.

References

1. H. Ammari, G. Ciraolo, H. Kang, H. Lee and G. Milton, Spectral theory of a Neumann-Poincaré-type operator and analysis of cloaking due to anomalous localized resonance, *Arch. Ration. Mech. Anal.* 208 (2013) 667–692.
2. H. Ammari, H. Kang, H. Lee, M. Lim and S. Yu, Enhancement of near cloaking for the full Maxwell equations, *SIAM J. Appl. Math.* 73 (2013) 2055–2076.
3. G. Bao, H. Liu and J. Zou, Nearly cloaking the full Maxwell equations: cloaking active contents with general conducting layers, *J. Math. Pures et Appl.* 101(5) (2014) 716–733.
4. A.S. Bonnet-Ben Dhia, L. Chesnel and P. Ciarlet Jr., Two-dimensional Maxwells equations with sign-changing coefficients, *Appl. Numer. Math.* 79 (2014) 29–41.

5. S.C. Brenner, J. Gedicke and L.-Y. Sung, An adaptive P_1 finite element method for two-dimensional transverse magnetic time harmonic Maxwell's equations with general material properties and general boundary conditions, *J. Sci. Comput.* 68(2) (2016) 848–863.
6. L. Demkowicz and J. Li, Numerical simulations of cloaking problems using a DPG method, *Comput. Mech.* 51 (2013) 661–672.
7. A. Greenleaf, Y. Kurylev, M. Lassas and G. Uhlmann, Cloaking devices, electromagnetics wormholes and transformation optics, *SIAM Review* 51 (2009) 3–33.
8. F. Guevara Vasquez, G.W. Milton and D. Onofrei, Broadband exterior cloaking, *Opt. Express* 17 (2009) 14800–14805.
9. Y. Hao and R. Mittra, *FDTD Modeling of Metamaterials: Theory and Applications*, Artech House Publishers, 2008.
10. Y. Huang and J. Li, Total reflection and cloaking by triangular defects embedded in zero index metamaterials, *Adv. Appl. Math. Mech.* 7(2) (2015) 1–10.
11. W.X. Jiang, T.J. Cui, G.X. Yu, X.Q. Lin, Q. Cheng, J.Y. Chin, Arbitrarily elliptical-cylindrical invisible cloaking, *J. Phys. D: Appl. Phys.* 41 (2008), 085504.
12. R.V. Kohn, D. Onofrei, M.S. Vogelius and M.I. Weinstein, Cloaking via change of variables for the Helmholtz equation, *Comm. Pure Appl. Math.* 63 (2010) 973–1016.
13. M. Lassas, M. Salo and L. Tzou, Inverse problems and invisibility cloaking for FEM models and resistor networks, *Math. Mod. Meth. Appl. Sci.* 25(2) (2015) 309–342.
14. U. Leonhardt, Optical conformal mapping, *Science* 312 (2006) 1777–1780.
15. U. Leonhardt and T. Tyc, Broadband invisibility by non-Euclidean cloaking, *Science* 323 (2009) 110–112.
16. J. Li, Well-posedness study for a time-domain spherical cloaking model, *Comput. Math. Appl.* 68 (2014) 1871–1881.
17. J. Li and Y. Huang, Mathematical simulation of cloaking metamaterial structures, *Adv. Appl. Math. Mech.* 4 (2012) 93–101.
18. J. Li and Y. Huang, *Time-Domain Finite Element Methods for Maxwell's Equations in Metamaterials*, Springer Series in Computational Mathematics, vol.43, Springer, 2013.
19. J. Li, Y. Huang and W. Yang, Developing a time-domain finite-element method for modeling of electromagnetic cylindrical cloaks, *J. Comp. Phys.* 231 (2012) 2880–2891.
20. J. Li, Y. Huang and W. Yang, An adaptive edge finite element method for electromagnetic cloaking simulation, *J. Comp. Phys.* 249 (2013) 216–232.
21. J. Li, Y. Huang and W. Yang, Well-posedness study and finite element simulation of time-domain cylindrical and elliptical cloaks, *Math. Comp.* 84 (2015) 543–562.
22. J. Li, Y. Huang, W. Yang and A. Wood, Mathematical analysis and time-domain finite element simulation of carpet cloak, *SIAM J. Appl. Math.* 74(4) (2014) 1136–1151.
23. W. Li, D. Liang and Y. Lin, A new energy-conserved S-FDTD scheme for Maxwell's equations in metamaterials, *Int. J. Numer. Anal. Mod.* 10 (2013) 775–794.
24. R. Liu, C. Ji, J.J. Mock, J.Y. Chin, T.J. Cui and D.R. Smith, Broadband ground-plane cloak, *Science* 323 (2009) 366–369.
25. S. Nicaise and J. Venel, A posteriori error estimates for a finite element approximation of transmission problems with sign changing coefficients, *J. Comput. Appl. Math.* 235 (2011) 4272–4282.
26. J.B. Pendry, D. Schurig and D.R. Smith, Controlling electromagnetic fields, *Science* 312 (2006) 1780–1782.
27. D.H. Werner and D.-H. Kwon (eds.), *Transformation Electromagnetics and Metamaterials: Fundamental Principles and Applications*, Springer, 2013.
28. Y. Wu and J. Li, Total reflection and cloaking by zero index metamaterials loaded with rectangular dielectric defects, *Applied Physics Letters* 102, 183105 (2013), 4 pages.
29. Z. Xie, J. Wang, B. Wang and C. Chen, Solving Maxwell's equation in meta-materials by a CG-DG method, *Commun. Comput. Phys.* 19(5) (2016) 1242–1264.

30. W. Yang, J. Li and Y. Huang, Mathematical analysis and finite element time domain simulation of arbitrary star-shaped electromagnetic cloaks, *SIAM J. Numer. Anal.* 56(1) (2018) 136–159.
31. Z. Yang and L.L. Wang, Accurate simulation of ideal circular and elliptic cylindrical invisibility cloaks, *Commun. Comput. Phys.* 17(3) (2015) 822–849.
32. Z. Yang, L.-L. Wang, Z. Rong, B. Wang and B. Zhang, Seamless integration of global Dirichlet-to-Neumann boundary condition and spectral elements for transformation electromagnetics, *Comput. Methods Appl. Mech. Engrg.* 301 (2016) 137–163.

Dynamic Pore-Network Models Development



X. Yin, E. T. de Vries, A. Raouf, and S. M. Hassanizadeh

1 Introduction

Along with the development of imaging techniques and micro-fluidics experiments, pore-network model has evolved over decades to represent more efficient and complex structures with more pore-scale mechanisms included.

Pore-network can be categorized as quasi-static or dynamic. Quasi-static models have been successfully applied for the prediction of relative permeabilities, and capillary pressure-saturation in multiphase system [18–20]. Dynamic pore-network model applies when capillary-dominance assumption is no longer valid and viscous forces need to be considered together with the capillary forces. Dynamic pore-network models mainly fall into two groups: single-pressure algorithm and two-pressure algorithm. We discussed several representative already developed pore-network models below and comprehensive reviews can be found in [1, 7, 10].

Blunt and King [6] simulated two-phase flow using isotropic pore network to study invasion fractal dimension and dynamic relative permeability. In their model, throats were completely filled with only fluid, while pores may contain two fluids simultaneously. Capillary pressure in pores was neglected.

van der Marck et al. [15] studied drainage by means of experiments and simulation methods. They extended the numerical pore-network model by Lenormand et al. [14]. They compared pressure buildup at boundary and saturation in the domain before breakthrough during flow-controlled drainage process. Pressure

X. Yin (✉) · E. T. de Vries · A. Raouf · S. M. Hassanizadeh
Department of Earth Sciences, Utrecht University, Utrecht, The Netherlands
e-mail: x.yin@uu.nl; e.t.devries@uu.nl; a.raouf@uu.nl; s.m.hassanizadeh@uu.nl

is defined at pores and pores could be occupied by two fluids. Only capillary pressure at channels was considered and capillary pressure at pores was neglected. The algorithm mainly involved pressure solving and saturation update.

Aker et al. [2] modeled the dynamics of drainage-dominated flow using a two-dimensional network. The flow front width from simulation was found to be consistent with a scaling relation and scaling exponents were compared with experimental data from the literature. Pressure was defined at nodes and only one fluid would occupy a node. Main terminal meniscus is modeled, while film flow and corner flow were not considered. Simultaneous flow of two liquids into one tube was allowed with maximum two menisci. Solving volume flux conservation at nodes provided pressure field. Rules were defined to describe how a meniscus will move into neighboring tubes when it reaches the end of a tube.

Al-Gharbi and Blunt [4] presented a dynamic network model for modeling of two-phase drainage. Wetting layer flow, meniscus oscillation, and the dynamics of snap-off are accounted for. In their model, pressure is assigned at pore centers and throat centers. Volume conservation equations for pores and throats determine the pressure field. In writing pressure drop between neighboring pore and throat, the number and orientation of interface are included. After solving pressure, location of interface can be updated by iteration. Rules regarding invasion of pore center and fusion of interface are defined to ensure relatively simple track of number and orientation of interface.

These single-pressure algorithms may be computationally efficient; however, they could suffer from one of the followings: (1) network with regular geometry/topology; (2) failure to include corner flow/film flow; (3) ignorance of capillary pressure at pores; (4) rule-based description of interface morphology; and (5) rule-based dynamic algorithm.

Using a two-pressure algorithm, both pore body and pore throat can be filled with different fluids, each with its own pressure. Interface position or shape is not explicitly described but included in pore body capillary pressure-saturation relationship. Pressure, fluxes, and saturations are calculated separately for each fluid. This algorithm was initially developed by Thompson [24] for study of imbibition process in fibrous materials. Later, this algorithm was improved and implemented onto structured network with fixed coordination number of six to study theories of two-phase flow in porous media [11–13] and wicking behavior [9].

In this work, we present the development of the two-pressure algorithm for unstructured network allowing for coordination number ranging from 1 to 26 introduced by Raouf and Hassanzadeh [23]. For two-phase flow involving two fluids with significant different viscosities, such as the water-air system, pressure drop in less viscous fluid may be negligible, and the above two-pressure algorithm may be simplified accordingly. Single-pressure algorithms for both drainage and imbibition are developed in this study.

We give description of our algorithms in Section 2 and simulations are carried out in Section 3.

2 Pore-Network Model Description

2.1 Model Features

2.1.1 Structure and Geometry

Pore bodies are cubic and pore throats have square cross sections. With these shapes of pore bodies and pore throats, we need to have inscribed radii of pore throats, inscribed radii of pore bodies, length of pore throats, and coordination number. We can either generate the unstructured network based on some distributions, such as log-normal distribution, and tune geometrical and topological data to match the available experimental measurements when necessary or we can model random geometry and topology directly using information from network extraction methods, like Avizo (FEI Visualization Sciences Group). Our algorithms can handle these two kinds of networks. Here, we will generate the unstructured network described in [23]. Figure 1 provides the distribution of inscribed radii sizes of pore throats, inscribed radii sizes of pore bodies, length sizes of pore throats, and coordination number.

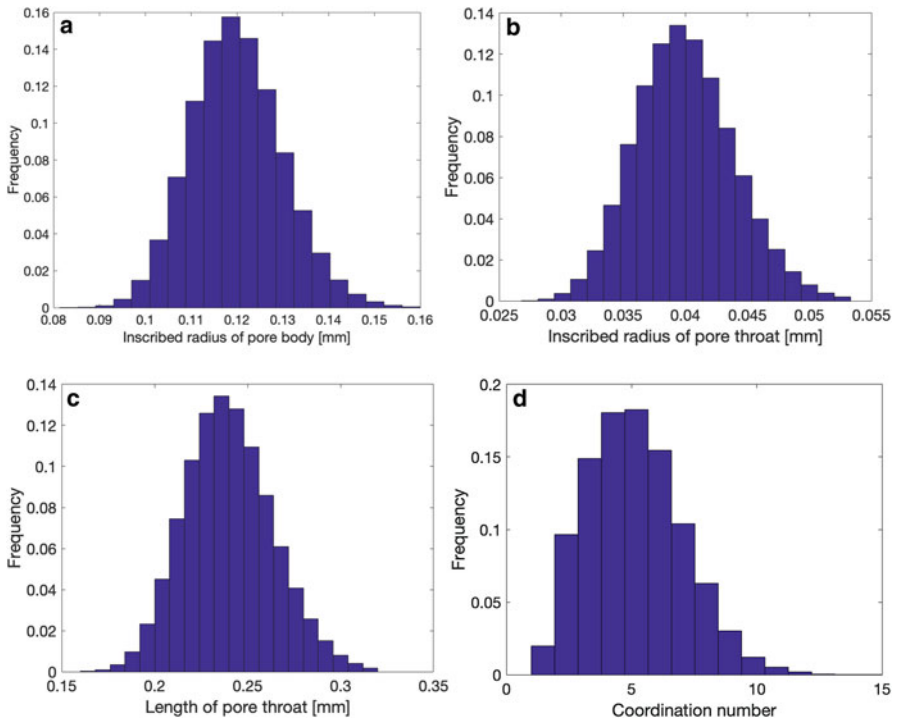


Fig. 1 Size distribution of: (a) inscribed radii of pore throats, (b) pore throats length, (c) inscribed radii of pore bodies, and (d) coordination number

2.1.2 Assumptions

Assumptions employed in the pore-network algorithm are as follows:

1. The volume of pore throats are assumed to be negligible compared to the volume of pore bodies. This means that the filling of a pore throat is assumed to occur in no time.
2. No hydraulic resistance is assigned to pore bodies; their resistance to the flow is assumed to be negligible compared to that of pore throats.
3. No gravity effects is included.
4. Flow of the wetting phase through corners of pore elements is taken into account. Therefore, any pore body or pore throat can be simultaneously occupied by both wetting phase and non-wetting phase.

2.1.3 System Parameters

Table 1 gives the fluid properties used in the simulations. Three values of non-wetting phase viscosities are adopted to have 3 viscosity ratios of $M(= \frac{\mu^n}{\mu^w})$: 10, 1, and 0.1.

2.2 Local Rules

2.2.1 Capillary Pressure for Pore Bodies

Assuming that the wetting phase resides symmetrically in all eight corners of pore bodies, capillary pressure-saturation relationship for cubic pore bodies can be defined as [13]:

$$p_i^c(s_i^w) = p_i^n - p_i^w = \frac{2\sigma^{wn} \cos\theta}{R_i(1 - \exp(-6.83s_i^w))} \quad (1)$$

where p_i^c is capillary pressure, p_i^α and s_i^α denote pressure and saturation of α phase in pore body i , and R_i is the radius of inscribed sphere of pore body i . Here, we

Table 1 Material properties

Specification	Symbol	Value	Unit
Contact angle	θ	0	degree
Interfacial tension	σ^{wn}	0.072	kg s ⁻²
Wetting fluid viscosity	μ^w	0.001	kg m ⁻¹ s ⁻¹
Non-wetting fluid viscosity	μ^n	0.01 or 0.001 or 0.0001	kg m ⁻¹ s ⁻¹

have not considered the possibility for different interface shapes in a pore body due to different filling states of its neighbors [14].

2.2.2 Entry Capillary Pressure for Pore Throats

Entry capillary pressure for a square pore throat is defined as [16, 17, 21]:

$$P_{ij}^{en} = \frac{\sigma^{wn}}{r_{ij}} \left(\frac{\theta + \cos^2\theta - \pi/4 - \sin\theta \cos\theta}{\cos\theta - \sqrt{\pi/4 - \theta + \sin\theta \cos\theta}} \right) \quad (2)$$

where r_{ij} is the radius of inscribed circle of pore throat cross section.

2.2.3 Minimum Wetting Phase Saturation in a Pore Body

Due to existence of corners, during drainage wetting phase will not be completely displaced from pore bodies. The minimum wetting phase saturation of each pore body is defined according to the global pressure difference, namely:

$$P_{global}^c = P_{inlet}^n - P_{outlet}^w \quad (3)$$

$$s_{i,min}^w = -\frac{1}{6.83} \ln\left(1 - \frac{1}{R_i} \frac{2\sigma^{wn} \cos\theta}{P_{global}^c}\right) \quad (4)$$

2.2.4 Invasion Criteria and Trapping

Drainage When non-wetting phase pressure in a pore body exceeds the entry capillary pressure of its neighboring pore throats, non-wetting phase will invade the pore throat.

Single-Phase Imbibition From one time step to another, a pore throat will be assumed to get invaded by the wetting phase only if at least one of its neighboring pore bodies has reached a wetting phase saturation of 0.477 (corresponding to the case where the non-wetting phase filling the inscribed sphere of the pore body). At first, only corners are assumed to become filled. The radius of the meniscus formed in the corner depends on the wetting phase pressure, and is given by Equation (9) below.

If the wetting phase pressure in one of the neighboring pore bodies is high enough, the whole pore cross section will be invaded by the wetting phase. So, the criteria for the full invasion of a pore throat is $p_{ij}^c < p_{ij}^{en}$ by wetting phase in unsaturated neighboring pore body. For wetting phase saturated pore body, wetting phase will invade the narrower pore throat through main terminal meniscus.

However, even if one or both invasion criteria, which were just described, are met, a pore throat will not be invaded if it is considered to be trapped [3]. Details about different scenarios of invasion and trapping can be found in Appendix A.

In general, when a pore throat is not considered to be trapped, it can always be invaded by the wetting phase (independent of the applied boundary pressure) since corner flow can always occur. However, we have imposed the requirement that the saturation of one of the neighboring pore bodies must exceed 0.477.

2.2.5 Conductivities of Pore Throats

The conductivities of phases through pore throats depend on fluids occupancy in the pore throat cross section. In general, we may have the following states:

1. Pore throat is fully occupied by wetting phase. Its conductivity is given by Azzam and Dullien [5].

$$K_{ij}^w = \frac{\pi}{8\mu^w l_{ij}} r_{ij}^{eff4} \quad (5)$$

where l_{ij} is length of pore throat and

$$r_{ij}^{eff} = \sqrt{\frac{\pi}{4}} r_{ij} \quad (6)$$

2. Pore throat is occupied by wetting phase in the corner, while non-wetting phase is in the middle [22].

$$K_{ij}^w = \frac{4 - \pi}{\beta \mu^w l_{ij}} r_{ij}^c{}^4 \quad (7)$$

$$K_{ij}^n = \frac{\pi}{8\mu^n l_{ij}} r_{ij}^{eff4} \quad (8)$$

where

$$r_{ij}^c = \frac{\sigma^{wn}}{p_{ij}^c} \left(\frac{\theta + \cos^2\theta - \pi/4 - \sin\theta \cos\theta}{\cos\theta - \sqrt{\pi/4 - \theta + \sin\theta \cos\theta}} \right) \quad (9)$$

$$r_{ij}^{eff} = \frac{1}{2} \left(\sqrt{\frac{r_{ij}^c{}^2 - (4 - \pi)r_{ij}^c{}^2}{\pi}} + r_{ij} \right) \quad (10)$$

Dimensionless resistance β is given by Zhou et al. [26]. For single-pressure algorithm, only wetting phase conductivity will be used.

2.2.6 Snap-Off

Snap-off may happen in a pore throat when capillary pressure in the pore throat drops below a threshold value so that stable corner interface is not supported any more. For square cross sectioned pore throat, ignoring dynamic contact angle effects, the criterion on snap-off is defined as [25]:

$$p_{ij}^c \leq \frac{\sigma^{wn}}{r_{ij}} (\cos\theta - \sin\theta) \quad (11)$$

Once snap-off happens in a pore throat, it will be fully occupied by the wetting phase and its conductivity will be changed accordingly.

2.3 Governing Equations

2.3.1 Governing Equations for Two-Pressure Drainage

For two-pressure algorithm, fluid volume balance equation in pore bodies can be written resulting in linear system of equations:

$$\sum_{j=1}^{N_i} [(K_{ij}^w + K_{ij}^n)p_i^w - (K_{ij}^w + K_{ij}^n)p_j^w] = - \sum_{j=1}^{N_i} K_{ij}^n (p_i^c - p_j^c) \quad (12)$$

where N_i is coordination number of pore body i .

Saturation Update After solving pressure field, saturation can be updated explicitly:

$$V_i \frac{\Delta s_i^w}{\Delta t} = -q_i^w = - \sum_{j=1}^{N_i} K_{ij}^w (p_i^w - p_j^w) \quad (13)$$

where V_i is volume of pore body i , Δt is time step, and q_i^w is total wetting phase flux of pore body i .

2.3.2 Governing Equations for Single-Pressure Imbibition

A single-pressure primary imbibition algorithm is developed for the case where a viscous fluid is the wetting phase and a much less viscous fluid (like air) is the non-wetting phase. Wetting phase pressure will be calculated, while the non-wetting phase is assumed to be at a constant and uniform pressure at all times. This is a valid assumption given the negligible viscosity of the non-wetting phase.

Local (i.e., the pore level) capillary pressure for a given pore body i is defined in Section 2.2. Initially, all pore bodies are assumed to have an initial wetting phase saturation (we set $s_{init}^w = 0.5\%$) except the inlet pore bodies which remain saturated at all times. Those boundary pore bodies are assumed to be fully saturated. Thus, initially all internal pore bodies have a large negative wetting phase pressure given by $p_i^c(s_i^w)$ relationship (we have imposed a maximum of p_i^c of 10^6 Pa). Saturated boundary pores are assigned the same pressure as inlet reservoir.

Infiltration starts by the wetting phase flowing from the saturated boundary pores into the internal pores. Their saturation, and thus their pressure rise and, subsequently, wetting phase can flow into the neighboring pores. The wetting phase flow occurs via pore throats and its rate can be calculated using Hagen-Poiseuille formula:

$$Q_{ij}^w = K_{ij}^w (p_i^w - p_j^w) \quad (14)$$

The saturation update for each pore body is made based on volume balance equation, the same as in Equation (13). The calculation is done fully explicitly. Note that the saturation update needs to be done for active unsaturated pore bodies, those for which at least one pore throat has been invaded by wetting phase. As long as no internal pore body has become fully saturated with wetting phase, the updating of pore body saturation and pressure, invasion of new pore throats and the next cycle of flow calculation, and the updating of saturation and pressure continue. However, at each step, we check whether a given pore throat can be invaded by the wetting phase, following the criteria discussed in Section 2.2.

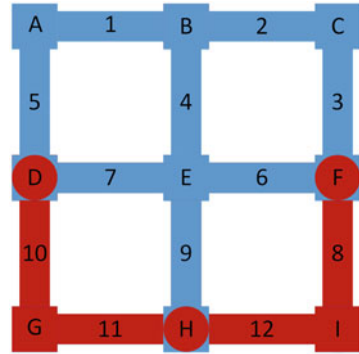
As soon as one or more internal pore bodies become fully saturated with the wetting phase, then we have to solve for the pressure of those pore bodies based on the following volume balance equation:

$$\sum_{j=1}^{N_i} K_{ij}^w (p_i^w - p_j^w) = 0 \quad (15)$$

Here, K_{ij}^w is calculated explicitly. The domain of saturated pore bodies is surrounded by unsaturated pore bodies. The pressure of unsaturated pore bodies (known from the current time step) is used as boundary condition values for the domain of saturated pore bodies. Once the pressure of all saturated pore bodies are calculated, steps described above for updating saturation of unsaturated pore bodies will be repeated.

Figure 2 gives the schematic diagram of the domain with boundary pores and internal saturated and unsaturated pores, where blue color shows the wetting phase and red color shows the non-wetting phase. In this figure, pore bodies are denoted from A to I and pore throats are denoted from 1 to 12. A, B, C are boundary pore bodies, E is internal saturated pore body, D, F, H are internal unsaturated pore bodies, and G, I are pore bodies that are not invaded yet. No initial tiny saturation is shown in the figure and no attempt is made to represent the real interface shape.

Fig. 2 Schematic diagram of boundary pores, internal unsaturated pores, and saturated pores in single-pressure imbibition algorithm



2.3.3 Governing Equations for Single-Pressure Drainage

Single-pressure algorithm is developed for drainage process, where non-wetting phase (like air) with a much lower viscosity value relative to the wetting fluid is assumed to be at a constant and uniform pressure distribution.

In this single-pressure algorithm, mass balance equation can be written resulting in linear system of equations:

$$V_i \frac{\Delta s_i^w}{\Delta t} = - \sum_{j=1}^{N_i} Q_{ij}^w \tag{16}$$

$$V_i \frac{ds_i^w}{dp_i^w} \frac{\Delta p_i^w}{\Delta t} = - \sum_{j=1}^{N_i} Q_{ij}^w \tag{17}$$

$$V_i \frac{ds_i^w}{dp_i^w} \frac{(p_i^w)^{k+1} - (p_i^w)^k}{\Delta t} = - \sum_{j=1}^{N_i} K_{ij}^w ((p_i^w)^{k+1} - (p_j^w)^{k+1}) \tag{18}$$

$$\left(\frac{V_i}{\Delta t} \frac{ds_i^w}{dp_i^w} + \sum_{j=1}^{N_i} K_{ij}^w \right) (p_i^w)^{k+1} - \sum_{j=1}^{N_i} K_{ij}^w (p_j^w)^{k+1} = \frac{V_i}{\Delta t} \frac{ds_i^w}{dp_i^w} (p_i^w)^k \tag{19}$$

For saturated pores, left-hand side of equations are zero.

Due to constant and uniform non-wetting phase pressure assumption, we have

$$\frac{ds_i^w}{dp_i^w} = - \frac{ds_i^w}{dp_i^c} \tag{20}$$

According to capillary pressure-saturation relationship defined in Section 2.2, above mass balance equations also hold for pore bodies with $s_i^w = 1$.

Saturation Update After solving pressure field, wetting phase flux of pore bodies can be calculated and saturation of pore bodies will be updated the same as in Equation (13).

2.4 Time Step

2.4.1 Imbibition

During imbibition, time step Δt is taken to be equal to the smallest filling time Δt_i of all active unsaturated pore bodies, but it is possible to have local drainage in some pore bodies. Allowing pore bodies to be infiltrated to $s_i^w = 1$ or drained to $s_i^w = 0$. Δt_i is determined as:

$$\Delta t_i = \begin{cases} \frac{V_i}{q_i^w} (s_i^w) & q_i^w > 0 \\ \frac{V_i}{q_i^w} (1 - s_i^w) & q_i^w < 0 \end{cases} \quad (21)$$

A global time step is then selected as the minimum time step of all pore bodies:

$$\Delta t = \min\{\Delta t_i\} \quad (22)$$

A saturation truncation value of 10^{-3} is adopted to ensure finite time step. When local saturation in a pore body is close to the target saturation values within the truncation value, that pore body will not be included in global time step determination.

2.4.2 Drainage

Time step is determined based on filling or emptying of pore bodies:

$$\Delta t_i = \begin{cases} \frac{V_i}{q_i^w} (s_i^w - s_{i,\min}^w) & q_i^w > 0 \\ \frac{V_i}{q_i^w} (1 - s_i^w) & q_i^w < 0 \end{cases} \quad (23)$$

The global time step is selected as the minimum time step of all pore bodies:

$$\Delta t = C_{oe} \min(\Delta t_i) \quad (24)$$

And, a saturation truncation of 10^{-3} is adopted to ensure finite time step. When local saturation in a pore body is close to the limits, that pore body will not involve in global time step determination. A time step coefficient C_{oe} smaller than 1 is adopted. Discussion on this coefficient can be found in Section 3.3

2.5 *Computational Procedure*

2.5.1 *Single-Phase Imbibition*

The procedure for dynamic primary imbibition simulation is:

1. Set boundary condition and initial condition for the pore network;
2. Determine conductivity of pore throats based on fluid occupancy as well as trapping and invasion criteria;
3. Solve wetting phase volume balance equations for the saturated pore bodies and get pressure field;
4. Calculate flux based on conductivity determined in step 2 and determine time step;
5. Update saturation and pressure of unsaturated active pore bodies;
6. Go to Step 2 and repeat the process.

2.5.2 *Drainage*

The procedure for dynamic drainage simulation is:

1. Set boundary condition and initial condition for the pore network;
2. Determine conductivity of pore throats based on invasion criteria;
3. Solve mass balance equations to get the pressure field;
4. Calculate flux, determine time step, and update saturation of pore bodies;
5. Go to step 2 and repeat the process.

3 *Simulations and Discussions*

3.1 *Boundary Conditions*

For 3D network, we assume that the network is connected with a wetting reservoir at one end and with a non-wetting phase reservoir at the other end. During drainage, the face connected with the non-wetting phase reservoir is considered as inlet and the other face is considered as the outlet; during imbibition, the place of inlet and outlet are exchanged.

3.2 *Averaging Procedure*

Pore-network simulations provide pore-scale information including pressure and saturation. To upscale these quantities to the macroscopic quantities, we need to define averaging operators for these variables.

$$S^\alpha = \frac{\sum_{i=1}^{N_b} s_i^\alpha V_i}{\sum_{i=1}^{N_b} V_i}, \alpha = n, w \tag{25}$$

$$P^\alpha = \frac{\sum_{i=1}^{N_b} p_i^\alpha s_i^\alpha V_i}{\sum_{i=1}^{N_b} s_i^\alpha V_i}, \alpha = n, w \tag{26}$$

$$P^c = \frac{\sum_{i=1}^{N_b} p_i^c A_i^{wn}}{\sum_{i=1}^{N_b} A_i^{wn}} \tag{27}$$

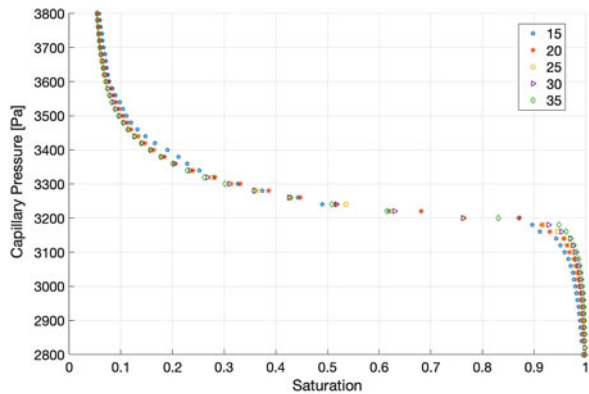
Here, pore-scale pressure is averaged over volume to obtain the macroscopic pressure, and capillary pressure is averaged over interfacial area to provide the macroscopic capillary pressure. Interfacial area in cubic pore bodies is given in Appendix B.

3.3 Two-Pressure Drainage Simulations

3.3.1 REV Size

Figure 3 shows quasi-static drainage simulations for different network sizes with the same statistical parameters provided in Figure 1. It is clear that a network size of 25 can be considered as REV (Representative Elementary Volume).

Fig. 3 Quasi-static $P^c - S^w$ curves for different network sizes



3.3.2 Time Step Independence

In dynamic pore-network models, time step is usually determined based on filling/emptying of pore bodies, in a manner that no more than one pore body is filled or emptied during one time step. A minimum time step value may also be imposed to save computational time. However, the effect of time step value is discussed. We check the impact of time step by altering coefficient C_{oe} in Equation (24).

Drainage process for viscosity ratio of 1 is simulated with inlet non-wetting phase reservoir and outlet wetting phase reservoir pressure drop of 7000 Pa. Figure 4 gives average phase pressure difference $P^n - P^w$ and average capillary pressure P^c during drainage for different values of C_{oe} . Here, average phase pressure is calculated based on Equation (26), and average capillary pressure is calculated based on Equation (27). Figure 5 shows saturation change over time for different values of C_{oe} . As can be seen, although difference exists between different values of C_{oe} , it may not be significant. Therefore, for simulations with large network, to improve computational time, a larger value of C_{oe} may be a reasonable choice.

Fig. 4 Dynamic drainage: Average phase pressure difference ($P^n - P^w$) and average capillary pressure P^c for different values of C_{oe}

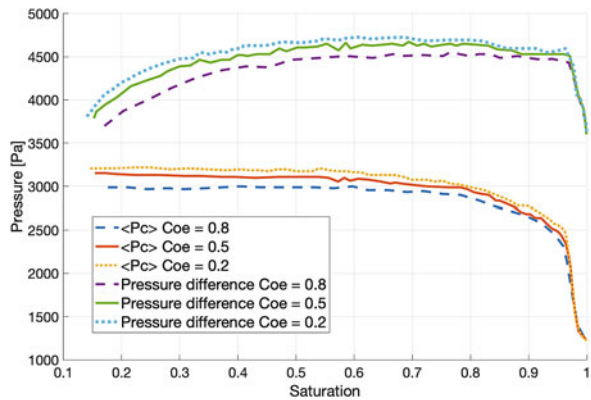
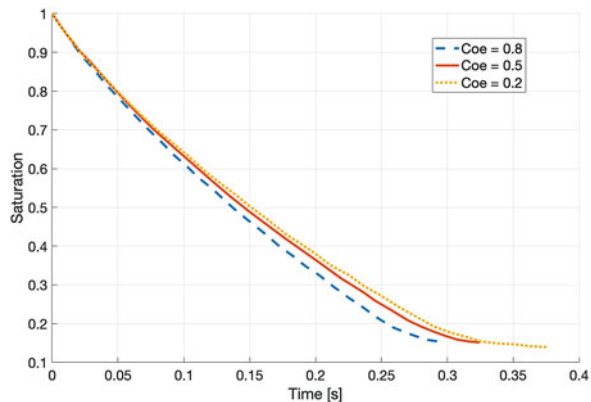


Fig. 5 Wetting phase saturation change with time during drainage for different values of C_{oe}



3.3.3 Nonequilibrium Effects in Average Phase Pressure for Different Viscosity Ratios

Figure 6 shows the average phase pressure difference $P^n - P^w$ and average capillary pressure P^c during drainage for different values of viscosity ratio. Pressure difference between inlet non-wetting phase reservoir and outlet wetting phase reservoir is set to 8000 Pa. As viscosity ratio M decreases, the invasion front becomes more unstable. For a wetting phase saturation S^w of higher than 0.7, lower viscosity ratio in general has higher phase pressure difference. Later as invasion front develops, this is no longer the case. Figure 7 gives total saturation change over time during drainage for different viscosity ratios. Process involving more viscous fluid takes longer time.

Fig. 6 Average phase pressure difference $P^n - P^w$ and average capillary pressure P^c during drainage for different viscosity ratios

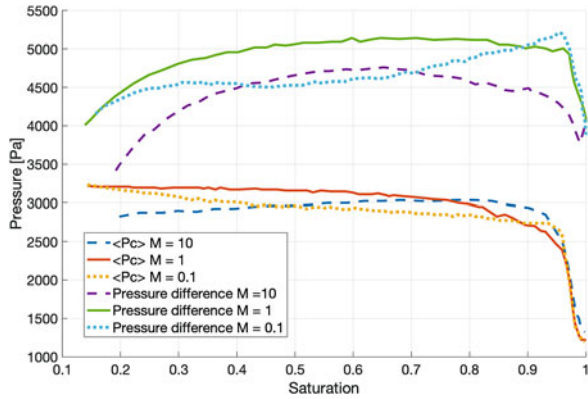


Fig. 7 Saturation change over time during drainage for different viscosity ratios

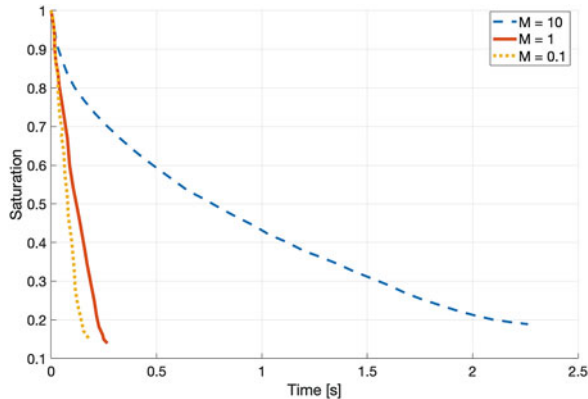
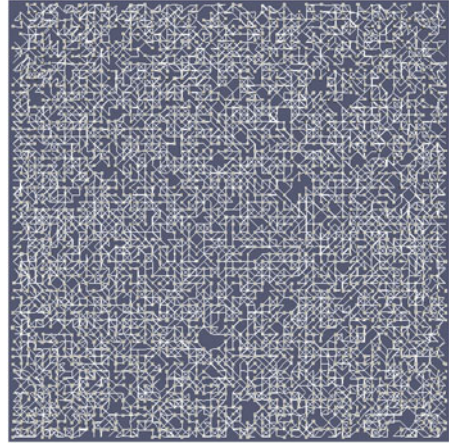


Fig. 8 Network used in simulations of saturation pattern during imbibition



3.4 Single-Pressure Imbibition: Flow Pattern for Different Boundary Pressure Drop

To compare different wetting advances during imbibition, we simulate imbibition into a two-layer network with statistical parameters given in Figure 1. Left side of network is connected with a wetting phase reservoir and the right side of network is connected with a non-wetting phase reservoir. Flow direction is from left to right (see Figure 8). To have different Ca numbers, different boundary pressure drops, $\Delta P = P_{inlet}^w - P_{outlet}^n$, are used: (1) $\Delta P = -1000$ Pa; (2) $\Delta P = 0$ Pa (namely spontaneous imbibition); and (3) $\Delta P = 2000$ Pa.

Figure 9 gives saturation pattern change for these three different pressure drops. As we can see, for negative inlet/outlet pressure drop, invasion is more capillary dominant. With increase of inlet/outlet pressure drop, wetting front gets sharper, which is in quantitative agreement with the reported experimental and numerical results [8, 14].

4 Conclusion

A two-pressure dynamic drainage algorithm is developed for three-dimensional unstructured network model. Time step dependency is discussed through drainage simulations. Dynamic effects in average phase pressure for fluid pairs with different viscosity ratios are studied using the code as an upscaling tool. For cases where the two fluids have significant viscosity difference, viscous pressure drop in one fluid may be negligible. This dynamic algorithm could be simplified to a single-pressure algorithm. Single-pressure algorithm for both drainage and imbibition are developed. Saturation pattern during imbibition for different boundary pressure

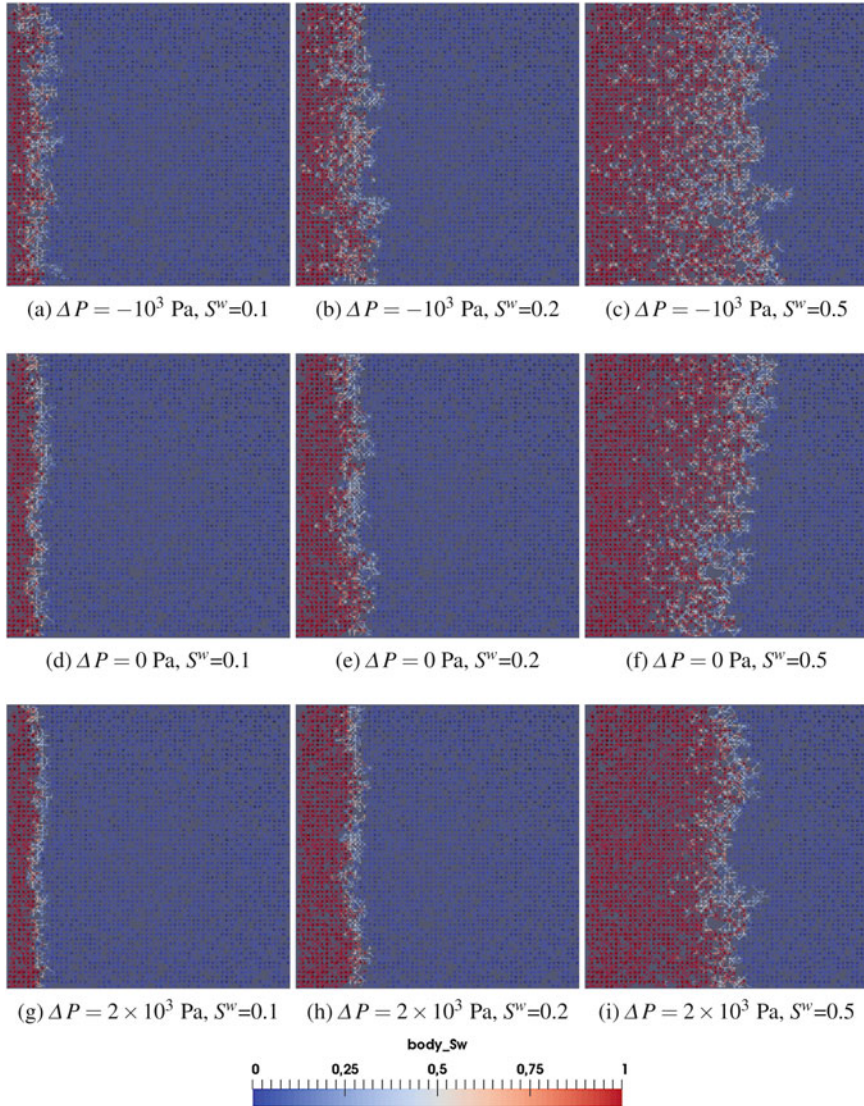


Fig. 9 Saturation pattern during imbibition for different boundary pressure drops

drops are studied. With the increase of boundary pressure, invasion becomes less capillary dominant and wetting front becomes sharper.

Our simulations show the efficiency and capacity of the developed models. These models can be employed in wider applications for study of two-phase flow in porous media.

Acknowledgements S. M. Hassanizadeh has received funding from the European Research Council under the European Union’s Seventh Framework Program (FP/2007–2013)/ERC Grant Agreement No. 341225.

Appendix A: Fluid Distribution Patterns During Primary Imbibition

In general, there are 18 possible fluid occupancy of pores, as shown in Figure 10, where blue color shows the wetting phase and red color shows the non-wetting phase. Here, no attempt is made to represent the real interface shape. However, because of assumptions in our primary imbibition algorithm, some of these configurations will not emerge and the rest ones are shown in Figure 11. To save computational time, we only search these liquid filling scenarios. We use A,B,C to denote three columns and 1–6 to denote the six rows. So, we have 18 possible distribution patterns A1 to C6 in Figure 10. Below, we explain which distribution pattern cannot be encountered and will be, therefore, excluded from considerations.

A2: As we do not consider volume of pore throats, so once there is wetting phase in the pore throat, neighboring pore body cannot be still fully saturated with non-wetting phase.

A5: See A2.

A6: It is not possible to trap wetting phase during primary imbibition in our algorithm.

C2: See A2.

C3: See A6.

C4: See A2.

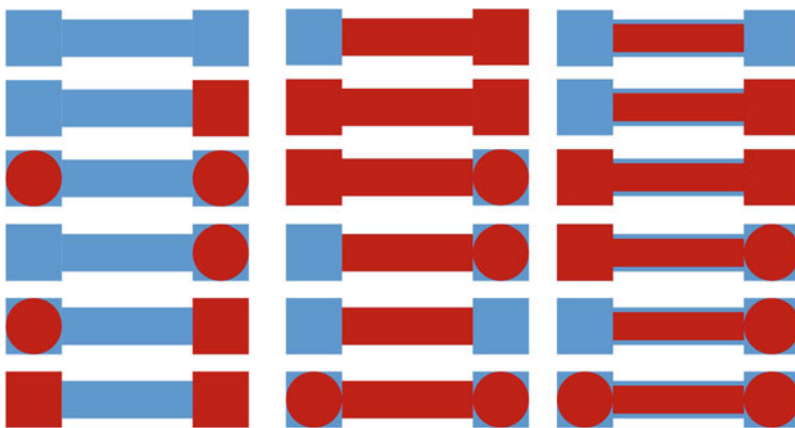


Fig. 10 All possible liquid fillings during imbibition

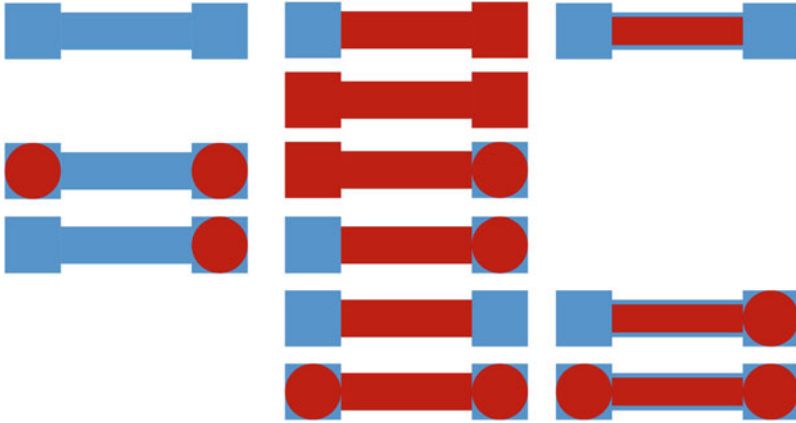


Fig. 11 Liquid fillings during primary imbibition

Appendix B: Interfacial Area for A Cubic Pore Body

Information of interfacial areas of corner interfaces and main terminal menisci for cubic pore bodies can be given as [13]:

Corner Interfaces

For a pore body with inscribed radius R_i filled with wetting and non-wetting phase, non-wetting phase volume can be bigger or smaller than the inscribed sphere volume.

$$R_{i,eq} = \begin{cases} R_i (\frac{6}{\pi} (1 - s_i^w))^{1/3} & s_i^w \geq 0.48 \\ R_i (1 - \exp(-6.83s_i^w))^{1/3} & s_i^w < 0.48 \end{cases} \quad (28)$$

$$A_i^{wn} = \begin{cases} 4\pi R_{i,eq}^2 & s_i^w \geq 0.48 \\ 4\pi R_{i,eq}^2 + 6\pi R_{i,eq} (R_i - R_{i,eq}) & s_i^w < 0.48 \end{cases} \quad (29)$$

Main Terminal Menisci

For main terminal meniscus, namely the interface between a pore body and its neighboring pore throat when non-wetting phase pressure in the pore body is not high enough to invade the pore throat, we have the area for this meniscus as

$$8\pi (\frac{\sigma^{wn}}{p_i^c})^2 (1 - \sqrt{1 - (\frac{r_{ij} D_i^c}{2\sigma^{wn}})^2}).$$

References

1. Aghaei A, Piri M (2015) Direct pore-to-core up-scaling of displacement processes: Dynamic pore network modeling and experimentation. *Journal of Hydrology* 522:488–509
2. Aker E, JØrgen MÅløy K, Hansen A, Batrouni G (1998) A two-dimensional network simulator for two-phase flow in porous media. *Transport in Porous Media* 32(2):163–186
3. Al-Futaisi A, Patzek TW (2003) Extension of Hoshen–Kopelman algorithm to non-lattice environments. *Physica A: Statistical Mechanics and its Applications* 321(3):665–678
4. Al-Gharbi MS, Blunt MJ (2005) Dynamic network modeling of two-phase drainage in porous media. *Physical Review E* 71(1):16,308
5. Azzam M, Dullien F (1977) Flow in tubes with periodic step changes in diameter: A numerical solution. *Chemical Engineering Science* 32(12):1445–1455
6. Blunt M, King P (1990) Macroscopic parameters from simulations of pore scale flow. *Phys Rev A* 42(8):4780–4787
7. Blunt MJ, Bijeljic B, Dong H, Gharbi O, Iglauer S, Mostaghimi P, Paluszny A, Pentland C (2013) Pore-scale imaging and modelling. *Advances in Water Resources* 51:197–216
8. Hughes RG, Blunt MJ (2000) Pore Scale Modeling of Rate Effects in Imbibition. *Transport in Porous Media* 40(3):295–322
9. Joekar-Niasar V, Hassanizadeh S (2012) Network modeling and its application to wicking: two-phase flow approach. In: Masoodi R, Pillai M (eds) *Wicking in Porous Materials: Traditional and Modern Modeling Approaches*, Taylor & Francis Inc.
10. Joekar-Niasar V, Hassanizadeh SM (2012) Analysis of Fundamentals of Two-Phase Flow in Porous Media Using Dynamic Pore-Network Models: A Review. *Critical Reviews in Environmental Science and Technology* 42(18):1895–1976
11. Joekar-Niasar V, Majid Hassanizadeh S (2011) Effect of fluids properties on non-equilibrium capillarity effects: Dynamic pore-network modeling. *International Journal of Multiphase Flow* 37(2):198–214
12. Joekar-Niasar V, Majid Hassanizadeh S (2011) Effect of fluids properties on non-equilibrium capillarity effects: Dynamic pore-network modeling. *International Journal of Multiphase Flow* 37(2):198–214
13. Joekar-Niasar V, Prodanović M, Wildenschild D, Hassanizadeh S (2010) Network model investigation of interfacial area, capillary pressure and saturation relationships in granular porous media. *Water Resources Research* 46(6)
14. Lenormand R, Touboul E, Zarcone C (1988) Numerical models and experiments on immiscible displacements in porous media. *Journal of Fluid Mechanics* 189(-1):165
15. van der Marck SC, Matsuura T, Glas J (1997) Viscous and capillary pressures during drainage: Network simulations and experiments. *Physical Review E* 56(5):5675–5687
16. Mason G, Morrow NR (1991) Capillary behavior of a perfectly wetting liquid in irregular triangular tubes. *Journal of Colloid and Interface Science* 141(1):262–274
17. Mayer RP, Stowe RA (1965) Mercury porosimetry breakthrough pressure for penetration between packed spheres. *Journal of Colloid Science* 20(8):893–911
18. Oren PE, Bakke S, Arntzen OJ (1998) Extending Predictive Capabilities to Network Models. *SPE Journal* 3(December):324–336
19. Patzek T (2001) Verification of a complete pore network simulator of drainage and imbibition. *SPE Journal* 6(2):144–156
20. Piri M, Blunt MJ (2005) Three-dimensional mixed-wet random pore-scale network modeling of two- and three-phase flow in porous media. I. Model description. *Physical Review E* 71(2):026,301
21. Princen H (1970) Capillary phenomena in assemblies of parallel cylinders: III. Liquid Columns between Horizontal Parallel Cylinders. *Journal of Colloid and Interface Science* 34(2):171–184

22. Ransohoff T, Radke C (1988) Laminar flow of a wetting liquid along the corners of a predominantly gas-occupied noncircular pore. *Journal of Colloid and Interface Science* 121(2):392–401
23. Raouf A, Hassanizadeh SM (2012) A new formulation for pore-network modeling of two-phase flow 48(January):1–13
24. Thompson KE (2002) Pore-scale modeling of fluid transport in disordered fibrous materials. *AIChE Journal* 48(7):1369–1389
25. Vidales AM, Riccardo JL, Zgrablich G (1998) Pore-level modelling of wetting on correlated porous media. *Journal of Physics D: Applied Physics* 31(20):2861–2868
26. Zhou D, Blunt M, Orr F (1997) Hydrocarbon Drainage along Corners of Noncircular Capillaries. *Journal of Colloid and Interface Science* 187(1):11–21

Mean Field Magnetohydrodynamic Dynamo in Partially Ionized Plasma: Nonlinear, Numerical Results



K. A. P. Singh

1 Introduction

The magnetic field in most low-mass stars, the Sun and other differentially rotating astrophysical objects such as accretion disks, molecular clouds, galaxies, and the interstellar matter to be generated by MHD dynamo process needs a seed field [1, 2]. In a mean field MHD dynamo, the magnetic field is split into small and large spatial scale components. After taking the average over the small scales, the magnetic induction equation describes the generation of the large-scale magnetic field.

The mean field dynamo equation contains the average induction and the turbulent transport and dissipation coefficients. The existence of the scale separation between a global scale and a microscopic scale (the scale of the turbulent fluctuations) is an integral part of the mean field dynamo.

The MHD phenomena in partially ionized plasma is usually different from that in the fully ionized plasma [3–5]. A major part of the solar photosphere [6, 7], the protoplanetary disks [8], and the molecular clouds [9] are partially ionized. The dynamo action in partially ionized plasma would be affected by the multi-fluid interactions between electrons, ions and neutrals. There is an interesting issue of the possible disconnect between the subsurface and surface solar magnetic field as emphasized (see, e.g., [10]), that may have some bearing on the neglect of the neutral fluid-plasma coupling in the flux generation and transport on the solar photosphere.

The MHD dynamo process, in principle, can take place through the entire convection zone and such a distributed dynamo is likely to be shaped largely by

K. A. P. Singh (✉)

Department of Physics, Institute of Science, Banaras Hindu University, Varanasi, India

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41, https://doi.org/10.1007/978-3-030-02487-1_22

357

the near-surface shear layer [11]. The high-resolution observations from the solar optical telescope (SOT) onboard Hinode satellite have shown mesogranular scale, internetwork magnetic field dominated by the horizontally inclined magnetic fields [12, 13]. The value of these horizontal magnetic fields typically is a few hundred Gauss and the ratio between the horizontal and vertical components of magnetic fields lies typically between 4 and 7. The ubiquitous existence of the mesogranular scale, horizontal magnetic fields on the Sun is probably due to the operation of a dynamo action in the near-surface zone [14]. Simulations including the effects of strong stratification, compressibility, partial ionization, radiative transfer have demonstrated an exponential growth of magnetic fields that approach a strength of about 25 G near the visible solar surface for magnetic Reynolds number of 2600 [14, 15]. The partial ionization in these simulations is treated in the equation of state, i.e. the mean molecular weight is not constant, and it is a function of both specific internal energy and density.

It is well known that the strong differential rotation at the base (tachocline) of the convection zone has a strong influence in the generation of magnetic fields via dynamo action [16]. The limitations associated with the tachocline dynamos are pointed out [11]. In the solar photosphere where the plasma is partially ionized, the presence of large-scale subsurface shear (c.f. [17]) can affect the evolution of toroidal and poloidal components of the magnetic field. In this paper, we study the role of shear flows in driving the mean field dynamo including the Hall drift and the ambipolar diffusion arising in partially ionized plasma. The description of modified magnetic induction equation in partially ionized plasma is provided in Section 2, the mean field dynamo equations are presented in Section 3, and finally the results and discussions are presented in the last section.

2 Mean field Dynamo in Three-Component Magnetofluid

The three-component partially ionized plasma consists of electrons (e), ions (i) of uniform mass density ρ_i and neutral particles (n) of uniform mass density ρ_n (c.f. [3, 4] for details). Neglecting the electron inertial force from the equation of motion of electrons, the electric field \mathbf{E} is found to be

$$\mathbf{E} = -\frac{\mathbf{V}_e \times \mathbf{B}}{c} - \frac{\nabla p_e}{en_e} - \frac{m_e}{e} \nu_{en} (\mathbf{V}_e - \mathbf{V}_n) - \frac{m_e}{e} \nu_{ei} (\mathbf{V}_e - \mathbf{V}_i). \quad (1)$$

Equation (1) is the modified Ohm's Law. The modification in Equation (1) arises due to the electron pressure gradient (∇p_e), electron-ion and electron-neutral collisions. In case of low ionization fraction, the ion dynamics can be ignored from the ion-momentum equation. The momentum equation for the ions then reduces to following form:

$$0 = -\nabla p_i + en_i \left(\mathbf{E} + \frac{\mathbf{V}_i \times \mathbf{B}}{c} \right) - \nu_{in} \rho_i (\mathbf{V}_i - \mathbf{V}_n) - \nu_{ie} \rho_i (\mathbf{V}_i - \mathbf{V}_e). \quad (2)$$

The ion-electron collision frequency (ν_{ie}) and the electron-ion (ν_{ei}) collision frequency are related by $\nu_{ei} = \left(\frac{\rho_i}{\rho_e}\right) \nu_{ie}$. Taking $\mathbf{J} = en_i(\mathbf{V}_i - \mathbf{V}_e)$ and $n_i = n_e$, the relative velocity between the ions and neutrals is found to be related through

$$\mathbf{V}_i - \mathbf{V}_n = -\nabla \frac{(\rho_i + \rho_e)}{\nu_{in}\rho_i} + \frac{\mathbf{J} \times \mathbf{B}}{c\nu_{in}\rho_i} \quad (3)$$

The magnetic induction equation, i.e. Equation (3) is modified as a result of the collisions in partially ionized plasma. The Faraday's law of induction can be written in terms of the magnetic field and electron velocity as

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times (\mathbf{V}_e \times \mathbf{B}) + \eta \nabla^2 \mathbf{B}$$

where the pressure term has been neglected for the incompressible case with constant temperature, η is the magnetic diffusivity that arises due to the electron-neutral collisions and the electron-ion collisions. In weakly ionized plasma, the electron-neutral collision dominates over the electron-ion collisions (c.f. [4]).

The contributions to the electron velocity flow (\mathbf{V}_e) comes from neutral velocity flow, the relative drift due to the neutrals and ions and the relative drift between the ions and the electrons, and it can be understood through the construction

$$(\mathbf{V}_e \times \mathbf{B}) = [\mathbf{V}_n - (\mathbf{V}_n - \mathbf{V}_i) - (\mathbf{V}_i - \mathbf{V}_e)] \times \mathbf{B}.$$

Substituting the term $(\mathbf{V}_e \times \mathbf{B})$ in the Faraday's law, one can arrive at the modified form of the magnetic induction equation that is shown below:

$$\frac{\partial \mathbf{B}}{\partial t} = \nabla \times \left(\mathbf{V}_n - \frac{\mathbf{J}}{en_e} + \frac{\mathbf{J} \times \mathbf{B}}{c\nu_{in}\rho_i} \right) \times \mathbf{B} + \eta \nabla^2 \mathbf{B} \quad (4)$$

Equation (4) contains the Hall term ($= \frac{\mathbf{J}}{en_e}$) and the ambipolar diffusion term ($= \frac{\mathbf{J} \times \mathbf{B}}{c\nu_{in}\rho_i}$). In the solar atmosphere, it is found that the Hall effect dominates in the solar photosphere whereas the ambipolar diffusion dominates in the chromosphere (c.f. [4]).

3 Dynamo Equations for Three-Component Magnetofluid

The magnetic induction equation is subjected to Hall drift and ambipolar diffusion, in addition to Ohmic diffusion, in partially ionized plasma. Following the procedure of mean field dynamo (c.f. [18]), the velocity \mathbf{V}_E and the magnetic field \mathbf{B} are split into their average parts and fluctuating parts. Refer [3] for further details on the derivations and detailed expressions.

The average and fluctuating parts in the mean field dynamo correspond to the large-scale and small-scale behaviours, respectively. So, we write

$$\mathbf{V}_E = \overline{\mathbf{V}_E} + \mathbf{V}_{E'} \quad (5)$$

$$\mathbf{B} = \overline{\mathbf{B}} + \mathbf{B}' \quad (6)$$

such that $\overline{\mathbf{V}_{E'}} = 0$ and $\overline{\mathbf{B}'} = 0$ where the bar on top of the physical quantities denotes spatial or the time average over small scales. One of the important parts of the mean field dynamo is to solve the magnetic induction equation for large- and small-scale fields. Some of the details of basic mean field dynamo equations relevant to our study are introduced here for the sake of complete understanding (refer [3] for details).

While deriving an expression for $\mathbf{V}_{E'}$, the terms containing the second or higher order quantities (of small-scale, fluctuating variables) are neglected, and only the terms containing small-scale variables (e.g., $\frac{\mathbf{J}' \times \overline{\mathbf{B}}}{c\nu_{in}\rho_i}$, $\frac{\overline{\mathbf{J}} \times \mathbf{B}'}{c\nu_{in}\rho_i}$, etc.) are retained. The small-scale fluctuation in density is ignored. Substituting Equations (5) and (6) into the magnetic induction equation, and under the first-order smoothing approximation (FOSA), the fluctuating velocity is given by

$$\mathbf{V}_{E'} = \mathbf{V}_{n'} - \frac{\mathbf{J}'}{en_e} + \frac{\mathbf{J}' \times \overline{\mathbf{B}}}{c\nu_{in}\rho_i} + \frac{\overline{\mathbf{J}} \times \mathbf{B}'}{c\nu_{in}\rho_i},$$

and the mean flow is found to be

$$\overline{\mathbf{V}_E} = \overline{\mathbf{V}_n} - \frac{\overline{\mathbf{J}}}{en_e} + \frac{\overline{\mathbf{J}} \times \overline{\mathbf{B}}}{c\nu_{in}\rho_i}.$$

An important physical quantity in the mean field dynamo is the electromotive force (ε) defined as

$$\varepsilon = \overline{\mathbf{V}_{E'} \times \mathbf{B}'}$$

and is a function of the mean magnetic induction $\overline{\mathbf{B}}$, and the mean quantities formed from the fluctuations. In the standard process of the mean field dynamo, the electromotive force is expressed as

$$\varepsilon = \alpha \overline{\mathbf{B}} - \beta \nabla \times \overline{\mathbf{B}},$$

where the kinetic helicity (α) is given by the mean field dynamo that contains turbulence, and additional contribution to it comes from the Hall drift and the ambipolar diffusion. The kinetic helicity is defined as

$$\alpha = -\frac{\tau_{cor}}{3} \overline{\mathbf{V}_{E'} \cdot (\nabla \times \mathbf{V}_{E'})}$$

and one can write $\alpha = \alpha_v + \alpha_H + \alpha_{Am}$. While getting an expression of α , the second and higher order terms of small-scale, fluctuating variables (e.g., \mathbf{J}' , \mathbf{B}' , etc.) are neglected.

The quantity α_v is a measure of the average kinetic helicity of the neutral fluid in the turbulence possessing correlations over timescale of τ_{cor} , and it is given by

$$\alpha_v = -\frac{\tau_{cor}}{3} \overline{\mathbf{V}_n' \cdot \boldsymbol{\Omega}_n'}$$

The contribution due to the Hall drift appears in the quantity α_H given by

$$\alpha_H = \frac{2}{3} \frac{\tau_{cor}}{en_e} \overline{\mathbf{J}' \cdot \boldsymbol{\Omega}_n'}$$

while the contribution due to the ambipolar effect appears in the quantity α_{Am} given by

$$\alpha_{Am} = \alpha_A \cdot \overline{\mathbf{B}}$$

with

$$\alpha_A = \frac{2}{3} \frac{\tau_{cor}}{c v_{in} \rho_i} \overline{\mathbf{J}' \times \boldsymbol{\Omega}_n'}$$

where $\boldsymbol{\Omega}_n' = \nabla \times \mathbf{V}_n'$ is the vorticity of the fluctuating, neutral component of the turbulent fluid. It may be noted that the Hall- α requires a component of the fluctuating current density along the fluctuating vorticity of the neutral fluid. The ambipolar- α results from the component of the fluctuating current density perpendicular to the fluctuating vorticity. The turbulent dissipation is given by

$$\beta = \frac{\tau_{cor}}{3} \overline{\mathbf{V}_E'^2} = \beta_v + \beta_H + \beta_{Am},$$

with $\beta_v = \frac{\tau_{cor}}{3} \overline{\mathbf{V}_n'^2}$ as a measure of the average turbulent kinetic energy of the neutral fluid in the turbulent regime possessing correlations over timescale τ_{cor} , and $\beta_H = -2 \frac{\tau_{cor}}{3en_e} \overline{\mathbf{J}' \cdot \mathbf{V}_n'}$ represents the contribution to turbulent dissipation due to the Hall effect. The coupling of the charged components with the neutral fluid is clearly manifested through the possible correlations between the current density fluctuations and the velocity fluctuations of the neutral fluid. The ambipolar term yields $\beta_{Am} = \beta_A \cdot \overline{\mathbf{B}}$ and $\beta_A = \frac{2\tau_{cor}}{3c v_{in} \rho_i} \overline{\mathbf{J}' \times \mathbf{V}_n'}$ with its essential nonlinear character manifest through its dependence on the average magnetic induction. One also observes that the Hall- β requires a component of current density fluctuations along the velocity fluctuations of the neutral fluid, whereas the ambipolar effect depends upon the component of the current density fluctuations perpendicular to the velocity fluctuations. The rigid or perfectly conducting boundary conditions

(all surface contributions vanish) are used while determining the averages of the quantities [c.f. 3]. Here the mean flow is taken as to be nonzero. The dynamo equation then reduces to a form

$$\frac{\partial \overline{\mathbf{B}}}{\partial t} = \nabla \times (\overline{\mathbf{V}_E} \times \overline{\mathbf{B}}) + \nabla \times (\alpha \overline{\mathbf{B}} - \beta \nabla \times \overline{\mathbf{B}}) + \overline{\eta} \nabla^2 \overline{\mathbf{B}}. \quad (7)$$

Let $\overline{\mathbf{B}}$ and $\frac{\partial \overline{A}}{\partial x}$ represent the toroidal and the poloidal components of the magnetic field, respectively. The mean flow (c.f. [19]) is given by

$$\overline{\mathbf{V}_E} = \overline{V(z)} \hat{\mathbf{e}}_y \quad (8)$$

We drop the bars on top in B and A for brevity. Assuming one-dimensional dependence, one can express $\mathbf{B} = (0, B, \frac{\partial A}{\partial x})$ in Cartesian coordinates. The coordinates (x, y, z) here correspond to the polar coordinates (θ, φ, r) . The boundary conditions then turn out to be the vanishing of B and A at the endpoints of a finite x -interval, say $x = 0$ and $x = \pi R$ that correspond to the two opposite poles of the sphere. We introduce a turbulent magnetic diffusivity η_1 . It is then convenient to present the magnetic induction equation in a dimensionless form using a normalizing magnetic field B_0 , a spatial scale R , a time-scale $\frac{R^2}{\eta_1}$, and writing $A = \tilde{A} R$. The rotation $\overline{\Omega}$ is made dimensionless using $\frac{\eta_1}{R^2}$. We obtain the following set of equations:

$$\begin{aligned} \frac{\partial B}{\partial t} = & \overline{\Omega} \frac{\partial \tilde{A}}{\partial x} + \frac{\partial^2 B}{\partial x^2} - R_\alpha \frac{\partial^2 \tilde{A}}{\partial x^2} - R_{\alpha A} \frac{\partial}{\partial x} \left[\left(B + a \frac{\partial \tilde{A}}{\partial x} \right) \frac{\partial \tilde{A}}{\partial x} \right] \\ & + R_{\beta A} \frac{\partial}{\partial x} \left[\left(B + b \frac{\partial \tilde{A}}{\partial x} \right) \frac{\partial B}{\partial x} \right] \end{aligned} \quad (9)$$

and

$$\frac{\partial \tilde{A}}{\partial t} = R_\alpha B + \frac{\partial^2 \tilde{A}}{\partial x^2} + R_{\alpha A} \left(B + a \frac{\partial \tilde{A}}{\partial x} \right) B + R_{\beta A} \frac{\partial}{\partial x} \left[\left(B + b \frac{\partial \tilde{A}}{\partial x} \right) \frac{\partial^2 \tilde{A}}{\partial x^2} \right] \quad (10)$$

where the coefficients $(R_\alpha, R_{\alpha A}, R_{\beta A}, a, b)$ appearing in the above equations and η_1 can be recovered from [3] and $\overline{\Omega} = \frac{\partial \overline{V}}{\partial z}$. One can assume, as an example, a model of turbulence composed of Alfvénic waves [3] and then solve Equations (9) and (10) numerically. The numerical solutions to the coupled, partial differential equations containing the physics of mean field dynamo in partially ionized astrophysical plasma are presented and discussed in the next section.

4 Discussions and Conclusions

The magnetic field is omnipresent in over a wide variety of astrophysical objects from clusters ($\sim 10^{20}$ km) to neutron stars (~ 10 km). The magnetic field strength in the astrophysical systems varies by several orders of magnitude and there exists a scale hierarchy, from large scale to small scale, in the astrophysical magnetic fields. The understanding behind the small-scale evolution of magnetic fields on the Sun is of crucial importance since the fundamental processes and interactions on various spatiotemporal scales are observationally not so accessible for other astrophysical objects.

The solar magnetic fields are organized in a hierarchy of structures that extend from large-scale active regions with a size of about 10^5 to 10^2 km [20]. With high-resolution observations of the fine structure of magnetic fields in the solar atmosphere becoming available, it became clear that the bulk of the magnetic flux has a quasi-fractal pattern at the solar surface (see [21] for the small-scale solar magnetic fields). The magnetic structures in Hinode magnetograms, for example, appear to scale in a self-similar power-law fashion over observable length scales [22]. The power-law pattern in the magnetic structure is typical of a multi-fractal behaviour due to turbulence which shows a self-similar behaviour over many length scales.

Following [23], a small-scale solar dynamo (or SSD) is considered to be any self-sustaining dynamo process in the photosphere or deeper in the convection zone that operates on scales small enough to be unaffected by solar rotation and to produce a dominance of fields with very small scales. In this sense, SSD is qualitatively different from the well-known dynamo models of the solar cycle that is believed to arise from a large-scale dynamo process that is driven by interaction of solar differential rotation and magnetic field and the large-scale fluid motions in the solar convection zone. Apart from the SSD, the only other plausible explanation of the small-scale magnetic fields in the quiet Sun is that they result from the interaction of turbulent convection with the large-scale fields associated with the global dynamo [23].

Much of our understanding of dynamo mechanisms has been based on the theoretical framework of MHD. For a low density and partially ionized plasma environment that prevails near the solar photosphere and overlying chromosphere, the role of Hall drift and the ambipolar diffusion on dynamo cannot be ignored. The role of Hall current in the magnetic field generation through the dynamo processes in a low density astrophysical system is studied [24]. The Hall currents are found to have a profound impact on the generation of macroscopic magnetic field. The effect of the Hall term is twofold: it transfers energy towards larger scales for scales larger than the Hall length, and it transfers energy towards smaller scales for scales smaller than the Hall length scale [25].

The evolution of an initially weak and small-scale magnetic field in a system maintained in a stationary regime of hydrodynamic turbulence is studied [26]. It was found that depending on the relative values of the length scales of the system, the

Hall current can strongly enhance or suppress the generation of large-scale magnetic energy. The amplification/generation of fast plasma flows by microscale (turbulent) magnetic fields via magnetofluid coupling has been investigated [27]. It has been found that the macroscopic magnetic fields and flows are generated simultaneously and proportionately from microscopic fields and flows. In this case, a strong macro-scale will be generated in the presence of a strong microscale driver.

The ambipolar diffusion (AD) was included in the dynamo studies [28, 29]. The role of ambipolar diffusion and the strong nonlinearity that AD introduces in the system has been studied in the context of astrophysical dynamos [30]. The fast dynamos in weakly ionized gases have also been studied in the context of galactic magnetic fields [31].

In mean field approach, the contribution to α and β -coefficients comes also due to the Hall effect and ambipolar diffusion (c.f. Section 3). The importance of Hall drift on α -dynamo action was noted [32], recalling with the inclusion of Hall drift in MHD, the freezing of the magnetic field is associated with the electron flow rather than the bulk (or centre-of-mass) velocity field.

The role of Hall effect and ambipolar diffusion on the stability of weakly ionized, magnetized planar shear flows is studied [33]. The development of instability in the presence of Hall drift and in large-scale shear flows in low density astrophysical plasmas is studied to find that a Hall magneto-shear (MSI) develops in the presence of a non-rotating shear flows [34]. The Hall magneto-rotational instability (MRI) develops in the presence of differentially rotating flows.

4.1 Numerical Results

In the present work, Equations (9) and (10) are solved numerically along with the appropriate boundary conditions including the shear flow. The term containing the shear appears in the y -component of the dynamo equations and the initial conditions are taken to be $\tilde{A}(x,0) = 0$, $B(x,0) = \sin x$. The vanishing of B and A at the endpoints of a finite x -interval, say $x = 0$ and $x = \pi R$, correspond to the two opposite poles of the sphere, has been taken as the boundary conditions. The dispersion relation shows how the dynamo waves are modified as a result of the surface shear flow and the Hall effect. The contribution to α due to the Hall drift appears in R_α through α_1 . The contribution due to the ambipolar diffusion appears in $R_{\alpha A}$ and $R_{\beta A}$. The ratios a and b provide the information of relative strengths with which $R_{\alpha A}$ and $R_{\beta A}$ act on the system (or system of equations). The turbulent diffusivity η_1 also has a contribution from the Hall effect in addition to β_v and the Ohmic diffusivity (η). For the dynamo waves in the absence of ambipolar diffusion ($R_{\alpha A} = 0$, $R_{\beta A} = 0$), Equations (9) and (10) become linear with a solution of the form $\exp[i(kx - \omega t)]$ giving the dispersion relation

$$\left(k^2 - i\omega\right)^2 - iR_\alpha \overline{\Omega} k - R_\alpha^2 k^2 = 0 \quad (11)$$

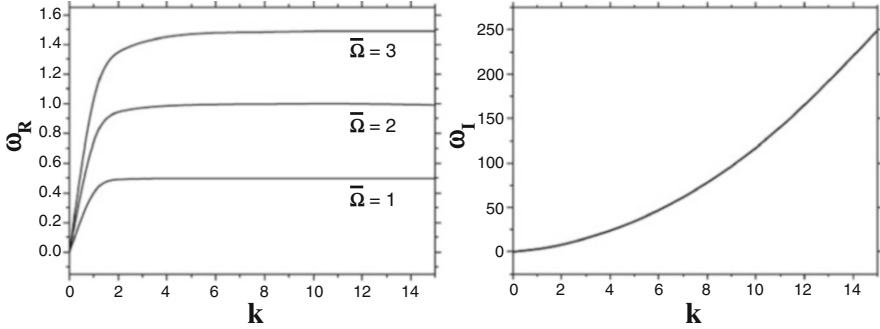


Fig. 1 The dispersion (ω_R) and growth (ω_I) as a function of wavenumber for $R_\alpha = 1.6$, $R_{\alpha A} = 0$, $R_{\beta A} = 0$ for three different values of $\bar{\Omega}$

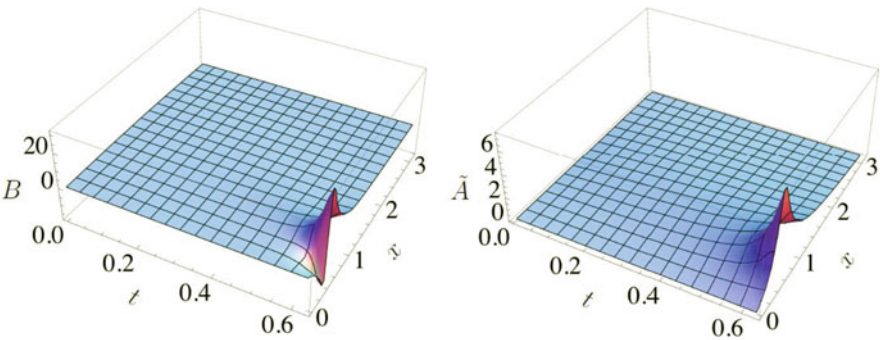


Fig. 2 Dependence of B and \bar{A} on t and x in three-dimensional representation for $\bar{\Omega} = 1$, $R_\alpha = 1.6$, $R_{\alpha A} = 1.7$, $R_{\beta A} = 0.1$, $a = 1$, $b = 0.7$

In the absence of shear flow ($\bar{\Omega} = 0$), we have $\omega = -ik (k \mp R_\alpha)$ and for $k > 0$, one of the two solutions shows that the field grows for $k < R_\alpha$. Equation (11) is solved numerically and Figure 1 shows the dispersion and the growth rate as a function of the wavenumber for $R_\alpha = 1.6$ for three different values of $\bar{\Omega}$. A nonzero shear affects the dispersion but the growth rate is not so sensitive to the values of $\bar{\Omega}$ that we have taken in this work. Overall, the magnetic field evolution is affected by the shear flow.

In this work, we have studied the mean field dynamo in partially ionized plasma by including the shear flow. The resulting variations of magnetic field as a function of t and x are shown in Figures. 2 and 3, using $\bar{\Omega} = 1$, $R_\alpha = 1.6$, $R_{\alpha A} = 1.7$, $R_{\beta A} = 0.1$, $a = 1$, $b = 0.7$. It is found that both the components of magnetic field (B and \bar{A}) grow in time, and this clearly shows the formation of sharp features. One can also see the reversal of the sign of B near $x \sim 0.4$.

The effect of ambipolar diffusion along with shear flow and Hall effect is shown in Figures 4 and 5. The magnetic field as a function of t and x is shown for a fixed value of shear flow ($\bar{\Omega} = 1$) and $R_\alpha = 0.2$, $R_{\alpha A} = 3.5$, $R_{\beta A} = 1.5$, $a = 1$,

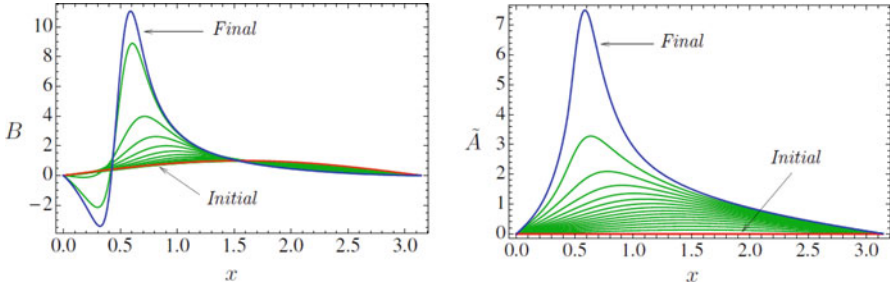


Fig. 3 Evolution of profiles of B and \tilde{A} as a function of x in two-dimensional representation for $\bar{\Omega} = 1$, $R_{\alpha} = 1.6$, $R_{\alpha A} = 1.7$, $R_{\beta A} = 0.1$, $a = 1$, $b = 0.7$. Red line represents the initial ($t = 0$) profile and blue the final. Green is for the intermediate ones at the time step of 0.042

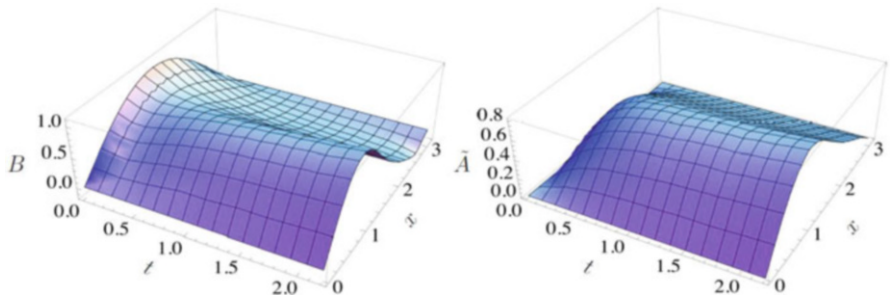


Fig. 4 B and \tilde{A} as a function t and x in three-dimensional representation for $\bar{\Omega} = 1$, $R_{\alpha} = 0.2$, $R_{\alpha A} = 3.5$, $R_{\beta A} = 1.5$, $a = 1$, $b = 1$

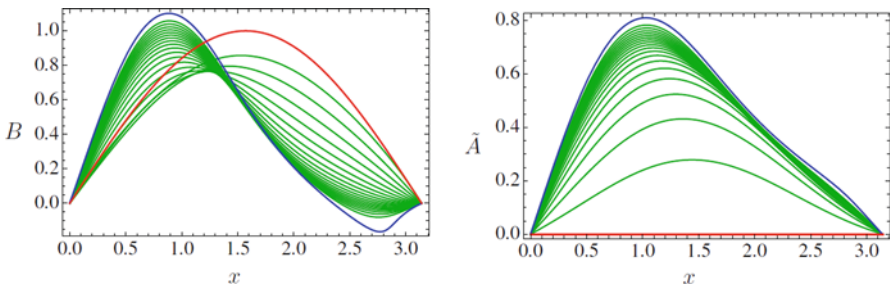


Fig. 5 Evolution of profiles of B and \tilde{A} as a function of x in two-dimensional representation for $\bar{\Omega} = 1$, $R_{\alpha} = 0.2$, $R_{\alpha A} = 3.5$, $R_{\beta A} = 1.5$, $a = 1$, $b = 1$. Red line represents the initial ($t = 0$) profile and blue the final. Green is for the intermediate ones at the time step of 0.11

$b = 1$. The inclusion of shear flow along with the Hall effect and ambipolar diffusion results in an increase of the magnetic field and also production of small-scale spatial structures. In the absence of the shear flow, there exists a critical wavenumber, $k_c = R_{\alpha}$ below which the field grows. For $\bar{\Omega} = 3$, the magnetic field strength

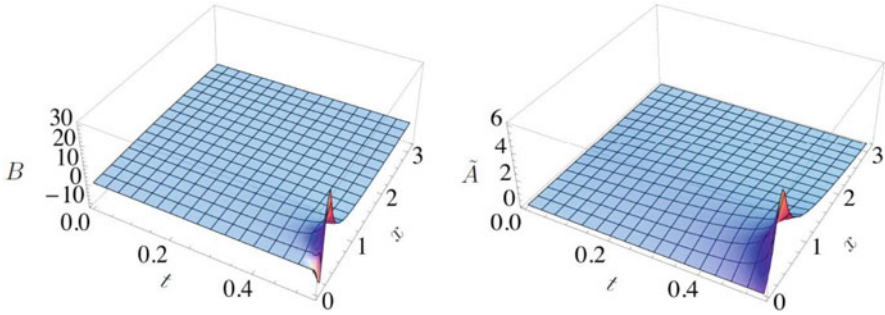


Fig. 6 B and \tilde{A} as a function t and x in three-dimensional representation for $\overline{\Omega} = 3, R_\alpha = 1.6, R_{\alpha A} = 1.7, R_{\beta A} = 0.1, a = 1, b = 0.7$

increases and it evolves comparatively on a faster timescale (Figure 6). It is clear that the presence of a finite amount of surface shear flow affects the magnetic field evolution significantly.

The high-resolution observations of the partially ionized layers of the Sun show internetwork magnetic field dominated by the horizontally inclined magnetic fields [12, 13]. It is observed that the ratio of the horizontal and vertical components of magnetic fields lies typically between 4 to 7. In the present work, the ratio of the components of magnetic field B_z/B_y , where $B_y = B$ and $B_z = \partial \tilde{A} / \partial x$, is computed numerically and then plotted in Figure 7. It is clear that between $x = 0.5$ and $x = 1.5$, the Hall and ambipolar diffusion in the presence of shear flow can explain the observed ratio of the components of magnetic field in the internetwork region of the Sun. It is an interesting result because in the classical view the magnetic field in the solar surface is concentrated into strong (\sim kG) vertical flux tubes and we have shown (through the mean field dynamo approach) how the horizontal magnetic fields are produced in the partially ionized surface of the Sun. It is important to mention that the surface dynamo simulations at the highest magnetic Reynolds number also yield a ratio of the horizontal and vertical flux density consistent with the observational results, but the overall amplitudes are low [35]. The ratio of mean horizontal and vertical component in such dynamo generated magnetic fields reaches values between 2 and 4 in the optical depth interval $-2 < \log \tau < -1$. It has been shown in the dynamo simulations of the quiet Sun that the ratio of horizontal to vertical field depends upon the height and it peaks at 450 km above the $\tau = 1$ [36]. Moreover, the ratio is found to decrease with the increasing field strength. This strong dependence of ratio of horizontal and vertical field on the magnetic field and height suggests that increasing resolution and reducing numerical diffusivities in the simulations is not sufficient to reach the observationally inferred field strengths on the Sun.

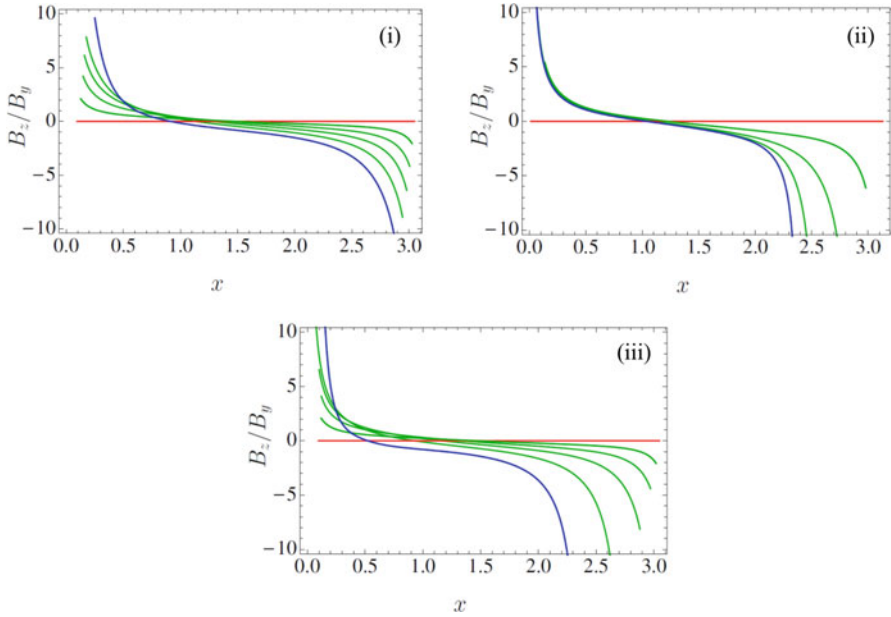


Fig. 7 (i) The ratio of magnetic field components B_z/B_y as a function of x in two-dimensional representation. Red line shows the initial and blue line shows the final case. Green line shows the intermediate ones at the time step of 0.1. The parameters chosen are $\bar{\Omega} = 1$, $R_\alpha = 1.6$, $R_{\alpha A} = 1.7$, $R_{\beta A} = 0.1$, $a = 1$, $b = 0.7$. (ii) Same as (i) but for $\bar{\Omega} = 1$, $R_\alpha = 0.2$, $R_{\alpha A} = 3.5$, $R_{\beta A} = 1.5$, $a = 1$, $b = 1$. The green line is for the intermediate ones at the time step of 0.5. (iii) Same as (i) but for $\bar{\Omega} = 3$, $R_\alpha = 1.6$, $R_{\alpha A} = 1.7$, $R_{\beta A} = 0.1$, $a = 1$, $b = 0.7$. The green line is for the intermediate ones at the time step of 0.2

4.2 Conclusions

The small-scale dynamo arising due to the shear flow, Hall drift and the ambipolar diffusion in the presence of the large-scale density gradient can change the transport of magnetic fields. It would be very interesting to understand how the small-scale dynamo operating in partially ionized solar and astrophysical plasma affects the physical properties and magnetic field evolution at the large spatiotemporal scales. Even though the dynamics of small-scale structures is often unobservable in the astrophysical plasmas, the physics of partially ionized plasma can play an important role through the nonlinear interactions with the large-scale density gradient present in the system. In conclusion, the present study of mean field dynamo in partially ionized plasma suggests that the surface shear, Hall drift and ambipolar diffusion can play an important role in driving a small-scale dynamo near the surface layers of the cool stars. The Hall and ambipolar terms appearing in the magnetic induction equation depend strongly upon the magnetic field, ion and neutral density. So, gravitational stratification can also affect the dynamo process and magnetic field evolution in the solar photosphere and solar chromosphere.

Acknowledgements The author gratefully acknowledges Joint ICTP-IAEA College on Advance Plasma Physics, 2014 and the financial support from International Centre for Theoretical Physics (ICTP), Trieste, Italy for providing an important platform for useful discussions on MHD Dynamo with Prof. Swadesh Mahajan. The author would also like to thank S.M. Chitre, Vinod Krishan and R.T. Gangadhara for useful discussions and support.

References

1. Moffatt, H.K.: Magnetic field generation in electrically conducting fluids, Cambridge University Press, England (1978)
2. Parker, E.N.: Cosmical magnetic fields: Their origin and their activity, Oxford University Press, New York (1979)
3. Krishan, V., Gangadhara, R.T.: Mean – field dynamo in partially ionized plasmas - I, Mon. Not. Royal Astron. Soc., **385**, 849 - 853 (2008)
4. Singh, K.A.P., Krishan, V.: Alfvén-like mode in partially ionized solar atmosphere, New Astron., **15**, 119 – 125 (2010)
5. Singh, K.A.P., Hillier, A., Isobe, H., Shibata, K.: Nonlinear Instability and intermittent nature of magnetic reconnection in solar chromosphere, Pub. Astron. Soc. Japan, **67**, 96 (1-11) (2015)
6. Leake, J.E., Arber, T.D.: The emergence of magnetic flux through a partially ionized solar chromosphere, Astron. Astrophys., **450**, 805-818 (2006)
7. Krishan, V., Varghese, B.A.: Cylindrical Hall-MHD waves: a nonlinear solution, Solar Phys., **247**, 343-349, (2008)
8. Krishan, V., Yoshida, Z.: Equilibrium structures in partially ionized rotating plasmas within Hall magnetohydrodynamics, Phys. Plasmas, **13**, 092303 (5pp.) (2006)
9. Brandenburg, A., Zweibel, E.G.: The formation of sharp structures by ambipolar diffusion Astrophys. J., **427**, L91-L94 (1994)
10. Schüssler, M.: Flux tubes, surface magnetism, and the solar dynamo: constraints and open problems, Astron. Nachr., **326**, 194-204 (2005)
11. Brandenburg, A.: The case for a distributed solar dynamo shaped by near-surface shear Astrophys. J., **625**, 539-547 (2005)
12. Orozco Suárez, D., Bellot Rubio, L.R., del Toro Iniesta, J.C., et al.: Quiet-Sun internetwork magnetic fields from the inversion of Hinode measurements, Astrophys. J., **670**, L61-L64 (2007)
13. Lites, B.W., Kubo, M., Socas-Navarro, H., et al.: The horizontal magnetic flux of the quiet-Sun internetwork as observed with the Hinode spectro-polarimeter, Astrophys. J., **672**, 1237-1253 (2008)
14. Vögler, A., Schüssler, M.: A solar surface dynamo, Astron. Astrophys., **465**, L43-L46 (2007)
15. Schüssler, M., Vögler, A.: Strong horizontal photospheric magnetic field in a surface dynamo simulation, Astron. Astrophys., **481**, L5-L8 (2008)
16. Tobias, S.M., Weiss, N.O.: The solar dynamo and tachocline (Eds. D.W. Hughes, R. Rosner, N.O. Weiss), Cambridge University Press, Cambridge, UK (2007)
17. Hindman, B.W., Gizon, L., Duvall, T., et al.: Comparison of solar subsurface flows assessed by ring and time-distance analyses, Astrophys. J., **613**, 1253-1262 (2004)
18. Krause, F. and Rädler, K. H.: Mean-field magnetohydrodynamics and dynamo theory, Pergamon Press, Oxford, New York (1980)
19. Stix, M.: Non-linear dynamo waves, Astron. Astrophys., **20**, 9-12 (1972)
20. Schüssler, M.: The Sun, a laboratory for astrophysics (Eds. J.T. Schmelz, J. C. Brown), Kluwer Academic Publishers, Netherlands, 191 (1992)
21. de Wijn, A.G., Stenflo, J.O., Solanki, S.K., Tsuneta, S.: Small-scale solar magnetic fields, Space Science Rev., **144**, 275-315 (2009)

22. Pietarila Graham, J., Danilovic, S., Schüssler, M.: Turbulent magnetic fields in the quiet Sun: implications of Hinode observations and small-scale dynamo simulations, *Astrophys. J.*, **693**, 1728-1735 (2009)
23. Lites, B.W.: Hinode observations suggesting the presence of a local small-scale turbulent dynamo, *Astrophys. J.*, **737**, 52 (9pp.) (2011)
24. Mininni, P. D., Gómez, D.O., Mahajan, S. M.: Role of the Hall current in magnetohydrodynamic dynamos, *Astrophys. J.*, **584**, 1120-1126 (2003)
25. Gómez, D.O., Mininni, P.D., Dmitruk, P.: Hall-magnetohydrodynamic small-scale dynamos, *Physical Rev. E*, **82**, 036406 (10pp.) (2010)
26. Mininni, P. D., Gómez, D.O., Mahajan, S. M.: Dynamo action in magnetohydrodynamics and Hall-magnetohydrodynamics, *Astrophys. J.*, **587**, 472-481 (2003)
27. Mahajan, S.M., Shatashvili, N.L., Mikeladze, S.V.: Acceleration of plasma flows due to reverse dynamo mechanism, *Astrophys. J.*, **634**, 419-425 (2005)
28. Zweibel, E.G.: Ambipolar diffusion drifts and dynamos in turbulent gases, *Astrophys. J.*, **329**, 384-391 (1988)
29. Proctor, M.R.E., Zweibel, E.G.: Dynamos with ambipolar diffusion drifts, *Geophys. Astrophys. Fluid Dyn.*, **64**, 145-161 (1992)
30. Brandenburg, A., Subramanian, K.: Large scale dynamos with ambipolar diffusion nonlinearity, *Astron. Astrophys.*, **361**, L33-L36 (2000)
31. Zweibel, E. G., Heitsch, F.: Fast Dynamos in weakly ionized gases, *Astrophys. J.*, **684**, 373-379 (2008)
32. Mininni, P. D., Gómez, D.O., Mahajan, S. M.: Dynamo action in Hall magnetohydrodynamics, *ApJ.*, **567**, L81 - L83 (2002)
33. Kunz, M.W.: On the linear stability of weakly ionized, magnetized planar shear flows, *Mon. Not. Royal Astron. Soc.*, **385**, 1494-151 (2008)
34. Bejarano C., Gómez D.O., Brandenburg, A.: Shear-driven instabilities in Hall-magnetohydrodynamic plasmas, *Astrophys. J.*, **737**, 62 (11pp.) (2011)
35. Danilovic, S., Schüssler, M., Solanki, S.: Probing quiet Sun magnetism using MURaM simulations and Hinode/SP results: support for a local dynamo, *Astron. Astrophys.*, **513**, A1 (8pp.) (2010)
36. Rempel, M.: Numerical simulations of quiet Sun magnetism: On the Contribution from a Small-scale Dynamo, *Astrophys. J.*, **789**, 132 (22pp.) (2014)

Outcome of Wall Features on the Creeping Sinusoidal Flow of MHD Couple Stress Fluid in an Inclined Channel with Chemical Reaction



Mallinath Dhange and Gurunath Sankad

2010 Mathematics Subject Classification 76V05, 76Z05, 76W99, 92C10

1 Introduction

The fluid mechanical study of peristalsis has received considerable attention in the last two decades mainly because of its potential applications to the biological systems and peristaltic movement of slurries, harmful fluids of nuclear industry. It is a mechanism for transport of fluids which is achieved when a progressive wave of area contraction or expansion propagates along the length of a distensible tube containing fluid. In the view of its importance, some workers [1–3] have explored the peristaltic movement of different liquids under various circumstances. In physiological structures, it is known that all vessels are not straight but have some inclination with axis. The gravitational strength is accounted due to the consideration of inclined channel. A few scholars have studied the peristaltic flow of Newtonian and non-Newtonian fluids in an inclined channel with various conditions [4–6].

It is seen that couple stress fluid behavior is exceptionally useful in understanding dissimilar physiological and mechanical procedures. Couple stress fluid is a fluid consisting of rigid, randomly oriented particles suspended in a viscous medium,

M. Dhange (✉) · G. Sankad
Department of Mathematics, B.L.D.E.A.S.V.P.Dr.P.G. Halakatti College of Engineering and Technology, (Affiliated to Visvesvaraya Technological University, Belagavi, India), Vijayapur 586103, Karnataka, India
e-mail: math.mallinath@bldeacet.ac.in; math.gurunath@bldeacet.ac.in

© Springer Nature Switzerland AG 2019
V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_23

371

such as blood, lubricants containing small amount of high polymer additive, electro-rheological fluids and synthetic fluids. The couple stress model introduced by Stokes [7] has distinct features. The main feature of couple stresses is to introduce a size-dependent effect. These fluids are able to describe blood, suspension fluids, and various types of lubricants. Such studies clarify the behavior of rheological complex liquids. Some studies on peristaltic transport of couple stress fluid have been reported in references [8–13]. After this study, few investigators have explored the wall effects on different fluids with peristalsis [14–17].

Magnetohydrodynamic (MHD) peristaltic flow nature of liquid is especially imperative in mechanical and physiological procedures. In the existence of magnetic field, many fluids possess an electrically conducting nature, which is an important aspect of the physical situation in the flow problems of magnetohydrodynamics. It is useful for tumor treatment, MRI glancing, blood pumping, reduction of bleeding during surgeries, targeted transportation of drugs, and so on. Magneto-therapy is an essential application to human body. This heals the diseases like ulceration, inflammations, and diseases of uterus. Some researchers [18–22] have explored the magnetohydrodynamic character of non-Newtonian liquids through different circumstances. They discussed the effects of magnetic field, permeability, micropolar, couple stress, and wall parameters.

Dispersion describes the spread of particles through random motion from regions of higher concentration to regions of lower concentration. The problem of dispersion in the presence of homogeneous and heterogeneous chemical reactions is of importance in several contexts, for example, in nuclear physics, gas absorption in agitated tank, biological systems, and the flow of nuclear fuel where heat is generated in the bulk. Dispersion plays a crucial task in chyme transport and other applications like environmental pollutant transportation, chromatographic separation, and the mixing and transport of drugs or toxic substances in physiological structures [23]. The basic theory on dispersion was first proposed by Taylor [24], who investigated theoretically and experimentally that the dispersion of a solute is miscible with a liquid flowing through a channel. Several workers [25–27] have investigated the dispersion of a solute in viscous fluid, under different limitations. Furthermore, some investigators [28–36] extended this analysis to non-Newtonian fluids.

Existing information on the topic witnessed that an analytical treatment of creeping sinusoidal flow and dispersion of an MHD couple stress fluid with chemical reaction and complaint wall has been never reported. Motivated from the reported literature, we have explored the impact of chemical and wall features on the MHD creeping sinusoidal stream and dispersion of couple stress fluid in an inclined channel. The principle outcomes are presented in the conclusion part. The present issue might be appropriate for the treatment on intestinal disorder, gallstones in gallbladder without surgery.

2 Mathematical Modelling

Consider the MHD couple stress fluid with peristalsis in the 2-dimensional inclined channel. Figure 1 depicts the wave shape.

The wave shape is given by the subsequent equation:

$$\pm \left[a \sin \frac{2\pi}{\lambda} (\mathcal{X} - ct) + d \right] = \pm h = \mathcal{Y}, \tag{1}$$

where the half width of the channel is d , the wavelength of the peristaltic wave is λ , the amplitude is a , and the speed is c .

The relating flow conditions of the current issue are

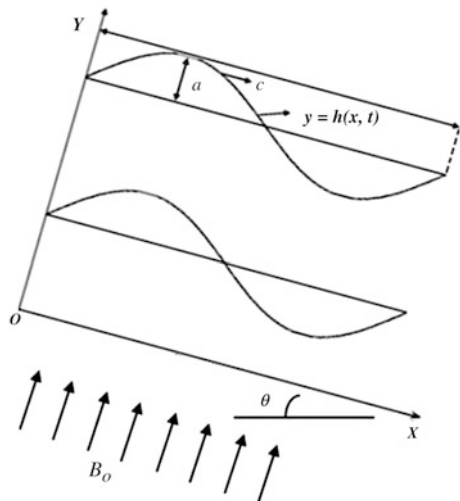
$$0 = \frac{\partial \mathcal{V}}{\partial \mathcal{Y}} + \frac{\partial \mathcal{U}}{\partial \mathcal{X}}, \tag{2}$$

$$-\frac{\partial p}{\partial \mathcal{X}} + \mu \nabla^2 \mathcal{U} - \eta' \nabla^4 \mathcal{U} - \sigma B_0^2 \mathcal{U} + \rho g \sin \theta = \rho \left[\frac{\partial}{\partial t} + \mathcal{U} \frac{\partial}{\partial \mathcal{X}} + \mathcal{V} \frac{\partial}{\partial \mathcal{Y}} \right] \mathcal{U}, \tag{3}$$

$$-\frac{\partial p}{\partial \mathcal{Y}} + \mu \nabla^2 \mathcal{V} - \eta' \nabla^4 \mathcal{V} - \rho g \cos \theta = \rho \left[\frac{\partial}{\partial t} + \mathcal{U} \frac{\partial}{\partial \mathcal{X}} + \mathcal{V} \frac{\partial}{\partial \mathcal{Y}} \right] \mathcal{V}, \tag{4}$$

where $\frac{\partial^2}{\partial \mathcal{X}^2} + \frac{\partial^2}{\partial \mathcal{Y}^2} = \nabla^2$, $\nabla^2 \nabla^2 = \nabla^4$, the constant associated with couple stress fluid is η' , the fluid density is ρ , the viscosity coefficient is μ , the velocity components in the \mathcal{X} , \mathcal{Y} direction is \mathcal{U} , \mathcal{V} , the pressure is p , the inclination of the channel is θ , and the magnetic field is B_0 .

Fig. 1 Geometry of the problem.



Referring Mitra and Prasad [14], the condition of the flexible wall movement is specified as:

$$p - p_0 = \mathcal{L}(h), \tag{5}$$

where the movement of the stretched membrane by the damping force is \mathcal{L} and is intended by the subsequent equation:

$$-\mathcal{T} \frac{\partial^2}{\partial \mathcal{X}^2} + m \frac{\partial^2}{\partial t^2} + \mathcal{C} \frac{\partial}{\partial t} = \mathcal{L}. \tag{6}$$

Here, the coefficient of sticky damping force is \mathcal{C} , the mass per/area is m , and the membrane tension is \mathcal{T} .

In the absence of body couples and body forces, and under long wavelength approximations the conditions (2) to (4) yield as:

$$0 = \frac{\partial \mathcal{Y}}{\partial \mathcal{Y}} + \frac{\partial \mathcal{U}}{\partial \mathcal{X}}, \tag{7}$$

$$0 = -\frac{\partial p}{\partial \mathcal{X}} + \mu \frac{\partial^2 \mathcal{U}}{\partial \mathcal{Y}^2} - \eta' \frac{\partial^4 \mathcal{U}}{\partial \mathcal{Y}^4} - \sigma B_0^2 \mathcal{U} + \rho g \sin \theta, \tag{8}$$

$$0 = -\frac{\partial p}{\partial \mathcal{Y}}. \tag{9}$$

The allied border conditions are

$$0 = \mathcal{U}, \quad 0 = \frac{\partial^2 \mathcal{U}}{\partial \mathcal{Y}^2}, \quad \text{at} \quad \mathcal{Y} = \pm h. \tag{10}$$

It is presumed that $p_0 = 0$ and the channel walls are inextensible; therefore, the straight displacement of the wall is nil and only lateral movement takes place, and

$$0 = \mu \frac{\partial^2 \mathcal{U}}{\partial \mathcal{Y}^2} - \eta' \frac{\partial^4 \mathcal{U}}{\partial \mathcal{Y}^4} - \sigma B_0^2 \mathcal{U} + \rho g \sin \theta = \frac{\partial}{\partial \mathcal{X}} \mathcal{L}(h), \quad \text{at} \quad \mathcal{Y} = \pm h, \tag{11}$$

where

$$\frac{\partial p}{\partial \mathcal{X}} = P' = -\mathcal{T} \frac{\partial^3 h}{\partial \mathcal{X}^3} + m \frac{\partial^3 h}{\partial \mathcal{X} \partial t^2} + \mathcal{C} \frac{\partial^2 h}{\partial \mathcal{X} \partial t} = \frac{\partial}{\partial \mathcal{X}} \mathcal{L}(h). \tag{12}$$

Solving the conditions (8) and (9) with (10) and (11), we obtain

$$-\frac{T'}{\sigma B_0^2} [A'_1 \cosh(m'_1 \mathcal{Y}) + A'_2 \cosh(m'_2 \mathcal{Y}) + 1] = \mathcal{U}(\mathcal{Y}), \tag{13}$$

where $m'_1 = \sqrt{\frac{\mu}{2\eta'} \left(1 + \sqrt{1 - \frac{4\eta'\sigma B_0^2}{\mu^2}} \right)}$, $m'_2 = \sqrt{\frac{\mu}{2\eta'} \left(1 - \sqrt{1 - \frac{4\eta'\sigma B_0^2}{\mu^2}} \right)}$.

The mean speed is specified as:

$$\frac{1}{2h} \int_{-h}^h \mathcal{U}(\mathcal{Y}) d\mathcal{Y} = \bar{\mathcal{U}}. \tag{14}$$

Conditions (13) and (14) yield as:

$$-\frac{T'}{\sigma B_0^2} \left[\frac{A'_1}{m'_1 h} \sinh(m'_1 h) + \frac{A'_2}{m'_2 h} \sinh(m'_2 h) + 1 \right] = \bar{\mathcal{U}}. \tag{15}$$

Utilizing Ravikiran and Radhakrishnamacharya [30], the liquid speed is given by the condition:

$$\mathcal{U} - \bar{\mathcal{U}} = \mathcal{U}_{\mathcal{X}}. \tag{16}$$

Conditions (13), (15), and (16) yield as:

$$-\frac{T'}{\sigma B_0^2} \left[A'_1 \cosh(m'_1 \mathcal{Y}) + A'_2 \cosh(m'_2 \mathcal{Y}) - \frac{A'_1}{m'_1 h} \sinh(m'_1 h) - \frac{A'_2}{m'_2 h} \sinh(m'_2 h) \right] = \mathcal{U}_{\mathcal{X}}, \tag{17}$$

where $A'_1 = \frac{(m'_2)^2}{[(m'_1)^2 - (m'_2)^2] \cosh(m'_1 h)}$, $A'_2 = \frac{-(m'_1)^2}{[(m'_1)^2 - (m'_2)^2] \cosh(m'_2 h)}$,

$$P' = -T \frac{\partial^3 h}{\partial x^3} + m \frac{\partial^3 h}{\partial x \partial t^2} + C \frac{\partial^2 h}{\partial x \partial t}, \quad T' = \frac{\partial p}{\partial \mathcal{X}} - \frac{\rho g}{\mu} \sin \theta.$$

3 Homogeneous - Heterogeneous Chemical Reactions with Diffusion

Alluding Taylor [24] and Gupta and Gupta [26], the scattering condition for the concentration of solution \mathcal{C} of the material for the current issue in isothermal circumstances is

$$\mathcal{D} \frac{\partial^2 \mathcal{C}}{\partial \mathcal{Y}^2} - k_1 \mathcal{C} = \frac{\partial \mathcal{C}}{\partial t} + \mathcal{U} \frac{\partial \mathcal{C}}{\partial \mathcal{X}} \tag{18}$$

Here, the rate constant of first-order chemical response is k_1 , the molecular diffusion coefficient is \mathcal{D} , and liquid concentration is \mathcal{C} .

The dimensionless quantities are specified as:

$$\eta = \frac{\mathcal{Y}}{d}, \xi = \frac{(\mathcal{X} - \bar{\mathcal{U}}t)}{\lambda}, R = \frac{\rho g}{\mu}, \mathcal{H} = \frac{h}{d}, \mathcal{P} = \frac{d^2}{\mu c \lambda} P',$$

$$\theta = \frac{t}{\bar{t}}, \bar{t} = \frac{\lambda}{\bar{\mathcal{U}}}, \mathcal{M} = \sqrt{\frac{\sigma B_0^2 d^2}{\mu}}. \tag{19}$$

For the regular estimations of physiologically essential parameters of this issue, it is normal that $\bar{\mathcal{U}} \approx \mathcal{C}$ (Ravikiran and Radhakrishnamacharya [30]).

To proceed further, we use $\bar{\mathcal{U}} \approx \mathcal{C}$, in condition (18) and the conditions (12), (17), (18) are nondimensionalized as:

$$-\epsilon \left[-\mathcal{E}_3 (2\pi)^2 \sin(2\pi\xi) + (\mathcal{E}_1 + \mathcal{E}_2) (2\pi)^3 \cos(2\pi\xi) \right] = \mathcal{P}, \tag{20}$$

$$-\frac{T}{\sigma B_0^2} [A_1 \cosh(m_1\eta) + A_2 \cosh(m_2\eta) + A_3] = \mathcal{U}\mathcal{X}, \tag{21}$$

$$\frac{d^2}{\lambda \mathcal{D}} \mathcal{U}\mathcal{X} \frac{\partial \mathcal{C}}{\partial \xi} = \frac{\partial^2 \mathcal{C}}{\partial \eta^2} - \frac{k_1 d^2}{\mathcal{D}} \mathcal{C}, \tag{22}$$

where

$$m_1 = m'_1 d = \sqrt{\frac{\gamma^2}{2} \left(1 + \sqrt{1 - \frac{4\mathcal{M}^2}{\gamma^2}} \right)}, \quad m_2 = m'_2 d = \sqrt{\frac{\gamma^2}{2} \left(1 - \sqrt{1 - \frac{4\mathcal{M}^2}{\gamma^2}} \right)},$$

the amplitude ratio is $\epsilon (= \frac{a}{d})$, the rigidity is $\mathcal{E}_1 (= -\frac{\mathcal{T} d^3}{\lambda^3 \mu c})$, the stiffness is

$\mathcal{E}_2 (= \frac{m c d^3}{\lambda^3 \mu})$, the viscous damping force in the wall is $\mathcal{E}_3 (= \frac{c d^3}{\mu \lambda^2})$, the

couple stress constraint is $\gamma (= d \sqrt{\frac{\mu}{\eta}})$, and the magnetic field constraint

is $\mathcal{M} (= B_0 d \sqrt{\frac{\sigma}{\mu}})$, $T = \mathcal{P} - R \sin \theta$.

The dispersion with first-order irreversible chemical response occurs in the mass of the liquid and at the channel walls. Referring Chandra and Philip [28], the wall conditions are specified as:

$$0 = \frac{\partial \mathcal{C}}{\partial \mathcal{Y}} + \mathcal{F}\mathcal{C} \quad \text{at} \quad \mathcal{Y} = h = [a \sin \frac{2\pi}{\lambda} (\mathcal{X} - \bar{\mathcal{U}}t) + d], \tag{23}$$

$$0 = \frac{\partial \mathcal{C}}{\partial \mathcal{Y}} - \mathcal{F}\mathcal{C} \quad \text{at} \quad \mathcal{Y} = -h = -[a \sin \frac{2\pi}{\lambda} (\mathcal{X} - \bar{\mathcal{U}}t) + d]. \tag{24}$$

Condition (19), (23), and (24) yield as:

$$0 = \frac{\partial \mathcal{C}}{\partial \eta} + \beta \mathcal{C} \quad \text{at} \quad \eta = \mathcal{H} = [\epsilon \sin(2\pi \xi) + 1], \tag{25}$$

$$0 = \frac{\partial \mathcal{C}}{\partial \eta} - \beta \mathcal{C} \quad \text{at} \quad \eta = -\mathcal{H} = -[\epsilon \sin(2\pi \xi) + 1], \tag{26}$$

where the heterogeneous response rate constraint is $\beta = \mathcal{F}d$, relating to catalytic response at the dividers.

Assuming that $\frac{\partial \mathcal{C}}{\partial \xi}$ is independent of η at any cross section and utilizing conditions (25) and (26), the primitive of (22) is obtained as:

$$-\frac{d^2}{\lambda \mathcal{D}} \frac{T}{\sigma B_0^2} \frac{\partial \mathcal{C}}{\partial \xi} \left[A_4 \cosh(m_1 \eta) + A_5 \cosh(m_2 \eta) + A_6 \cosh(\alpha \eta) + A_7 \right] = \mathcal{C}(\eta). \tag{27}$$

The volumetric flow rate \mathcal{Q} is specified as:

$$\int_{-\mathcal{H}}^{\mathcal{H}} \mathcal{C} \mathcal{U}_x d\eta = \mathcal{Q}. \tag{28}$$

Using conditions (21) and (27) in (28), we obtain

$$-2 \frac{d^6}{\lambda \mu^2 \mathcal{D}} \frac{\partial \mathcal{C}}{\partial \xi} \mathcal{G}(\xi, \epsilon, \alpha, \beta, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{M}, \theta, \gamma) = \mathcal{Q}, \tag{29}$$

where

$$\begin{aligned} \mathcal{G}(\xi, \epsilon, \alpha, \beta, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{M}, \theta, \gamma) = & -\frac{T^2}{\mathcal{M}^4} \left[\frac{A_1 A_4}{2} B_1 + \frac{A_2 A_5}{2} B_2 \right. \\ & + (A_1 A_5 + A_2 A_4) B_3 + A_1 A_6 B_4 \\ & \left. + A_2 A_6 B_5 + (A_1 A_7 + A_3 A_4) B_6 + (A_2 A_7 + A_3 A_5) B_7 + A_3 A_6 B_8 + A_3 A_7 \mathcal{H} \right], \\ A_1 = & \frac{(m_2)^2}{[(m_1)^2 - (m_2)^2] \cosh(m_1 \mathcal{H})}, \quad A_2 = \frac{-(m_1)^2}{[(m_1)^2 - (m_2)^2] \cosh(m_2 \mathcal{H})}, \\ A_3 = & \frac{-(m_2)^2 \sinh(m_1 \mathcal{H})}{m_1 \mathcal{H} [(m_1)^2 - (m_2)^2] \cosh(m_1 \mathcal{H})} + \frac{(m_1)^2 \sinh(m_2 \mathcal{H})}{m_2 \mathcal{H} [(m_1)^2 - (m_2)^2] \cosh(m_2 \mathcal{H})}, \\ A_4 = & \frac{(m_2)^2}{[(m_1)^2 - (\alpha)^2] [(m_1)^2 - (m_2)^2] \cosh(m_1 \mathcal{H})}, \quad A_6 = A_3 L_1 - A_4 L_2 - A_5 L_3, \end{aligned}$$

$$\begin{aligned}
 A_5 &= \frac{-(m_1)^2}{[(m_2)^2 - (\alpha)^2][(m_1)^2 - (m_2)^2] \cosh(m_2 \mathcal{H})}, & A_7 &= -\frac{A_3}{\alpha^2}, \\
 L_1 &= \frac{\beta}{\alpha^2(\alpha \sinh(\alpha \mathcal{H}) + \beta \cosh(\alpha \mathcal{H}))}, & L_2 &= \frac{(m_1 \sinh(m_1 \mathcal{H}) + \beta \cosh(m_1 \mathcal{H}))}{(\alpha \sinh(\alpha \mathcal{H}) + \beta \cosh(\alpha \mathcal{H}))}, \\
 L_3 &= \frac{(m_2 \sinh(m_2 \mathcal{H}) + \beta \cosh(m_2 \mathcal{H}))}{(\alpha \sinh(\alpha \mathcal{H}) + \beta \cosh(\alpha \mathcal{H}))}, & B_1 &= \frac{2m_1 \mathcal{H} + \sinh(2m_1 \mathcal{H})}{2m_1}, \\
 B_2 &= \frac{2m_2 \mathcal{H} + \sinh(2m_2 \mathcal{H})}{2m_2}, & B_6 &= \frac{\sinh(m_1 \mathcal{H})}{m_1}, & B_7 &= \frac{\sinh(m_2 \mathcal{H})}{m_2}, & B_8 &= \frac{\sinh(\alpha \mathcal{H})}{\alpha}, \\
 B_3 &= \frac{m_1 \sinh(m_1 \mathcal{H}) \cosh(m_2 \mathcal{H}) - m_2 \cosh(m_1 \mathcal{H}) \sinh(m_2 \mathcal{H})}{[(m_1)^2 - (m_2)^2]}, \\
 B_4 &= \frac{m_1 \sinh(m_1 \mathcal{H}) \cosh(\alpha \mathcal{H}) - \alpha \cosh(m_1 \mathcal{H}) \sinh(\alpha \mathcal{H})}{[(m_1)^2 - (\alpha)^2]}, \\
 B_5 &= \frac{m_2 \sinh(m_2 \mathcal{H}) \cosh(\alpha \mathcal{H}) - \alpha \cosh(m_2 \mathcal{H}) \sinh(\alpha \mathcal{H})}{[(m_2)^2 - (\alpha)^2]}, & \alpha &= \sqrt{\frac{k_1}{\mathcal{D}}}d.
 \end{aligned}$$

Looking at condition (29) with Fick’s law of scattering, the dispersing coefficient D^* was computed to such an extent that the solute disperses near to the plane moving with the typical speed of the flow and is specified as:

$$2 \frac{d^6}{\mu^2 \mathcal{D}} \mathcal{G}(\xi, \epsilon, \alpha, \beta, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{M}, \theta, \gamma) = \mathcal{D}^*. \tag{30}$$

The mean of \mathcal{G} is $\bar{\mathcal{G}}$ and is attained as:

$$\int_0^1 \mathcal{G}(\xi, \epsilon, \alpha, \beta, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{M}, \theta, \gamma) d\xi = \bar{\mathcal{G}}. \tag{31}$$

4 Outcomes and Discussion

The expression for $\bar{\mathcal{G}}(\xi, \epsilon, \alpha, \beta, \mathcal{E}_1, \mathcal{E}_2, \mathcal{E}_3, \mathcal{M}, \theta, \gamma)$ as shown in Equation (31) has been obtained by numerical integration using the software MATHEMATICA and the domino effects are presented through graphs. It may ensure that $\mathcal{E}_1, \mathcal{E}_2,$ and \mathcal{E}_3 cannot be zero all together.

The effects of the magnetic constraint (\mathcal{M}) and the couple stress constraint (γ) on the scattering coefficient ($\bar{\mathcal{G}}$) are depicted in Figures 2, 3, 4 and 5. It is observed

Fig. 2 Illustration of magnetic constraint (\mathcal{M}) with scattering coefficient ($\bar{\mathcal{G}}$) when $\mathcal{E}_1 = 0.1$, $\mathcal{E}_2 = 4.0$, $\mathcal{E}_3 = 0.06$, $\epsilon = 0.2$, $\alpha = 1.0$, $\gamma = 6.0$, and $\theta = \pi/6$.

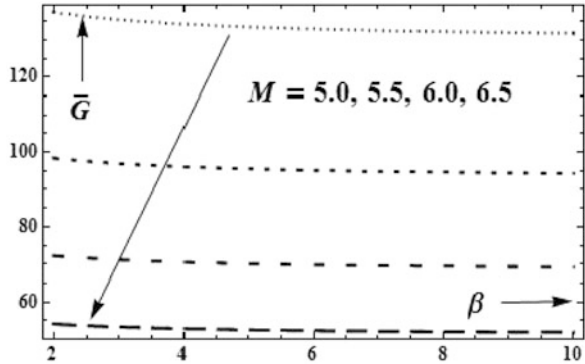


Fig. 3 Illustration of magnetic constraint (\mathcal{M}) with scattering coefficient ($\bar{\mathcal{G}}$) when $\mathcal{E}_1 = 0.1$, $\mathcal{E}_2 = 0.0$, $\mathcal{E}_3 = 0.06$, $\epsilon = 0.2$, $\beta = 5.0$, $\gamma = 6.0$, and $\theta = \pi/6$.

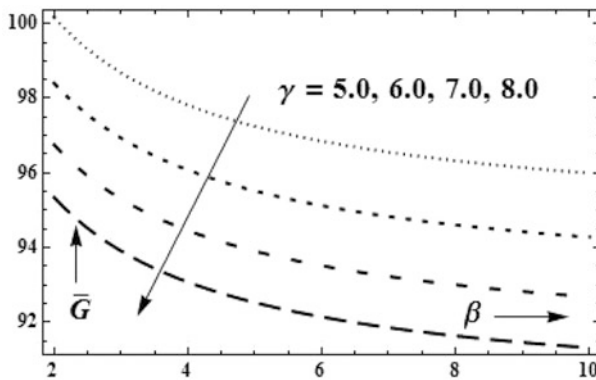
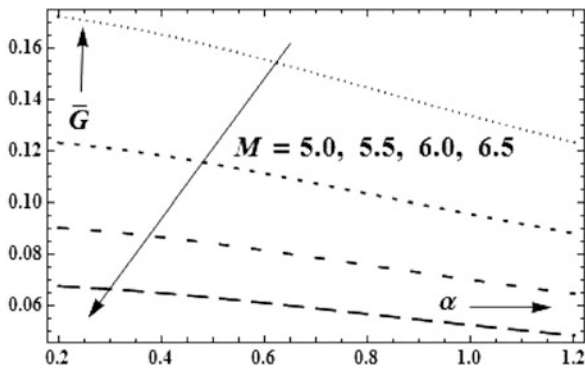


Fig. 4 Illustration of couple stress constraint (γ) with scattering coefficient ($\bar{\mathcal{G}}$) when $\mathcal{E}_1 = 0.1$, $\mathcal{E}_2 = 4.0$, $\mathcal{E}_3 = 0.06$, $\epsilon = 0.2$, $\alpha = 1.0$, $\mathcal{M} = 5.5$, and $\theta = \pi/6$.

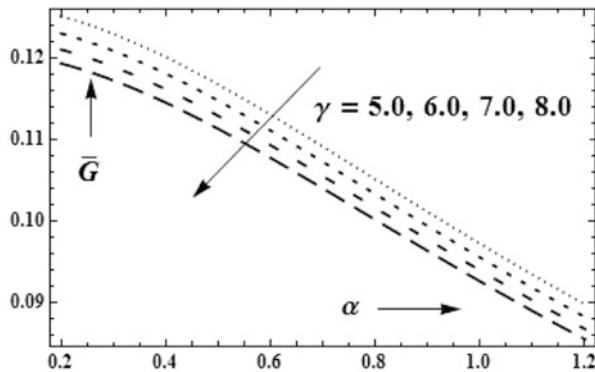


Fig. 5 Illustration of couple stress constraint (γ) with scattering coefficient ($\bar{\mathcal{G}}$) when $\mathcal{E}_1 = 0.1$, $\mathcal{E}_2 = 0.0$, $\mathcal{E}_3 = 0.06$, $\epsilon = 0.2$, $\beta = 5.0$, $\mathcal{M} = 5.5$, and $\theta = \pi/6$.

that $\bar{\mathcal{G}}$ descends with rise in magnetic constraint (\mathcal{M}) (Figures 2 and 3). This is because of that the application of a magnetic field normal to the flow direction has a tendency to slow down the movements of the fluid in the channel, and it gives rise to a resistance force called the Lorentz force which acts opposite to the flow direction and as a result dispersion diminishes. Figures 4 and 5 depict that $\bar{\mathcal{G}}$ descends with an increase in couple stress constraint (γ), as a result dispersion may reduce. This finding agrees with the conclusions of Alemayehu and Radhakrishnamacharya [29] and Abbas et al. [34].

Figures 6 and 7 depict that concentration profile ($\bar{\mathcal{G}}$) and inclination of the channel (θ) behave alike. This result concurs with arguments of Sankad and Radhakrishnamacharya [20] and Rathod et al. [21]. The impacts of the rigidity constraint (\mathcal{E}_1), stiffness constraint (\mathcal{E}_2), and viscous damping force (\mathcal{E}_3) of the wall on the dissipating coefficient ($\bar{\mathcal{G}}$) are illustrated in Figures 8, 9, 10, 11, 12 and 13. It is experiential that $\bar{\mathcal{G}}$ ascends monotonically with an expansion in \mathcal{E}_1 , \mathcal{E}_2 , and \mathcal{E}_3 . This understanding might be true that increment in the flexibility of the channel walls helps the stream moment which causes to enhance the dispersion. This result agrees with the outcomes of Ravikiran and Radhakrishnamacharya [30].

It is seen that $\bar{\mathcal{G}}$ descends with an increase in the homogeneous compound response rate constraint (α) (Figures 3, 5, 7, 9, 11, and 13). Also, it is noticed from the Figures 2, 4, 6, 8, 10, and 12 that the scattering diminishes with heterogeneous substance response rate constraint (β), and the decrease in the effective scattering coefficient is sharp in a section near to the wall. This agrees with chemical point of view because the reactions which affect diffusion happen only at the surface for heterogeneous substance response. This implies that heterogeneous substance response tends to decrease the scattering of the solute. This outcome is reliable with the contentions of Padma and Ramana Rao [25] and Hayat et al. [32].

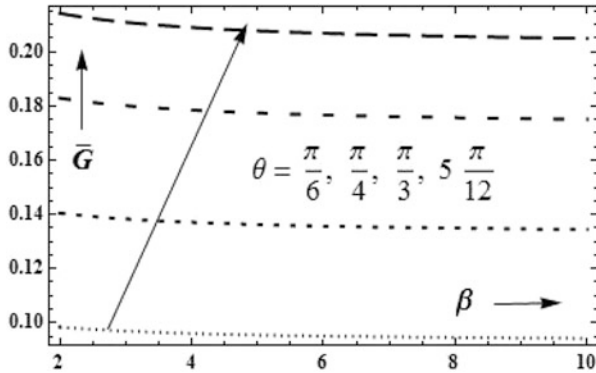


Fig. 6 Illustration of angle of inclination (θ) with scattering coefficient (\bar{G}) when $\mathcal{E}_1 = 0.1, \mathcal{E}_2 = 0.0, \mathcal{E}_3 = 0.00, \epsilon = 0.2, \alpha = 1.0, \mathcal{M} = 5.5,$ and $\gamma = 6.0$.

Fig. 7 Illustration of angle of inclination (θ) with scattering coefficient (\bar{G}) when $\mathcal{E}_1 = 0.1, \mathcal{E}_2 = 0.0, \mathcal{E}_3 = 0.06, \epsilon = 0.2, \beta = 5.0, \mathcal{M} = 5.5,$ and $\gamma = 6.0$.

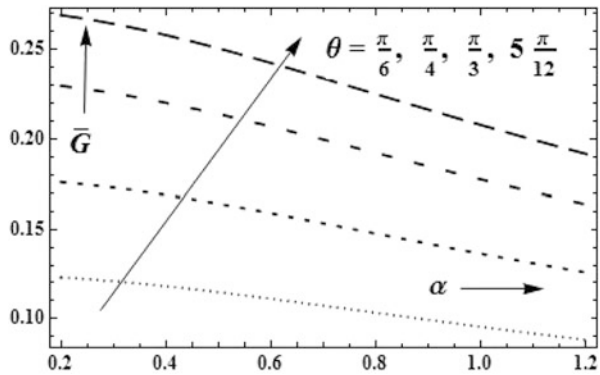
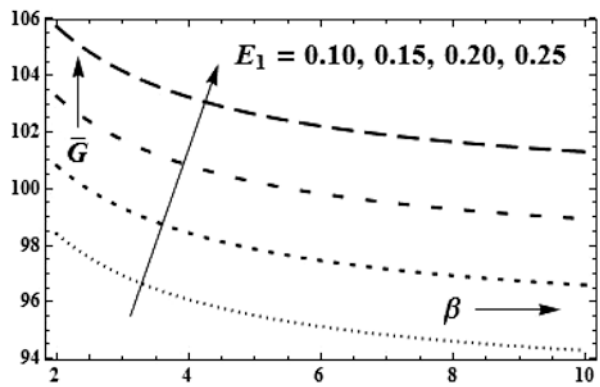


Fig. 8 Illustration of rigidity (\mathcal{E}_1) with scattering coefficient (\bar{G}) when $\mathcal{E}_2 = 4.0, \mathcal{E}_3 = 0.00, \epsilon = 0.2, \alpha = 1.0, \gamma = 6.0, \mathcal{M} = 5.5,$ and $\theta = \pi/6$.



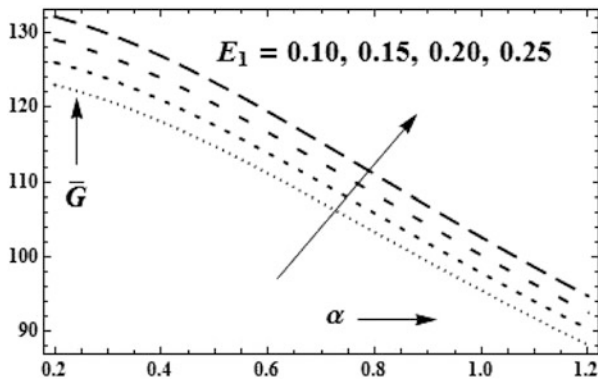


Fig. 9 Illustration of rigidity (\mathcal{E}_1) with scattering coefficient ($\bar{\mathcal{G}}$) when $\mathcal{E}_2 = 4.0, \mathcal{E}_3 = 0.06, \epsilon = 0.2, \beta = 5.0, \gamma = 6.0, \mathcal{M} = 5.5,$ and $\theta = \pi/6$.

Fig. 10 Illustration of stiffness (\mathcal{E}_2) with scattering coefficient ($\bar{\mathcal{G}}$) when $\mathcal{E}_1 = 0.1, \mathcal{E}_3 = 0.00, \epsilon = 0.2, \alpha = 1.0, \gamma = 6.0, \mathcal{M} = 5.5,$ and $\theta = \pi/6$.

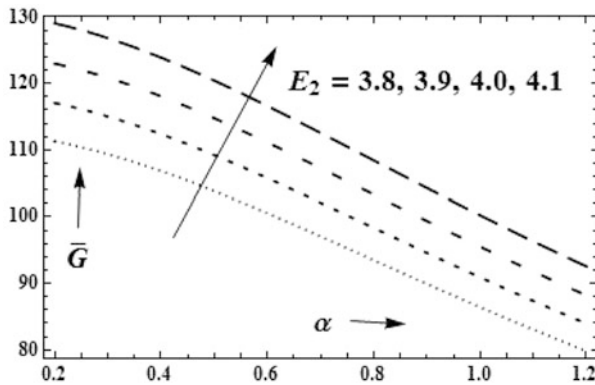
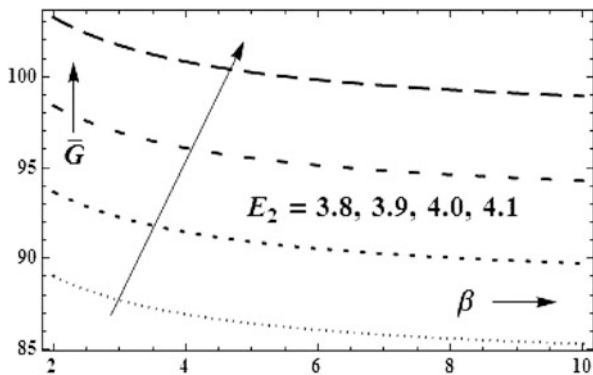


Fig. 11 Illustration of stiffness (\mathcal{E}_2) with scattering coefficient ($\bar{\mathcal{G}}$) when $\mathcal{E}_1 = 0.1, \mathcal{E}_3 = 0.06, \epsilon = 0.2, \beta = 5.0, \gamma = 6.0, \mathcal{M} = 5.5,$ and $\theta = \pi/6$.

Fig. 12 Illustration of damping force of the wall (\mathcal{E}_3) with scattering coefficient ($\bar{\mathcal{G}}$) when $\mathcal{E}_1 = 0.1$, $\mathcal{E}_2 = 0.00$, $\epsilon = 0.2$, $\alpha = 1.0$, $\gamma = 6.0$, $\mathcal{M} = 5.5$, and $\theta = \pi/6$.

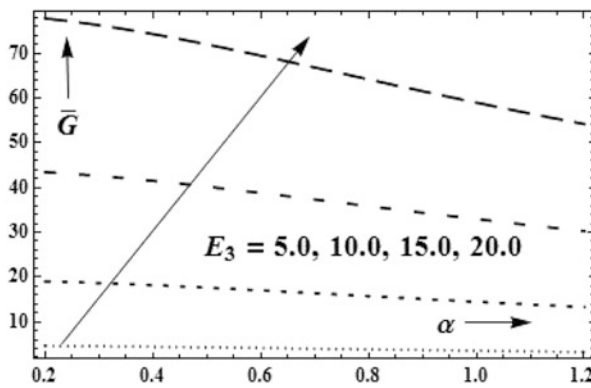
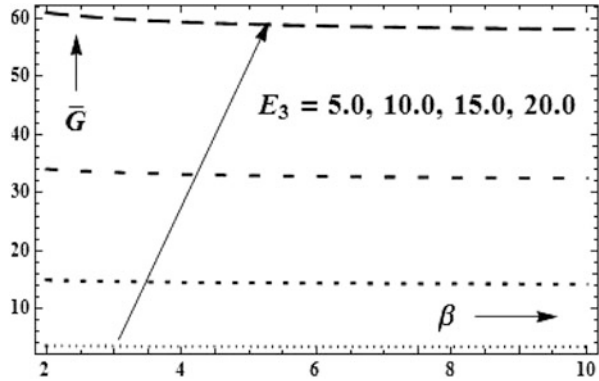


Fig. 13 Illustration of damping force of the wall (\mathcal{E}_3) with scattering coefficient ($\bar{\mathcal{G}}$) when $\mathcal{E}_1 = 0.1$, $\mathcal{E}_2 = 0.06$, $\epsilon = 0.2$, $\beta = 5.0$, $\gamma = 6.0$, $\mathcal{M} = 5.5$, and $\theta = \pi/6$.

5 Conclusions

The effects of magnetic constraint (\mathcal{M}), couple stress constraint (γ), angle of inclination (θ), homogeneous response rate (α), heterogeneous response rate (β), rigidity (\mathcal{E}_1), stiffness (\mathcal{E}_2), and damping characteristic (\mathcal{E}_3) of the wall on scattering coefficient ($\bar{\mathcal{G}}$) have been inspected for peristaltic movement of a couple stress fluid. It is of great importance for the movement of blood in artery, bolus in esophagus, bile in bile duct, and chyme in small intestine of the digestive system.

- It is seen that the concentration profile ($\bar{\mathcal{G}}$) rises with an increase in inclination of the channel and wall features.
- It is noticed that concentration profile ($\bar{\mathcal{G}}$) descends with rise in heterogeneous response rate, homogeneous response rate, couple stress, and magnetic constraints.

- Finally, rigidity (\mathcal{E}_1), stiffness (\mathcal{E}_2), damping force (\mathcal{E}_3) of the wall, and angle of inclination (θ) favor the dispersion, while couple stress constraint (γ), magnetic constraint (\mathcal{M}), homogeneous response rate constraint (α), and heterogeneous response rate constraint (β) resist the dispersion.

References

1. Jaffrin, M. Y., Shapiro, A. H., Weinberg, S. L.: Peristaltic pumping with long wavelengths at low Reynolds number. *J. Fluid Mech.* **37**, 799–825 (1969)
2. Fung, Y. C., Yih, C. S.: Peristaltic transport, *ASME J. Appl. Mech.* **35(4)**, 669–675 (1968)
3. Misra, J. C., Ghosh, S. K.: A mathematical model for the study of blood flow through a channel with permeable walls. *Acta Mech.* **122**, 137–153 (1997)
4. Jayarami Reddy, B., Subba Reddy, M. V., Nadhamuni Reddy, C., Yogeswar Reddy, P.: Peristaltic flow of a Williamson fluid in an inclined planar channel under the effect of magnetic field. *Adv. Appl. Sci. Res.* **3(1)**, 452–461 (2012)
5. Ibrahim, A., Abdulhadi, A.M.: Peristaltic transport of a viscoelastic fluid with fractional Maxwell model in an inclined channel. *Qi J. Sci.* **4(B)**, 1962–1975 (2014)
6. Kavitha, A., Reddy, H. R., Saravana, R., Sreenath, R.: Peristaltic transport of a Jeffrey fluid in contact with a Newtonian fluid in an inclined Channel. *Ain Sham Engg. J.* **8**, 683–687 (2017)
7. Stokes, V. K.: Couple stresses in fluids. *Phy. Fluids.* **25**, 1709–1715 (1966)
8. Srivastava, L. M.: Peristaltic transport of a couple stress fluid. *Rheol. Acta.* **25**, 638–641 (1986)
9. Elshehawey, E. F., Mekheimer, Kh. S.: Couple stress in peristaltic transport of fluids. *J. Phy. D.* **27**, 1163–1170 (1994)
10. Mekheimer, Kh. S.: Peristaltic flow of a couple stress fluid in a non-uniform channels. *J. Biorheol.* **39(6)**, 755–765 (2002)
11. Mekheimer, Kh. S.: Effect of the induced magnetic field on peristaltic flow of a couple stress fluid. *Phy. Lett. A.* **373(23)**, 4271–4278 (2008)
12. Ali, N., Hayat, T., Sajid, M.: Peristaltic flow of a couple stress fluid in asymmetric channel. *J. Biorheol.* **44(2)**, 125–138 (2007)
13. Hayat, T., Sajjad, R., Alsaedi, A., Muhammad, T., Ellahi, R.: On squeezing flow of couple stress nanofluid between two parallel plates. *Results in Phy.* **7**, 553–561 (2017)
14. Mittra, T. K., Prasad, N. S.: On the influence of wall properties and Poiseuille flow in Peristalsis. *J. Biomech.* **6**, 681–693 (1973)
15. Sankad, G. C., Radhakrishnamacharya, G.: Effect of magnetic Field on the peristaltic transport of couple stress fluid in a channel with wall effects, *Int. J. Biomath.* **4(3)**, 365–378 (2011)
16. Pandey, S. K., Chaube, M. K.: Study of wall properties on peristaltic transport of a couple stress fluid. *Meccanica.* **46**, 1319–1330 (2011)
17. Hina, S., Mustafa, M., Hayat, T.: On the exact solution for peristaltic flow of couple stress fluid with wall properties. *Bulgarian Chem. Commun.* **47(1)**, 30–37 (2015)
18. Mekheimer, Kh. S.: Effect of the induced magnetic field on the peristaltic flow of a couple stress. *Phy. Lett. A.* **372**, 4271–4278 (2008)
19. Sankad, G. C., Radhakrishnamacharya, G.: Effect of Magnetic Field on Peristaltic Motion of Micropolar Fluid with Wall Effects. *J. Appl. Math. Fluid Mech.* **1**, 37–50 (2009)
20. Sankad, G. C., Radhakrishnamacharya, G.: Influence of wall properties on the peristaltic transport of a micropolar fluid in an inclined channel. *J. New Results Sci.* **6**, 62–75 (2014)
21. Rathod, V. P., Manikrao, P., Sridhar, N. G.: Peristaltic flow of a couple stress fluid in an inclined channel under the effect of magnetic field. *Pelagia Research Libr.* **6(9)**, 101–109 (2015)
22. Tripathi, D., Beg, O.A.: Magnetohydrodynamic peristaltic flow of a couple stress fluid through coaxial channels containing a porous medium. *J. Mech. Med. Biol.* **12(5)**, 1250088-1-20 (2012)

23. Ng, N. O.: Dispersion in steady and oscillatory flows through a tube with reversible and irreversible wall reactions. *Proc. Roy. Soc. Lond.* **463(A)**, 481–515 (2006)
24. Taylor, G. I.: Dispersion of soluble matter in solvent flowing slowly through a tube. *Proc. Roy. Soc. Lond.* **219(A)**, 186–203 (1953)
25. Padma, D., Ramana Rao, V. V.: Homogeneous and heterogeneous reaction on the dispersion of a solute in MHD Couette flow I. *Curr. Sc.* **44**, 803–804 (1975)
26. Gupta, P. S., Gupta, A. S.: Effect of homogeneous and heterogeneous reactions on the dispersion of a solute in the laminar flow between two plates. *Proc. Roy. Soc. Lond.* **330(A)**, 59–63 (1972)
27. Sankad, G., Dhange, M.: Peristaltic pumping of an incompressible viscous fluid in a porous medium with wall effects and chemical reactions. *Alex. Engg. J.* **55**, 2015–2021 (2016)
28. Chandra, P., Philip, D.: Effect of heterogeneous and homogeneous reactions on the dispersion of a solute in simple microfluid. *Indian J. Pure Appl. Math.* **24**, 551–561 (1993)
29. Alemayehu, H., Radhakrishnamacharya, G.: Dispersion of solute in peristaltic motion of a couple stress fluid through a porous medium. *Tamkang J. Math.* **43(4)**, 541–555 (2012)
30. Ravi Kiran, G., Radhakrishnamacharya, G.: Peristaltic flow and hydrodynamic dispersion of a reactive micropolar fluid - simulation of chemical effects in the digestive process. *J. Mech. Med. Biol.* **17(1)** (2017), 1750013–1750030, <https://doi.org/10.1142/S0219519417500130>.
31. Hayat, T., Sajjad, R., Ellahi, R., Alsaedi, A., Muhammad, T.: Homogeneous-heterogeneous reactions in MHD flow of micropolar fluid by a curved stretching surface. *J. Mol. Liquid.* **240**, 209–220 (2017)
32. Hayat, T., Tanveer, A., Yasmin, H., Alsaedi, A.: Homogeneous-heterogeneous reactions in peristaltic flow with convective conditions. *PLOS one*, **9(12)** (2014), <https://doi.org/10.1371/journal.pone.0113851>.
33. Hayat, T., Tanveer, A., Alsaedi, A.: Mixed convective peristaltic flow of Carreau-Yasuda fluid with thermal deposition and chemical reaction. *Int. J. Heat and Mass Trans.* **96**, 474–481 (2016)
34. Abbas, Z., Jafer, H., Muhammad, S.: Hydromagnetic mixed convective two-phase flow of couple stress and viscous fluids in an inclined channel. *The Smithsonian/NASA Astrophysics Data System.* **69**, 553–561 (2014)
35. Abbas, Z., Sheikh, M.: Stagnation-point flow of a hydromagnetic viscous fluid over stretching/shrinking sheet with generalized slip condition in the presence of homogeneous-heterogeneous reactions. *Taiwan J. Chem. Engg.* **55**, 69–75 (2015)
36. Abbas, Z., Sheikh, M.: Numerical study of homogeneous-heterogeneous reactions on stagnation point flow of ferrofluid with non-linear slip condition. *Chinese J. Chem. Engg.* **25(1)**, 11–17 (2017)

A Fractional Inverse Initial Value Problem



Amin Boumenir and Vu Kim Tuan

To the memory of our friend Fadhel Al-Musallam, Kuwait University

2000 Mathematics Subject Classification 26A33, 34K29, 35R30

1 Introduction

Consider the inverse problem of reconstructing the coefficients $\{\alpha, k, \ell, p(x)\}$, appearing in the fractional evolution equation, with $0 < \alpha < 1$:

$$\begin{cases} {}_0^C \mathcal{D}_t^\alpha u(x, t) = \partial_x (p(x) \partial_x u(x, t)) & \text{for } 0 < x < \ell < \infty, \quad t > 0, \\ \lim_{x \rightarrow 0^+} p(x) \partial_x u(x, t) = 0, \quad t > 0, \\ k \lim_{x \rightarrow \ell^-} p(x) \partial_x u(x, t) + u(\ell, t) = 0, \quad t > 0, \quad k > 0, \\ u(x, 0) = a(x), \quad 0 < x < \ell, \end{cases} \quad (1)$$

from readings of $u(0, t)$ and $u(\ell, t)$, and where ${}_0^C \mathcal{D}_t^\alpha f(t)$ denotes the Caputo derivative [12]:

$${}_0^C \mathcal{D}_t^\alpha f(t) = \int_0^t \frac{(t - \xi)^{-\alpha}}{\Gamma(1 - \alpha)} f'(\xi) d\xi, \quad 0 < \alpha < 1.$$

A. Boumenir · V. K. Tuan (✉)

Department of Mathematics, University of West Georgia, Carrollton, GA 30118, USA
e-mail: boumenir@westga.edu; vu@westga.edu

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_24

387

A good survey of anomalous diffusion phenomena in environmental engineering leading to the fractional evolution equation (1) can be found in [5, 10, 11]. Under rather restrictive conditions $p \in C^2(0, \ell)$ and $p(x) > 0$, J. Cheng et al. [5] proved the uniqueness of a weak solution from its observation $u(0, t)$, $0 < t < T$, generated by a distributional initial condition, namely $u(x, 0) = \delta(x)$, by rewriting the Sturm-Liouville operator in the impedance form into the standard form $\frac{d^2}{dx^2} - q(x)$, and using the Gelfand-Levitan inverse spectral theory. In [13], the Neumann fractional diffusion equation in the standard form ${}_0^C \mathcal{D}_t^\alpha u = u_{xx} - q(x)u$, $u_x(0, t) = u_x(\ell, t) = 0$, has been considered under classical initial conditions. It was proved that the observations $\{u(0, t), u(\ell, t)\}_{0 < t < T}$, under finitely many specially chosen initial conditions $a(x) \in L^2(0, \ell)$, determine $q(x)$ uniquely, and a constructive reconstruction algorithm was proposed.

In this paper, we offer a different and new treatment to problem (1) where our emphasis is on the reconstruction of p under minimal restrictions:

$$\frac{1}{p} \in L(0, \ell), \quad p(x) \geq 0, \quad \text{and } a \in L^2(0, \ell). \tag{2}$$

This would allow engineers and practitioners in the field of applied control theory to work with classical solutions. We point out that the Dirac delta function δ means concentrating finite energy at a single point and may not be appropriate to use in nondestructive testing or noninvasive monitoring. This is the main reason why engineers would avoid working with distributions and weak solutions. However, it greatly simplifies the mathematical treatment as we shall explain at the end of Section 2. Although we make measurements at both end points, we assume that we do not know the distance ℓ between them.

Note that condition (2) would prevent the use of the Gelfand-Levitan theory as the Liouville transformation is not applicable. To circumvent this issue, we call on the M.G. Krein inverse spectral theory of the string, and we refer the reader to [7]. Also since α is unknown, we cannot use the Laplace transform to recast the problem into a standard heat equation. We are concerned with the following question.

Statement of the Problem

Reconstruct $\{\alpha, k, \ell, p(x)\}$ uniquely by a single measurement $a \rightarrow \{u(0, t), u(\ell, t)\}_{0 < t < T}$.

2 Direct Problem

Equation (1) can be written as a fractional evolution equation:

$${}_0^C \mathcal{D}_t^\alpha u = Au,$$

where A is the Sturm-Liouville operator in the impedance form:

$$\begin{cases} Ay(x) = -(p(x)y'(x))' & \text{for } 0 < x < \ell, \\ B_0(y) := \lim_{x \rightarrow 0^+} p(x)y'(x) = 0 \quad \text{and} \quad B_\ell(y) := k \lim_{x \rightarrow \ell^-} p(x)y'(x) + y(\ell) = 0, & k > 0. \end{cases} \tag{3}$$

The operator A acts in $L^2(0, \ell)$ and its domain is simply given by:

$$\begin{aligned} \text{Dom}(A) = & \left\{ y \in L^2(0, \ell) : y, py' \in AC(0, \ell), (py')' \in L^2(0, \ell) \right. \\ & \left. \text{and } B_0(y) = B_\ell(y) = 0 \right\}. \end{aligned}$$

The operator A under the conditions:

$$p^{-1} \in L(0, \ell) \text{ and } p(x) \geq 0 \tag{4}$$

is regular at both end points, $x = 0, \ell$, and the limits $\lim_{x \rightarrow 0^+} p(x)y'(x)$ and $\lim_{x \rightarrow \ell^-} p(x)y'(x)$ exist, see [14, section 2.3]. Here, the derivative y' exists almost everywhere, however $py'(x)$ is a continuous function which is called a quasi-derivative. Under the condition $k > 0$, it has a positive discrete spectrum, and generates a normalized eigenbasis, $\{\phi_n\}_{n \geq 1}$ say, where:

$$A\phi_n(x) = \lambda_n\phi_n(x),$$

with $\|\phi_n\| = 1$, and $\lambda_n \uparrow \infty$. Since for each $t > 0$, $u(\cdot, t) \in L^2(0, \ell)$, we then look for the coefficients $c_n(t)$ such that

$$u(x, t) = \sum_{n \geq 1} c_n(t) \phi_n(x), \quad \text{where } c_n(t) = \int_0^\ell u(x, t) \phi_n(x) dx. \tag{5}$$

Clearly, it follows from (5) and (1) that

$$\begin{aligned} {}_0^C \mathcal{D}_t^\alpha c_n(t) &= \int_0^\ell {}_0^C \mathcal{D}_t^\alpha u(x, t) \phi_n(x) dx = \int_0^\ell \partial_x (p(x) \partial_x u(x, t)) \phi_n(x) dx \\ &= \lim_{\varepsilon \rightarrow 0^+} \left[p(x) \partial_x u(x, t) \phi_n(x) - p(x) u(x, t) \phi_n'(x) \right] \Big|_\varepsilon^{\ell-\varepsilon} \\ &\quad + \int_0^\ell u(x, t) (p(x) \phi_n'(x))' dx \\ &= -\lambda_n c_n(t). \end{aligned}$$

So, we have

$${}^C_0\mathcal{D}_t^\alpha c_n(t) = -\lambda_n c_n(t) \quad \text{and} \quad c_n(0) = \int_0^\ell a(x)\phi_n(x)dx. \tag{6}$$

Equation (6) has the solution [13]:

$$c_n(t) = c_n(0) E_\alpha(-\lambda_n t^\alpha),$$

where $E_\alpha(x)$ is the Mittag-Leffler function [12]:

$$E_\alpha(x) := \sum_{k=0}^\infty \frac{x^k}{\Gamma(\alpha k + 1)}.$$

Therefore, we have the following representation of the classical solution:

$$u(x, t) = \sum_{n \geq 1} c_n(0) E_\alpha(-\lambda_n t^\alpha) \phi_n(x). \tag{7}$$

From the asymptotics of the Mittag-Leffler function [8]:

$$E_\alpha(-\lambda_n t^\alpha) = O\left(\frac{1}{\lambda_n t^\alpha}\right), \quad t > 0, \quad n \rightarrow \infty, \tag{8}$$

and $\lambda_n \uparrow \infty$, $\{c_n(0)\}_{n \geq 1} \in \ell^2$, we see that for each $t > 0$, the series (7) converges in $L^2(0, \ell)$. For the Neumann Sturm-Liouville operator in the standard form $\frac{d^2}{dx^2} - q(x)$, the eigenvalues λ_n have the asymptotics $\lambda_n \sim (\frac{\pi}{\ell}(n - 1))^2$, hence the uniform convergence of the series (7) on $[0, \ell]$ is easy to derive. For operators in the impedance form (3), no asymptotic formula for λ_n is known, so the uniform convergence of (7) in x is not obvious. We will prove in the next section that, for any $t > 0$, the series (7) converges absolutely and uniformly in x on $[0, \ell]$, and therefore, the measurements or readings at $x = 0$ and $x = \ell$ can be expressed as:

$$\begin{aligned} u(0, t) &= \sum_{n \geq 1} c_n(0) \phi_n(0) E_{\alpha,1}(-\lambda_n t^\alpha) \quad \text{and} \\ u(\ell, t) &= \sum_{n \geq 1} c_n(0) \phi_n(\ell) E_{\alpha,1}(-\lambda_n t^\alpha), \quad t > 0. \end{aligned} \tag{9}$$

The task now is to:

1. Prove the uniform convergence of the series (7) in x on $[0, \ell]$ by finding an estimate on $\{\phi_n(x)\}_{n \geq 1}$.
2. Extract the complete spectral data from (9) by showing that all $c_n(0)$, $\phi_n(0)$, $\phi_n(\ell) \neq 0$ for $a(x) = 1$.

Remark We now explain the main advantage of using the Dirac delta function as an initial condition [5]. Normalize the eigenfunctions by $\phi_n(0) = 1$. The Fourier coefficients are

$$c_n(0) = \frac{1}{\|\phi_n\|_2^2} \int_0^\ell \delta(x)\phi_n(x)dx = \frac{\phi_n(0)}{\|\phi_n\|_2^2} = \frac{1}{\|\phi_n\|_2^2},$$

and so

$$u(x, t) = \sum_{n \geq 1} \frac{1}{\|\phi_n\|_2^2} E_\alpha(-\lambda_n t^\alpha) \phi_n(x), \tag{10}$$

and obviously, the observation at one end point $x = 0$ only

$$u(0, t) = \sum_{n \geq 1} \frac{1}{\|\phi_n\|_2^2} E_\alpha(-\lambda_n t^\alpha), \quad 0 < t < T,$$

already guarantees the presence of all the spectral data $\{\lambda_n, \|\phi_n\|_2\}_{n \geq 1}$. Unfortunately, this initial condition $a(x) = \delta(x)$ can hardly be used in practice and we show that we need to observe both end points, when we use classical functions.

3 Convergence

In this section, we collect some identities on eigenfunctions that would help us prove convergence of the series (7). Successive integration of the eigenfunction equation:

$$-(p(x)\phi'_n(x))' = \lambda_n \phi_n(x) \quad \text{with} \quad \lim_{x \rightarrow 0^+} p(x)\phi'_n(x) = 0$$

yields

$$-\frac{p(x)\phi'_n(x)}{\lambda_n} = \int_0^x \phi_n(\eta) d\eta = \int_0^\ell \chi_{[0,x]}(\eta)\phi_n(\eta)d\eta, \tag{11}$$

$$\frac{\phi_n(0) - \phi_n(x)}{\lambda_n} = \int_0^x \phi_n(\eta) \int_\eta^x \frac{1}{p(t)} dt d\eta = \int_0^\ell K(x, \eta)\chi_{[0,x]}(\eta)\phi_n(\eta)d\eta, \tag{12}$$

where $\chi_{[0,x]}(\eta)$ is the characteristic function of the interval $[0, x]$, and

$$K(x, \eta) := \int_\eta^x \frac{1}{p(t)} dt \tag{13}$$

is continuous in $0 \leq \eta \leq x \leq \ell$ by (2) and clearly, $0 \leq K(x, \eta) \leq K(\ell, \eta)$.

Taking the limit when $x \rightarrow \ell -$ in (11) and recalling that

$$\phi_n(\ell) = -k \lim_{x \rightarrow \ell -} p(x)\phi'_n(x), \tag{14}$$

we obtain

$$\frac{\phi_n(\ell)}{k\lambda_n} = \int_0^\ell \phi_n(\eta)d\eta. \tag{15}$$

Formula (11) says that $-\frac{p(x)\phi'_n(x)}{\lambda_n}$ is the Fourier coefficient of the characteristic function $\chi_{[0,x]}(\eta)$ of the interval $[0, x]$. Parseval's formula for the Fourier series of $\chi_{[0,x]}(\eta)$ then yields

$$p^2(x) \sum_{n \geq 1} \frac{\phi_n'^2(x)}{\lambda_n^2} = x. \tag{16}$$

Applying Parseval's formula for the Fourier series of $f(x) = 1$ and using formula (15), we arrive at

$$\sum_{n \geq 1} \frac{\phi_n^2(\ell)}{\lambda_n^2} = k^2l < \infty. \tag{17}$$

Consequently, $\left\{ \frac{\phi_n(\ell)}{\lambda_n} \right\}_{n \geq 1} \in \ell^2$.

Combining (11) and (12), we have

$$\frac{\phi_n(0) - \phi_n(x) - kp(x)\phi'_n(x)}{\lambda_n} = \int_0^x [K(x, \eta) + k]\phi_n(\eta)d\eta.$$

Taking the limit when $x \rightarrow \ell -$ and using (14), we arrive at

$$\frac{\phi_n(0)}{\lambda_n} = \int_0^\ell [K(\ell, \eta) + k]\phi_n(\eta)d\eta. \tag{18}$$

Formula (18) says that $\frac{\phi_n(0)}{\lambda_n}$ is the Fourier coefficient of the function $K(\ell, \eta) + k$, and Parseval's formula for the Fourier series of $K(\ell, \eta) + k$ then yields

$$\sum_{n \geq 1} \frac{\phi_n^2(0)}{\lambda_n^2} = \int_0^\ell (K(\ell, \eta) + k)^2 d\eta < \infty. \tag{19}$$

Hence, $\left\{ \frac{\phi_n(0)}{\lambda_n} \right\}_{n \geq 1} \in \ell^2$.

Similarly, formula (12) says that $\frac{\phi_n(0) - \phi_n(x)}{\lambda_n}$ is the Fourier coefficient of $K(x, \eta)\chi_{[0,x]}(\eta)$, and Parseval’s formula for the Fourier series of $K(x, \eta)\chi_{[0,x]}(\eta)$ then yields

$$\sum_{n \geq 1} \frac{[\phi_n(0) - \phi_n(x)]^2}{\lambda_n^2} = \int_0^x K^2(x, \eta)d\eta \leq \int_0^\ell K^2(\ell, \eta)d\eta. \tag{20}$$

Combining (19) and (20), we arrive at

$$\begin{aligned} \sum_{n \geq 1} \frac{\phi_n^2(x)}{\lambda_n^2} &\leq 2 \sum_{n \geq 1} \frac{\phi_n^2(0)}{\lambda_n^2} + 2 \sum_{n \geq 1} \frac{[\phi_n(0) - \phi_n(x)]^2}{\lambda_n^2} \\ &\leq 2 \int_0^\ell [K^2(\ell, \eta) + (K(\ell, \eta) + k)^2] d\eta < \infty, \quad x \in [0, \ell]. \end{aligned} \tag{21}$$

The bound of $\sum_{n \geq 1} \frac{\phi_n^2(x)}{\lambda_n^2}$ in (21) is independent of $x \in [0, \ell]$. Together with (8) and $\{c_n(0)\}_{n \geq 1} \in \ell^2$, it yields the absolute and uniform convergence of the series (7) in x on $[0, \ell]$. Hence, for any fixed $x \in [0, \ell]$, $u(x, t)$ is analytic in $t > 0$, and the representations at the boundaries (9) hold.

We show now that

$$\phi_n(0) \neq 0, \quad n \geq 1. \tag{22}$$

Otherwise, integrating from $x = 0$ and using that $\lim_{x \rightarrow 0^+} p(x)\phi'_n(x) = 0$ we would get the following Volterra integral equation for ϕ_n :

$$\begin{aligned} \phi_n(x) &= \phi_n(0) + \lim_{x \rightarrow 0^+} p(x)\phi'_n(x) \int_0^x \frac{1}{p(t)} dt - \lambda_n \int_0^x \phi_n(\eta) \int_\eta^x \frac{1}{p(t)} dt d\eta \\ &= -\lambda_n \int_0^x K(x, \eta) \phi_n(\eta) d\eta, \quad x \in [0, \ell]. \end{aligned} \tag{23}$$

The Volterra integral equation (23) with the continuous kernel $K(x, \eta)$ would have only the trivial solution $\phi_n(x) = 0$, that is impossible. Thus, (22) holds.

Similarly, one can show that

$$\phi_n(\ell) \neq 0, \quad n \geq 1. \tag{24}$$

Otherwise, it would imply, by (14), that $\lim_{x \rightarrow \ell^-} p(x)\phi'_n(x) = 0$ and integrating from $x = \ell$, we would get the following Volterra integral equation for ϕ_n :

$$\begin{aligned} \phi_n(x) &= \phi_n(\ell) + \lim_{x \rightarrow \ell^-} p(x)\phi_n'(x) \int_{\ell}^x \frac{1}{p(t)} dt + \lambda_n \int_{\ell}^x \phi_n(\eta) \int_x^{\eta} \frac{1}{p(t)} dt d\eta \\ &= -\lambda_n \int_x^{\ell} K(\eta, x) \phi_n(\eta) d\eta, \quad x \in [0, \ell]. \end{aligned} \tag{25}$$

The Volterra integral equation (25) with the continuous kernel $K(\eta, x)$ would have only the trivial solution $\phi_n(x) = 0$, that is impossible. Thus, (24) holds, and it follows from (15) and (24) that

$$\lambda_n \int_0^{\ell} \phi_n(\eta) d\eta = \frac{\phi_n(\ell)}{k} \neq 0,$$

which implies that both

$$\lambda_n \neq 0 \quad \text{and} \quad \int_0^{\ell} \phi_n(\eta) d\eta \neq 0 \quad \text{for all } n \geq 1. \tag{26}$$

Recall that when $a(x) = 1$

$$c_n(0) = (1, \phi_n) = \int_0^{\ell} \phi_n(\eta) d\eta = \frac{\phi_n(\ell)}{k\lambda_n} \neq 0. \tag{27}$$

Thus, we have proved.

Proposition 1 *We have the following:*

- a) $\phi_n(0), \phi_n(\ell) \neq 0$ for any $n \geq 1$.
- b) For any $a(x) \in L^2(0, \ell)$ and for each $t > 0$ the series (7) converges absolutely and uniformly in x on $[0, \ell]$, and for each fixed $x \in [0, \ell]$ the series (7) is an analytic function in $t > 0$. In particular, the series (9) converge absolutely and represent analytic functions in $t > 0$.
- c) If $a(x) = 1$, then $c_n(0) \neq 0$ for any $n \geq 1$, and $u(0, t)$ and $u(\ell, t)$ both contain all the boundary spectral data $\{\lambda_n, \phi_n(0), \phi_n(\ell)\}_{n \geq 1}$.

4 The Data Processing

Due to Proposition 1c) from now on we take the special initial condition $a(x) = 1$. Representations (9) are the key to solving our inverse problem. First, since $c_n(0), \phi_n(0), \phi_n(\ell) \neq 0$ for any $n \geq 1$, the series (9) contain all eigenvalues λ_n , which are crucial for the completeness of the spectral data. We now explain how to extract the fractional order α of the equation, the eigenvalues λ_n , and the boundary values $\phi_n(0), \phi_n(\ell)$, from the pair $\{u(0, t), u(\ell, t)\}_{0 < t < T}$. The analyticity of $u(0, t)$

and $u(\ell, t)$ means that the observations on $(0, T)$ can be extended uniquely on $(0, \infty)$, so we will assume that we observe $u(0, t)$ and $u(\ell, t)$ for any $t > 0$.

Applying the Laplace transform:

$$F(s) = \mathcal{L}(f)(s) := \int_0^\infty e^{-st} f(t) dt, \tag{28}$$

to (9), using [8, formula (3.7.7)]

$$\mathcal{L}(E_\alpha(-\lambda t^\alpha))(s) = \frac{s^{\alpha-1}}{s^\alpha + \lambda},$$

we obtain

$$\begin{aligned} U(0, s) &= \mathcal{L}(u(0, t))(s) = s^{\alpha-1} \sum_{n \geq 1} \frac{c_n(0) \phi_n(0)}{s^\alpha + \lambda_n}, \\ U(\ell, s) &= \mathcal{L}(u(\ell, t))(s) = s^{\alpha-1} \sum_{n \geq 1} \frac{c_n(0) \phi_n(\ell)}{s^\alpha + \lambda_n}. \end{aligned} \tag{29}$$

4.1 Finding α

From (29), we have

$$U(0, s) - U(\ell, s) = s^{\alpha-1} \sum_{n \geq 1} \frac{c_n(0)}{s^\alpha + \lambda_n} \{\phi_n(0) - \phi_n(\ell)\}.$$

Consequently:

$$\ln(|U(0, s) - U(\ell, s)|) = (\alpha - 1) \ln(s) + \ln \left| \sum_{n \geq 1} \frac{c_n(0)}{s^\alpha + \lambda_n} \{\phi_n(0) - \phi_n(\ell)\} \right|. \tag{30}$$

By Parseval’s equality on (6) and (12), we have by (13)

$$\begin{aligned} &\lim_{s \rightarrow 0^+} \sum_{n \geq 1} \frac{c_n(0)}{s^\alpha + \lambda_n} \{\phi_n(0) - \phi_n(\ell)\} \\ &= \sum_{n \geq 1} c_n(0) \frac{(\phi_n(0) - \phi_n(\ell))}{\lambda_n} = \int_0^\ell K(\ell, \eta) d\eta > 0. \end{aligned}$$

Taking the limits as $s \rightarrow 0^+$ in (30) yields α

$$\alpha = 1 + \lim_{s \rightarrow 0^+} \frac{\ln |U(0, s) - U(\ell, s)|}{\ln s}. \tag{31}$$

4.2 Finding λ_n and $\frac{\phi_n(\ell)}{\phi_n(0)}$

Once α is known, the formulas:

$$s^{\frac{1}{\alpha-1}} U(0, s^{\frac{1}{\alpha}}) = \sum_{n \geq 1} \frac{c_n(0) \phi_n(0)}{s + \lambda_n}, \quad s^{\frac{1}{\alpha-1}} U(\ell, s^{\frac{1}{\alpha}}) = \sum_{n \geq 1} \frac{c_n(0) \phi_n(\ell)}{s + \lambda_n}, \tag{32}$$

say that $-\lambda_n$ are simply the poles of $s^{\frac{1}{\alpha-1}} U(0, s^{\frac{1}{\alpha}})$ and $s^{\frac{1}{\alpha-1}} U(\ell, s^{\frac{1}{\alpha}})$ with the corresponding residues $c_n(0) \phi_n(0)$ and $c_n(0) \phi_n(\ell)$. Thus, (32) would reveal all eigenvalues λ_n and the values $\frac{\phi_n(\ell)}{\phi_n(0)}$ that are necessary in Krein’s string inverse problem. However, we will employ the method of limits, developed in [3, 4, 13], to explicitly determine λ_n and $\frac{\phi_n(\ell)}{\phi_n(0)}$. Denote

$$p_1(t) = \mathcal{L}^{-1} \left(s^{\frac{1}{\alpha-1}} U(0, s^{\frac{1}{\alpha}}) \right) (t), \quad \tilde{p}_1(t) = \mathcal{L}^{-1} \left(s^{\frac{1}{\alpha-1}} U(\ell, s^{\frac{1}{\alpha}}) \right) (t),$$

then $p_1(t), \tilde{p}_1(t)$ are known from the measurements $u(0, t), u(\ell, t)$. Clearly:

$$p_1(t) = \sum_{n \geq 1} c_n(0) \phi_n(0) e^{-\lambda_n t}, \quad \tilde{p}_1(t) = \sum_{n \geq 1} c_n(0) \phi_n(\ell) e^{-\lambda_n t}. \tag{33}$$

Since the eigenvalues are increasing, $\lambda_1 < \lambda_2 < \dots < \lambda_n < \dots$, then for large k , we have $p_1(k) \sim c_1(0) \phi_1(0) e^{-\lambda_1 k}$ and so the limit:

$$\lim_{k \rightarrow \infty} \frac{p_1(k)}{p_1(k+1)} = \lim_{k \rightarrow \infty} \frac{e^{-\lambda_1 k}}{e^{-\lambda_1(k+1)}} = e^{\lambda_1} \Rightarrow \lambda_1 = \lim_{k \rightarrow \infty} \ln \left(\frac{p_1(k)}{p_1(k+1)} \right). \tag{34}$$

We can also recover $c_1(0) \phi_1(0)$ and $c_1(0) \phi_1(\ell)$ by other limits:

$$\lim_{k \rightarrow \infty} p_1(k) e^{\lambda_1 k} = c_1(0) \phi_1(0), \quad \lim_{k \rightarrow \infty} \tilde{p}_1(k) e^{\lambda_1 k} = c_1(0) \phi_1(\ell), \tag{35}$$

since $\lambda_1 - \lambda_n < 0$ for all $n \geq 2$, and $\frac{\phi_1(\ell)}{\phi_1(0)} = \frac{c_1(0) \phi_1(\ell)}{c_1(0) \phi_1(0)}$. By removing the first terms $c_1(0) \phi_1(0) e^{-\lambda_1 k}$ and $c_1(0) \phi_1(\ell) e^{-\lambda_1 k}$ from the series defining the originals $p_1(k)$ and $\tilde{p}_1(k)$, we obtain new functions:

$$p_2(k) = p_1(k) - c_1(0) \phi_1(0) e^{-\lambda_1 k} = \sum_{n \geq 2} c_n(0) \phi_n(0) e^{-\lambda_n k},$$

$$\tilde{p}_2(k) = \tilde{p}_1(k) - c_1(0) \phi_1(\ell) e^{-\lambda_1 k} = \sum_{n \geq 2} c_n(0) \phi_n(\ell) e^{-\lambda_n k}.$$

To recover the next pair $\left(\lambda_2, \frac{\phi_2(\ell)}{\phi_2(0)}\right)$, we only need to repeat steps (34) and (35) to the new functions $p_2(k)$ and $\tilde{p}_2(k)$, and by doing so we can eventually determine all the sequence $\left\{\lambda_n, \frac{\phi_1(\ell)}{\phi_1(0)}\right\}_{n \geq 1}$.

We thus have proved.

Proposition 2 *Assume that the initial condition is $a(x) = 1$. Then, we can extract the complete boundary spectral data $\left\{\left(\lambda_n, \frac{\phi_n(\ell)}{\phi_n(0)}\right)\right\}_{n \geq 1}$.*

To avoid the costly Liouville transformation, which is not applicable under (2), we shall transmute the operator A into a similar string [2]. We refer to the theory as outlined in the book by Dym and McKean [7] for its beautiful connection with the Fourier analysis and de Branges spaces. We mention also the book by Atkinson, and the works by Coleman, McLaughlin, and Hald, McLaughlin [1, 6, 9].

5 M.G. Krein’s String

We now show that the operator A is similar to a string operator [7]. Denote by:

$$\tau(x) = \int_0^x \frac{1}{p(t)} dt, \quad 0 \leq x \leq \ell, \tag{36}$$

which is an increasing one-to-one function, $[0, \ell] \xrightarrow{\tau} [0, b]$, where $b = \tau(\ell)$. Next, define a nonnegative weight function w by:

$$w(\tau(x)) = p(x) \geq 0, \tag{37}$$

and a unitary operator $\mathbb{T} : L_w^2(0, b) \rightarrow L^2(0, \ell)$ by a composition operation $\mathbb{T}(h) = h \circ \tau$,

$$\begin{aligned} \|h\|_w^2 &= \int_0^b |h(s)|^2 w(s) ds = \int_0^\ell |h(\tau(x))|^2 p(x)\tau'(x) dx \\ &= \int_0^\ell |h(\tau(x))|^2 dx = \|\mathbb{T}(h)\|^2. \end{aligned}$$

To proceed, we need the following lemma.

Lemma 3 $w \in L(0, b)$.

Proof From (37), we have

$$\int_0^b w(\tau) d\tau = \int_0^\ell w(\tau(x))\tau'(x) dx = \int_0^\ell p(x) \frac{1}{p(x)} dx = \ell < \infty.$$

Since $w \in L(0, b)$, we can define the string operator:

$$\begin{cases} \mathbb{S}\psi(\tau) = \frac{-1}{w(\tau)} \frac{d^2}{d\tau^2} \psi(\tau), & 0 < \tau < b, \\ \psi'(0) = 0 & k\psi'(b) + \psi(b) = 0, \end{cases} \tag{38}$$

which models the vibration of a string whose mass between $(0, x]$ is $\int_0^x w(\eta)d\eta$, which is why w is called the density of the string. The operator \mathbb{S} acts in

$$L_w^2(0, b) = \left\{ f \text{ measurable} : \int_0^b |f(x)|^2 w(x)dx < \infty \right\},$$

and its domain is given by:

$$\text{Dom}(\mathbb{S}) = \left\{ \psi \in L_w^2(0, b) : \psi(\tau) = a + \int_0^\tau (t - \tau) g(t)w(t)dt, \right. \\ \left. \text{where } g \in L_w^2(0, b) \text{ and } k\psi'(b) + \psi(b) = 0 \right\}.$$

We now prove the following proposition.

Proposition 4 *Assume that (2) holds. Then, the operator A in (3) is similar to the string operator \mathbb{S} defined by (38) with a mass density w given by (37), length $b = \tau(\ell)$, and where τ is defined as in (36).*

Proof Define the function ψ by $\varphi = \mathbb{T}\psi$, where $\varphi \in \text{Dom}(A)$, that is:

$$\psi(\tau(x)) = \varphi(x). \tag{39}$$

Then, it follows from

$$\frac{d\psi}{d\tau}(\tau(x)) = p(x)\varphi'(x) \tag{40}$$

that

$$\frac{d^2\psi}{d\tau^2}(\tau(x)) = p(x) (p(x)\varphi'(x))', \text{ with } \psi'(0) = 0, \text{ and } k\psi'(b) + \psi(b) = 0.$$

Use (37) to write

$$\frac{-1}{w(\tau(x))} \frac{d^2}{d\tau^2} \psi(\tau(x)) = - (p(x)\varphi'(x))' \tag{41}$$

and so we deduce from (41) and (40) that

$$\mathbb{T}\mathbb{S}\psi = A\varphi = A\mathbb{T}\psi \quad \text{for any } \psi \in \text{Dom}(\mathbb{S}). \tag{42}$$

It is also easily seen from (40) that $\psi \in \text{Dom}(\mathbb{S})$ if and only if $\mathbb{T}\psi \in \text{Dom}(A)$ and (42) establishes that \mathbb{S} and A are similar. Thus, they have the identical spectral data. □

5.1 Finding $k, \ell,$ and w

Consider the solution to the initial value problem:

$$\begin{cases} -\psi''(\xi, \lambda) = \lambda w(\xi) \psi(\xi, \lambda), & 0 < \xi < b, \\ \psi(0, \lambda) = 1 \text{ and } \psi'(0, \lambda) = 0. \end{cases} \tag{43}$$

The solutions $\psi(b, \lambda)$ and $\psi'(b, \lambda)$ are entire functions of λ of order $1/2$, while the eigenvalues λ_n of \mathbb{S} , (38), are precisely the zeros of

$$k\psi'(b, \lambda_n) + \psi(b, \lambda_n) = 0.$$

Thus, we can write it as:

$$k\psi'(b, \lambda) + \psi(b, \lambda) = \gamma \prod_{n \geq 1} \left(1 - \frac{\lambda}{\lambda_n}\right), \tag{44}$$

where

$$\gamma = k\psi'(b, 0) + \psi(b, 0) = 1, \tag{45}$$

since $\psi(x, 0) = 1$, which is easily obtained by setting $\lambda = 0$ in (43). Next, we need the norming constants $\|\psi(\cdot, \lambda_n)\|$ which are obtained from the identity:

$$w\psi^2 = (\psi' \partial_\lambda \psi - \psi \partial_\lambda \psi')',$$

and the fact that $\partial_\lambda \psi'(0, \lambda_n) = \psi'(0, \lambda_n) = 0$ and $k\psi'(b, \lambda_n) + \psi(b, \lambda_n) = 0$. Thus, we deduce that

$$\begin{aligned} \alpha_n &:= \int_0^b \psi^2(\eta, \lambda_n) w(\eta) d\eta = \psi'(b, \lambda) \partial_\lambda \psi(b, \lambda) - \psi(b, \lambda_n) \partial_\lambda \psi'(b, \lambda_n) \tag{46} \\ &= -\psi(b, \lambda_n) \partial_\lambda [k\psi'(b, \lambda_n) + \psi(b, \lambda_n)] = -\frac{\phi_n(\ell)}{\phi_n(0)} \frac{\partial}{\partial \lambda} \prod_{j \geq 1} \left(1 - \frac{\lambda}{\lambda_j}\right) \Bigg|_{\lambda=\lambda_n}, \end{aligned}$$

where the last equation follows from (44) and the fact that $\psi(b, \lambda_n) = \frac{\phi_n(\ell)}{\phi_n(0)}$.

Having the complete spectral data $\{\lambda_n, \alpha_n\}_{n \geq 1}$ from (46), we can construct a unique spectral function:

$$\Gamma(\lambda) = \sum_{\lambda_n \leq \lambda} \frac{1}{\alpha_n}, \tag{47}$$

which then yields a unique mass $m(\xi) = \int_0^\xi w(\eta)d\eta$, by the well-known M.G. Krein inverse spectral theory [7, section 5.8, page 198] as long as

$$\sum_{n \geq 1} \frac{1}{(1 + \lambda_n) \alpha_n} < \infty, \tag{48}$$

and happens to be a necessary and sufficient condition for the existence of a string. The mass is obtained as a limit of a sequence of finite strings generated by continued fractions, each corresponding to a partial sum of the spectral function by the formula:

$$\int_0^{\lambda_n} \frac{\Gamma(\lambda)}{\lambda + b} d\lambda = \frac{1}{bm_0 + \frac{1}{x_1 - x_0 + \frac{1}{bm_1 + \frac{1}{x_2 - x_1 + \dots + \frac{1}{bm_n + \frac{1}{k_n}}}}}}, \tag{49}$$

where m_k is the approximate mass at x_k , and $x_0 = 0$ [7, page 205]. By letting $\lambda_n \rightarrow \infty$, condition (48) ensures the convergence of the sequence of finite strings defined by (49) to the sought string. Thus, the sought mass m is the limit of the finite step functions $\sum_{x_j \leq x \leq x_n} m_j$, where x_j, m_j are read off (49). Note that since we started with an absolutely continuous mass, then by the uniqueness of its recovery, its spectral data should yield back an absolutely continuous mass. Then, $w(\xi) = m'(\xi)$.

5.2 Finding p

Once the density w is obtained, we can reconstruct p by solving the functional equation (37), $w(\tau(x)) = p(x) = \frac{1}{\tau'(x)}$ for $0 \leq x \leq \ell$, simply by bringing the inverse function of τ , which is defined by $x(\tau(x)) = x$, or $x'(\tau) = \frac{1}{\tau'(x)}$. Thus, we have $x'(\tau) = w(\tau)$ with $x(0) = 0$ or

$$x(\tau) = \int_0^\tau w(\eta) d\eta,$$

which we can invert back to get $\tau(x)$ and then finally:

$$p(x) = w(\tau(x)).$$

5.3 Finding ℓ

The length b of the string in (38) is simply the last point of growth of the mass $m(x) = \int_0^x w(\eta)d\eta$, i.e., $\text{supp } w \subset [0, b]$. Having b and the map τ , we get

$$\ell = x(b) = \int_0^b w(\eta)d\eta.$$

5.4 Finding k

Using (17) and (27), we have

$$k = \frac{1}{\ell} \sum_{n \geq 1} \frac{\phi_n^2(\ell)}{k\lambda_n^2} = \frac{1}{\ell} \sum_{n \geq 1} \frac{c_n(0)\phi_n(\ell)}{\lambda_n}. \tag{50}$$

Since ℓ , $c_n(0)\phi_n(\ell)$, and λ_n have been found, we get k from (50). Thus, we have proved.

Theorem 5 Assume that $0 \leq p^{-1} \in L(0, b)$, $0 < k$, $0 < \alpha < 1$, and $a(x) = 1$. Then:

- a) One single measurement $\{u(0, t), u(\ell, t)\}_{0 < t < T}$ determines $\{\alpha, \ell, k, p(x)\}$ uniquely.
- b) The unknowns $\{\alpha, \ell, k, p(x)\}$ are reconstructed explicitly and uniquely from the single measurement of $\{u(0, t), u(\ell, t)\}_{t > 0}$.

References

1. F.V. Atkinson, *Discrete and Continuous Boundary Problems*, Volume 8, Mathematics in Science and Engineering: A Series of Monographs and Textbooks, Academic Press, 1964.
2. A. Boumenir, The Gelfand-Levitan theory for strings, *Topics in Operator Theory, Volume 2, Systems and Mathematical Physics*, 115–136, Oper. Theory Adv. Appl., 203, Birkhäuser Verlag, Basel, 2010.
3. A. Boumenir and Vu Kim Tuan, An inverse problem for the heat equation, *Proc. Amer. Math. Soc.* 138 (2010), 3911–3921.
4. A. Boumenir and Vu Kim Tuan, Recovery of the heat coefficient by two measurements, *Inverse Problems and Imaging* 5 (2011), no. 4, 775–791.
5. J. Cheng, Nakagawa, M. Yamamoto, and T. Yamazaki, Uniqueness in an inverse problem for a one-dimensional fractional diffusion equation, *Inverse Problems* 25(2009), 115002 (16).

6. C.F. Coleman and J.R. McLaughlin, Solution of the inverse spectral problem for an impedance with integrable derivative, I, *Comm. Pure Appl. Math.* 46 (1993), 145–184; II, *Comm. Pure Appl. Math.* 46 (1993), 185–212.
7. H. Dym, and H.P. McKean, *Gaussian Processes, Function Theory, and the Inverse Spectral Problem*, Probability and Mathematical Statistics, Vol. 31, Academic Press, New York-London, 1976.
8. R. Gorenflo, A. A. Kilbas, F. Mainardi, and S. V. Rogosin, *Mittag-Leffler Functions, Related Topics and Applications*, Springer, 2014.
9. O. Hald and J. McLaughlin, Recovery of BV coefficients from nodes, *Inverse Problems* 14 (1998), no. 2, 245–273.
10. M. Kirane and S. Malik, Determination of an unknown source term and the temperature distribution for the linear heat equation involving fractional derivative in time, *Applied Mathematics and Computation* 218 (2011), no. 1, 163–170.
11. Z. Li et al, Uniqueness in inverse boundary value problems for fractional diffusion equations, *Inverse Problems* 32 (2016), 015004.
12. I. Podlubny, *Fractional Differential Equations*, Academic Press, 1999.
13. Vu Kim Tuan, Inverse problem for fractional diffusion equation, *Fract. Calc. Appl. Anal.* 14 (2011), no. 1, 31–55.
14. A. Zettl, *Sturm-Liouville Theory*, Mathematical Surveys and Monographs, Vol. 121, 2005.

Three-Dimensional Biomagnetic Flow and Heat Transfer over a Stretching Surface with Variable Fluid Properties



M. G. Murtaza, E. E. Tzirtzilakis, and M. Ferdows

1 Introduction

The flow of biomagnetic fluid dynamics (BFD), a mathematical model, was first developed by Haik et al. [1]. This model conforms with the principles of ferrohydrodynamics (FHD) by Rosensweig [2]. This model considered biomagnetic fluid to be Newtonian, electrically nonconducting magnetic fluid. An extended mathematical model was developed by Tzirtzilakis [3]. According to this model, the biofluid flow under the influence of an applied magnetic field is consistent with the principles of FHD and magnetohydrodynamics (MHD). Tzirtzilakis and Kafoussias [4] analyzed the mathematical model of the flow of a biomagnetic fluid over a linearly stretching sheet under the action of a magnetic field, which is generated by a magnetic dipole. Further, Misra and Shit [5] investigated the biomagnetic viscoelastic fluid flow over a stretching sheet and indicated that the presence of an external magnetic field appreciably influences the flow of biomagnetic fluid.

The magnetization property M is the behavior of a biological fluid when it is exposed to a magnetic field. Andersson and Valnes [6] used a magnetization equation that is linear and temperature dependent, whereas Tzirtzilakis and

M. G. Murtaza (✉) · M. Ferdows

Research Group of Fluid Flow Modeling and Simulation, Department of Applied Mathematics,
University of Dhaka, Dhaka 1000, Bangladesh
e-mail: murtaza@cou.ac.bd; ferdows@du.ac.bd

E. E. Tzirtzilakis

Fluid Dynamics and Turbomachinery Laboratory, Department of Mechanical Engineering,
Technological Educational Institute of Western Greece, 1 M. Aleksandrou Str, Koukouli, 26334
Patras, Greece
e-mail: etzirtzilakis@teimes.gr

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_25

403

Kafaussias [4] used one that is nonlinear and temperature dependent. Haik et al. [7] studied the viscosity of human blood in a high static magnetic field. They used the magnetization equation that is not temperature dependent. In all the aforementioned studies, uniform fluid viscosity and thermal conductivity are considered. However, it is evident that the physical properties of fluid may change with temperature, especially fluid viscosity and fluid thermal conductivity. Vajravelu et al. [8] investigated the effects of variable fluid properties on the thin film flow of Ostwald-de Waele fluid over a stretching surface. Prasad et al. [9] also analyzed the effects of variable fluid properties on MHD flow and heat transfer over a nonlinear stretching sheet. Salawu and Dada [10] studied the radiative heat transfer of variable viscosity and thermal conductivity effects on an inclined magnetic field with dissipation in a non-Darcy medium. Makinde et al. [11] investigated the MHD variable viscosity reacting flow over a convectively heated plane in a porous medium with thermophoresis and radiative heat transfer. All the afore-mentioned authors assume that fluid viscosity and thermal conductivity vary as a linear function of temperature. Kafoussias et al. [12] investigated the free-forced convective boundary-layer flow of a biomagnetic fluid under the action of a localized magnetic field. They concluded that when the viscosity parameter increases, the skin friction coefficient also increases, whereas the Nusselt number decreases.

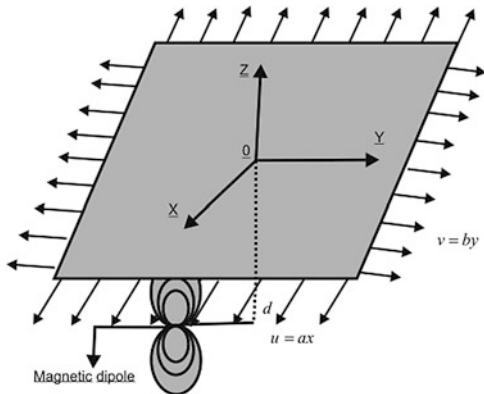
The aim of this study is to examine the temperature-dependent viscosity and thermal conductivity of biomagnetic fluid flow over a three-dimensional stretching sheet with variable surface temperature. Here, we conclude that the effects of variable viscosity and thermal conductivity and the flow characteristic are significantly changed compared with constant physical properties. This study will help the development of medical treatment by controlling blood velocity and blood temperature.

2 Formulation of the Problem

Let us consider a steady three-dimensional boundary-layer flow and heat transfer of a viscous incompressible biomagnetic fluid over a stretching surface. Assume that the flat surface stretches in two lateral directions x and y with the velocities ax and by , respectively. The stretching sheet is placed in the plane $z = 0$, whereas the fluid occupies the upper half of plane $z \geq 0$. Here, we consider that the plates are kept at a constant temperature T_w , whereas the fluid is at temperature T_c , such that $T_w < T_c$. Let the surface be maintained at a power law temperature. The viscous and electrically nonconducting magnetic fluid is subject to the action of a magnetic field H , which is generated by a magnetic dipole located at a distance d below the sheet and parallel to the x axis. The geometry and magnetic dipole of the problem is shown in Figure 1.

The boundary-layer equations of the fluid and energy equation in the presence of variable fluid properties can be written as

Fig. 1 Physical configuration and coordinate system



Continuity equation:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \tag{1}$$

Momentum equation:

$$u \frac{\partial u}{\partial x} + w \frac{\partial u}{\partial z} = \frac{1}{\rho_\infty} \frac{\partial}{\partial z} \left(\mu \frac{\partial u}{\partial z} \right) \tag{2}$$

$$v \frac{\partial v}{\partial y} + w \frac{\partial v}{\partial z} = \frac{1}{\rho_\infty} \frac{\partial}{\partial z} \left(\mu \frac{\partial v}{\partial z} \right) + \frac{1}{\rho_\infty} \mu_0 M \frac{\partial H}{\partial y} \tag{3}$$

$$w \frac{\partial w}{\partial z} = - \frac{1}{\rho_\infty} \frac{\partial p}{\partial z} + \frac{1}{\rho_\infty} \frac{\partial}{\partial z} \left(\mu \frac{\partial w}{\partial z} \right) + \frac{1}{\rho_\infty} \mu_0 M \frac{\partial H}{\partial z} \tag{4}$$

Energy equation:

$$\rho_\infty c_p \left(u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} + w \frac{\partial T}{\partial z} \right) + \mu_0 T \frac{\partial M}{\partial T} \left(v \frac{\partial H}{\partial y} + w \frac{\partial H}{\partial z} \right) = \frac{\partial}{\partial z} \left(k \frac{\partial T}{\partial z} \right) \tag{5}$$

With boundary conditions are

$$\left. \begin{aligned} u = u_w = ax, v = v_w = by, w = 0, T = T_w = T_c + Ax^m y^n \text{ at } z = 0 \\ u \rightarrow 0, v \rightarrow 0, T \rightarrow T_c, p + \frac{1}{2} \rho q^2 \text{ as } z \rightarrow \infty \end{aligned} \right\} \tag{6}$$

Here $u, v,$ and w are the velocity components along the x, y and z axes respectively. μ is the fluid viscosity, ρ_∞ is the fluid density far away from the sheet, μ_0 is the magnetic permeability, H is the magnetic field. The terms $\mu_0 M \frac{\partial H}{\partial y}$ and $\mu_0 M \frac{\partial H}{\partial z}$ in (3) and (4) respectively represent the components of the magnetic force per unit volume of the fluid and depend on the existence of the magnetic gradient. These two terms are well known from FHD, which are the so-called Kelvin forces

and the term $\mu_0 T \frac{\partial M}{\partial T} \left(v \frac{\partial H}{\partial y} + w \frac{\partial H}{\partial z} \right)$ of the thermal energy equation (5) represents the thermal power per unit volume due to magnetization that takes place as an adiabatic process. The power indices m and n indicate the variable surface temperature in the (x, y) plane.

Assuming that the viscosity and thermal conductivity of fluid is temperature dependent and is of the form described by Salawu and Dada [10]

$$\frac{1}{\mu} = \frac{1}{\mu_\infty} [1 + \gamma(T - T_\infty)] \text{ or } \frac{1}{\mu} = s(T - T_r) \tag{7}$$

where $s = \frac{\gamma}{\mu}$ and $T_r = T_\infty - \frac{1}{\gamma}$. Here s and T_r are the constants and their values depend on the reference state and γ is a constant connected with the thermal property of the fluid. Generally, for the liquids $s > 0$ and for gases $s < 0$.

On the other hand, for most liquids, the thermal conductivity k is assumed to vary as a linear function of temperature in the form described by Salawu and Dada [10]

$$k = k_\infty(1 + a\theta) \tag{8}$$

where $a = \frac{k_w - k_\infty}{k_\infty}$ is the thermal conductivity parameter.

The biomagnetic fluid flow is affected by the magnetic field generated by the presence of a magnetic dipole and it is assumed that the magnetic dipole is located at distance d below the sheet. The magnetic dipole gives rise to a magnetic field that is sufficiently strong to saturate the fluid and its scalar potential for the magnetic dipole whose components H_y, H_z of the magnetic field $H = (H_y, H_z)$, due to the magnetic dipole, are provided by Tzirtzilakis and Kafoussias [13]

$$H(x, y, z) = \frac{I}{2\pi} \frac{1}{\sqrt{y^2 + (z + d)^2}} = \frac{I}{2\pi} \left(\frac{1}{(z + d)} - \frac{1}{2} \frac{y^2}{(z + d)^3} \right)$$

A linear equation involving the magnetic intensity H and temperature T is provided by Tzirtzilakis and Kafaussias [13]

$$M = KH(T_c - T), \text{ where } K \text{ is a constant.} \tag{9}$$

3 Transformation of Equations

We are now introducing the nondimensional coordinates according to Tzirtzilakis and Kafoussias [13].

$$\xi(x) = \sqrt{\frac{a}{\nu}}x, \quad \zeta(y) = \sqrt{\frac{a}{\nu}}y, \quad \eta(z) = \sqrt{\frac{a}{\nu}}z$$

The dimensionless velocity, pressure, and temperature of the magnetic fluid are provided by the following expressions:

$$u = \sqrt{av} f'(\eta), v = \sqrt{av} f'(\eta), w = -\sqrt{av}(f(\eta) + g(\eta)) \tag{10}$$

$$P(\xi, \zeta, \eta) = a\mu_\infty P(\eta), \quad \theta(\xi, \zeta, \eta) = \frac{T_c - T}{T_c - T_w} = \theta(\eta) \tag{11}$$

The magnitude H of the magnetic field strength is given by the expression

$$H(\zeta, \eta) = \frac{I}{2\pi} \sqrt{\frac{a}{v}} \left[\frac{1}{\eta + \alpha} - \frac{1}{2} \frac{\zeta^2}{(\eta + \alpha)^3} \right] \tag{12}$$

where α is the dimensionless distance of the electric wire from the ξ axis and $\alpha = d\sqrt{\frac{a}{v}}$

By substituting Equation (9) and all the above expressions (10) to (12) into the momentum equation (2) to (4) and energy equation (5), and equating the coefficients of equal power of ξ, ζ we obtain the following system of ordinary differential equations.

$$f''' + \frac{\theta - \theta_r}{\theta_r} [f'^2 - (f + g)f''] - \frac{\theta'}{\theta - \theta_r} f'' = 0 \tag{13}$$

$$g''' + \frac{\theta - \theta_r}{\theta_r} [g'^2 - (f + g)g''] - \frac{\theta'}{\theta - \theta_r} g'' + \frac{\theta - \theta_r}{\theta_r} \frac{\beta\theta}{(\eta + \alpha)^4} = 0 \tag{14}$$

$$P' - \frac{\theta_r}{\theta - \theta_r} (f'' + g'') + \frac{\theta_r \theta'}{(\theta - \theta_r)^2} (f' + g') + \frac{\beta\theta}{(\eta + \alpha)^3} = 0 \tag{15}$$

$$(1 + a\theta)\theta'' + a\theta'^2 - Pr[mf'\theta + ng'\theta - (f + g)\theta'] - \frac{\beta\lambda(f + g)(\theta - \epsilon)}{(\eta + \alpha)^3} = 0 \tag{16}$$

The boundary conditions are

$$f' = 1, g' = \frac{b}{a} = \delta, \theta = 1, f = g = 0 \quad \text{at} \quad \eta = 0 \tag{17}$$

$$f' \rightarrow 0, g' \rightarrow 0, P \rightarrow -P_\infty, \theta \rightarrow 0 \quad \text{as} \quad \eta \rightarrow \infty \tag{18}$$

The dimensionless parameters appearing in these equations are like the Prandtl number, $Pr = \frac{\mu_\infty c_p}{k_\infty}$, viscous dissipation parameter, $\lambda = \frac{a\mu_\infty^2}{\rho_\infty k_\infty (T_c - T_w)}$, dimensionless Curie temperature, $\epsilon = \frac{T_c}{(T_c - T_w)}$, ferromagnetic interaction parameter, $\beta = \frac{I^2}{4\pi^2} \frac{K\mu_0(T_c - T_w)\rho_\infty}{\mu_\infty^2}$, dimensionless distance, $\alpha = d\sqrt{\frac{a}{v}}$, viscosity parameter,

$\theta_r = \frac{T_r - T_c}{T_w - T_c} = -\frac{1}{\gamma(T_w - T_c)}$, where θ_r is negative for liquids and θ_r is positive for gases.

The local skin friction coefficient and local Nusselt number are important physical parameters of this flow and heat transfer, which is defined respectively

$$C_f = -\frac{\tau_w}{\frac{1}{2}\rho_\infty u_w^2} \quad \text{and} \quad Nu = -\frac{xq_w}{k_\infty(T_w - T_c)}$$

where $\tau_w = -\mu\left(\frac{\partial u}{\partial z}\right)_{z=0}$ and $q_w = -k\left(\frac{\partial T}{\partial z}\right)_{z=0}$.

and the corresponding dimensionless quantities can be written as

$$C_{fx} = C_f(Re_x)^{1/2} = -\frac{2\theta_r}{\theta_r - 1} f''(0) \quad \text{and} \quad Nu_x = Nu(Re_x)^{-1/2} = -\theta'(0)$$

4 Numerical Method

The set of Equations (13), (14), and (16) are highly nonlinear and coupled and therefore the system cannot be solved analytically. Thus, the Equations (13), (14), and (16), with boundary conditions (17) and (18), are solved numerically using the essential features of this technique are based on: (i) the common finite difference method with central differencing; (ii) a tridiagonal matrix manipulation; and (iii) an iterative procedure. This numerical method is described in detail in Kaffoussias and Williams [14] and is used in Murtaza et al. [15].

5 Results

To assess the validity and accuracy of the numerical results, we computed the numerical values for the wall temperature gradient and compared with those of Liu and Andersson [16] by setting $\beta = 0$, $f'(0) = 1$, $g'(0) = 0.5$, $Pr = 1$. The comparisons are found to be in good agreement (Table 1).

Table 1 Comparisons with the published literature with regard to wall heat transfer rate coefficients

Stretching ratio	$m = 0, n = 0$		$m = 2, n = 0$		$m = 0, n = 2$	
	Present	Liu and Anderson [16]	Present	Liu and Anderson [16]	Present	Liu and Anderson [16]
$\delta = 0.25$	-0.66721	-0.665933)	-1.36331	-1.364890	-0.88301	-0.883125
$\delta = 0.5$	-0.73546	-0.735334	-1.39377	-1.395356	-1.10544	-1.106491
$\delta = 0.75$	-0.79599	-0.796472	-1.42341	-1.425038	-1.29056	-1.292003

As the fluid is blood, we considered a human body temperature $T_w = 37^\circ\text{C}$, whereas the body Curie temperature is $T_c = 41^\circ\text{C}$ according to Loukopoulos and Tzirtzilakis [18]. For these values of temperature, the dimensionless temperature number is $\epsilon = 78.5$. Also, we assume $\rho = 1050 \text{ kgm}^{-3}$, $\mu = 3.2 \times 10^{-3} \text{ kgm}^{-3}\text{s}^{-1}$ according to Tzirtzilakis [17]. Generally, the specific heat under a constant pressure c_p and thermal conductivity k of any fluid are temperature dependent. For the temperature range, consider this problem, $c_p = 14.65 \text{ Jkg}^{-1}\text{K}^{-1}$ and $k = 2.2 \times 10^{-3} \text{ Jm}^{-1}\text{s}^{-1}\text{K}^{-1}$ respectively according to Tzirtzilakis and Xenos [19] and hence $Pr = 21$. We consider the ferromagnetic interaction parameter $\beta = 0$ to 10 as in Tzirtzilakis and Kafoussias [4]. Note that $\beta = 0$ corresponds to hydrodynamic flow. The viscous dissipation parameter is $\lambda = 6.4 \times 10^{-14}$.

Figures 2, 3, and 4 display the influence of the ferromagnetic parameter, the viscosity parameter, and the thermal conductivity parameter on velocity and temperature distributions. It is evident from the figures that with an increase in the ferromagnetic parameter, the velocity profiles $f'(\eta)$ are greater than the corresponding hydrodynamics case. However, the opposite is true for the velocity component $g'(\eta)$ (Figure 3). This fact is due to the influence of Kelvin forces on the flow field in the y direction. From Figure 4, it is observed that an increase in the ferromagnetic field parameter increases the temperature profiles. The reason behind this is that an increase in the magnetic field reduces the boundary-layer thickness and enhances the thermal conductivity of the fluid. These figures also

Fig. 2 Velocity profile along x -axis for various values of β, a, θ_r

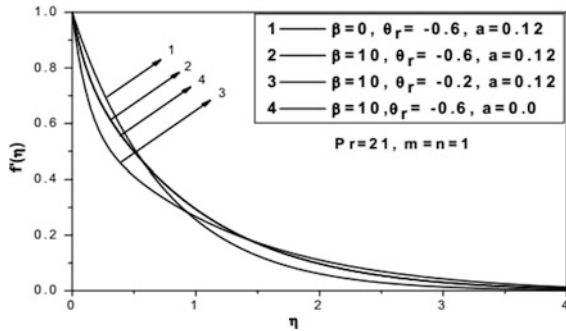


Fig. 3 Velocity profile along the y -axis for various values of β, a, θ_r

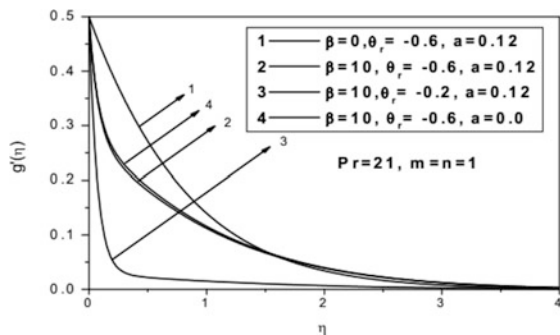


Fig. 4 Temperature profile for various values of β, a, θ_r

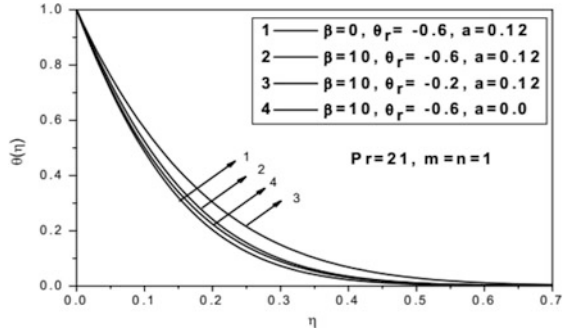


Fig. 5 Velocity profile along x-axis for various values of δ, m

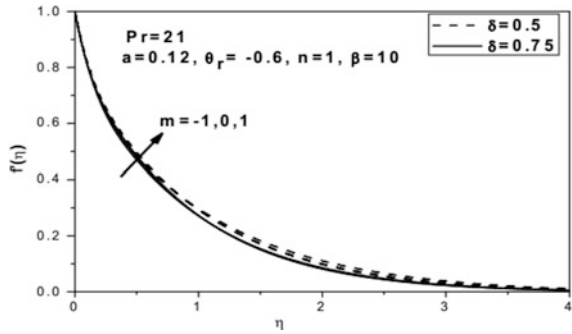
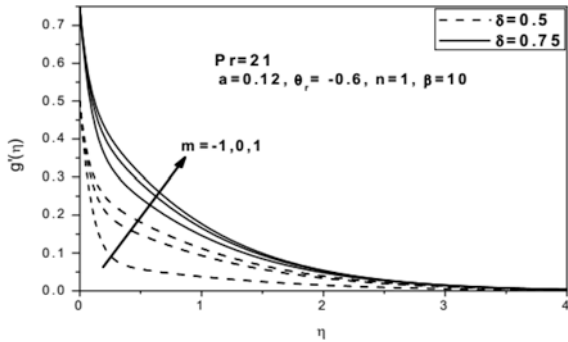


Fig. 6 Velocity profile along y-axis for various values of δ, m



indicate that with increases in the values of θ_r , the velocity decreases and enhances the temperature profile. This is because the increase θ_r results in an increase in the thermal boundary-layer thickness, which results in a decrease in the velocity and an increase in the temperature.

Figures 5, 6, 7, 8, and 9 exhibit the effect of the wall temperature parameter on velocity and temperature distribution. From these figures, we observed that the variation of sheet temperature has a significant effect on the velocity and temperature profile. From the figures, we conclude that the velocity profile increases with an increase in the wall temperature parameter, but the opposite behavior is shown for temperature profile. This is because when $m, n > 0$ heat flows from

Fig. 7 Temperature profile for various values of δ, m

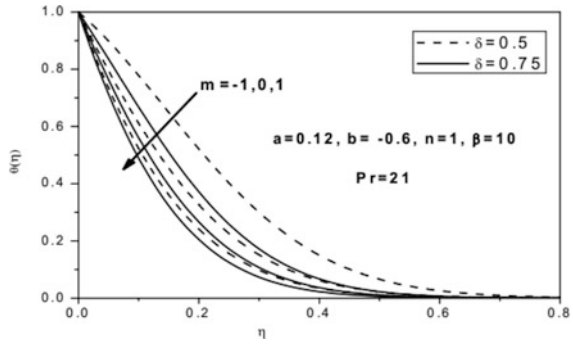


Fig. 8 Velocity profile along y-axis for various values of δ, n

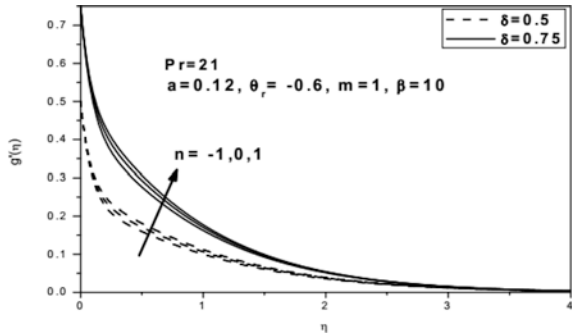
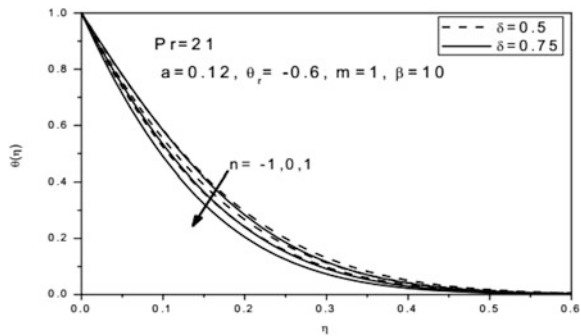


Fig. 9 Temperature profile for various values of δ, n ,



the stretching sheet into the fluid and when $m, n < 0$, the temperature gradient is positive and heat flows into the stretching sheet from the fluid. When m, n are both increased, then the temperature profile is decreased and the velocity is increased, i.e., the thermal boundary layer becomes thinner and the momentum boundary layer is thicker.

Figures 10, 11, 12, and 13 depict the skin friction coefficient and rate of wall heat transfer with regard to the viscosity parameter and thermal conductivity parameter for various values of ferromagnetic number β . It is observed from these figures that by increasing the viscosity parameter, the velocity gradient at the wall is increased and the reverse trend was found for the wall temperature gradient. In

Fig. 10 Skin friction coefficient for various values of β with regard to θ_r

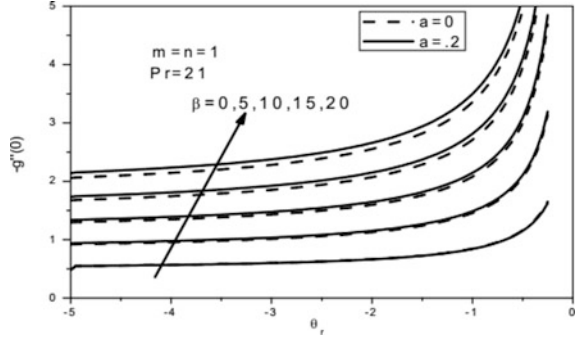


Fig. 11 Rate of heat transfer for various values of β with regard to θ_r

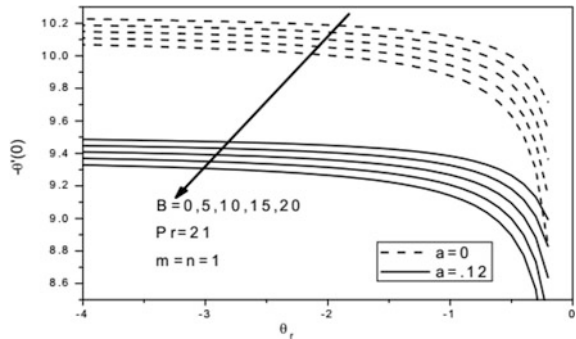
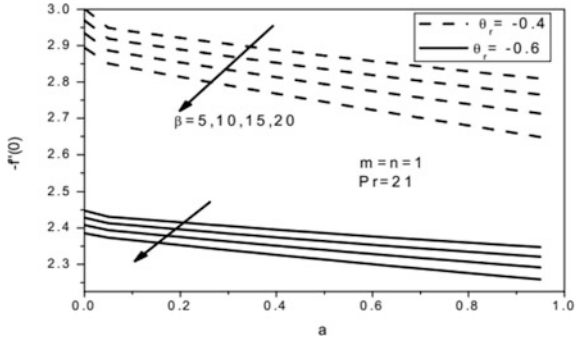
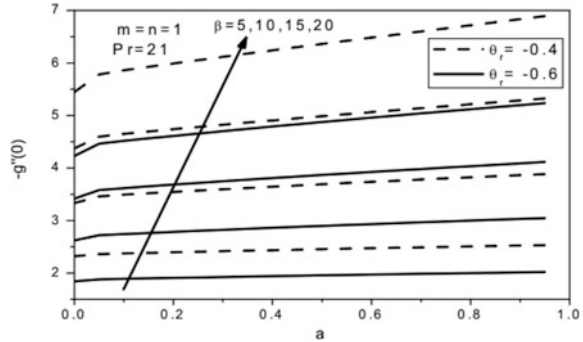


Fig. 12 Skin friction for various values of β and θ_r with regard to a



Figures 12, and 13, we see that the skin friction $f''(0)$ decreases with increases in the thermal conductivity parameter, whereas $g''(0)$ is increasing. The rate of wall heat transfer decreases with increases in the thermal conductivity parameter. We also observe that the viscosity parameter is less affected by temperature distribution than skin friction. It is noted that the effect of thermal conductivity is greater in the temperature gradient than other physical significance. On the other hand, the viscosity parameter is more affected in skin friction than the temperature gradient.

Fig. 13 Skin friction for various values of β and θ_r with regard to a



6 Conclusion

In this chapter, the effect of variable fluid properties on BFD in the presence of an applied magnetic field is analyzed. The results are presented graphically to investigate the influence of pertinent parameters on the velocity and temperature fields. Some of the important results are summarized here.

- 1) The effect of the variable thermal conductivity parameter is to enhance the temperature in the flow region and is reversed in the case of the wall temperature parameter. This parameter effect is negligible for velocity and skin friction.
- 2) The effect of the increasing value of the viscosity parameter θ_r is to enhance the temperature but decrease the velocity. This parameter has a greater effect on the velocity profile and skin friction, but the effect on the wall temperature gradient is negligible.
- 3) For the effect of the ferromagnetic parameter, as the ferromagnetic number increases, the velocity profile $f'(\eta)$ increases, but the velocity profile $g'(\eta)$ decreases with increased ferromagnetic number. This occurs because of the Kelvin forces that act on the y axis.
- 4) The effect of thermal conductivity is greater in the temperature gradient than other physical significance. On the other hand, the viscosity parameter is more affected in skin friction than the temperature gradient.

Acknowledgements The author would like to thank the Ministry of Science and Technology, Bangladesh, for providing financial support under the NST fellowship.

References

1. Haik, Y., Pai, V. and Chen, C. J.: Biomagnetic fluid dynamics, in fluid dynamics at interfaces, edited by W. Shyy and R. Narayanan (Cambridge University Press, Cambridge), 439–452 (1999)
2. Rosensweig, R. E.: Magnetic fluids. Annual Review of Fluid Mechanics, **19**, 437–461(1987).

3. Tzirtzilakis, E.E.: A mathematical model for blood flow in magnetic field, *Physics of fluid*, **17**, 077103 (2005)
4. Tzirtzilakis, E. E. and Kafoussias, N. G.: Bio-magnetic fluid flow over a stretching sheet with nonlinear temperature dependent magnetization, *Z. angew. Math. Phys.*, **54**, 551–565, (2003)
5. Misra, J. C., Shit, G. C.: Biomagnetic viscoelastic fluid flow over a stretching sheet. *Applied mathematics and computation*, **210**, 350–361 (2009)
6. Andersson, H. I., Valnes, O. A.: Flow of a heated ferrofluid over a stretching sheet in the presence of a magnetic dipole, *Acta Mech.* **128**, 39–47 (1998)
7. Haik, Y., Pai, V., Chen, C. J.: Apparent viscosity of human blood in a high static magnetic field, *J. Magn. Magn. Mater.* **225**, 180 (2001)
8. Vajravelu, K., Prasad, K.V. and Raju, B.T.: Effects of variable fluid properties on the thin film flow of Ostwald-de Waele fluid over a stretching surface, *Journal of Hydrodynamics Series B*, **25**, 10–19 (2013)
9. Prasad, K. V., Vajravelu, K. and Datti, P. S.: The effects of variable fluid properties on the hydromagnetic flow and heat transfer over a non-linearly stretching sheet, *International Journal of Thermal Sciences*, **49**, 603–610 (2010)
10. Salawu, S. O., Dada, M. S.: The radiative heat transfer of variable viscosity and thermal conductivity effects on inclined magnetic field with dissipation in a non-Darcy medium. *Journal of Nigerian Mathematical Society*, **35**, 93–106 (2016)
11. Makinde O. D., Khan, W. A., Culham J. R.: MHD variable viscosity reacting flow over a convectively heated plane in a porous medium with thermophoresis and radiative heat transfer. *Int. j. of Heat Mass Transfer*, **93**, 595–604 (2016)
12. Kafoussias, N. G., Tzirtzilakis, E. E., Raptis, A.: Free-force convective boundary layer flow of a biomagnetic fluid under the action of a localized magnetic field. *Canadian J. of Physics*, **86**, 447–457 (2008)
13. Tzirtzilakis, E. E., Kafoussias, N. G.: Three dimensional magnetic fluid boundary layer flow over a linearly stretching sheet, *Journal of heat transfer*, **132**, 011702-1-8 (2010)
14. Kafoussias, N. G., Williams, E. W.: An Improved Approximation Technique to Obtain Numerical Solution of a Class of Two-Point Boundary Value Similarity Problems in Fluid Mechanics, *Int. J. For Numerical Methods in Fluids*, **17**, 145–162 (1993)
15. Murtaza, M. G., Tzirtzilakis, E. E., Ferdows, M.: Effect of electrical conductivity and magnetization on the biomagnetic fluid flow over a stretching sheet. *Z. Angew. Math, Phys (ZAMP)*, **68**, 93 (2017)
16. Liu, I. C., Andersson, H. I.: Heat transfer over a bidirectional stretching sheet with variable thermal conditions, *International journal of heat and mass transfer*, **51**, 4018–4024, (2008)
17. Tzirtzilakis, E. E.: A simple numerical methodology for BFD problems using stream function vorticity formulation. *Commun numer methods Eng*, **24**, 683–700 (2008)
18. Loukopoulos, V. C., Tzirtzilakis, E. E.: Biomagnetic channel flow in spatially varying magnetic field, *International journal of engineering science*, **42**, 571–590 (2004)
19. Tzirtzilakis, E. E., Xenos, M. A. : Biomagnetic fluid flow in a driven cavity, *Meccanica*, **48**, 187–200 (2013)

Effects of Slip on the Peristaltic Motion of a Jeffrey Fluid in Porous Medium with Wall Effects



Gurunath Sankad and Pratima S. Nagathan

1 Introduction

In physiology, the investigation of peristaltic movement has enhanced the interest of researchers for its wide applications as seen in the transportation of urine produced by the kidneys along the ureters to the bladder, the movements of the chyme in gastrointestinal tract, transport of harmful acidic sanitary fluids, and also fluids in the nuclear industry. Latham [1] did the pioneering work on peristalsis. Weinberg et al. [2] observed the peristaltic pumping experimentally. The wall effect on the Poiseuille flow influenced by peristalsis has been undertaken by Mitra and Prasad [3]. Misra and Pandey [4] analyzed the transportation of blood through small vessels by mathematically modeling the motion under peristalsis. Study of fluids behaving non-Newtonianly has recently gained importance as the usual viscous fluids fail to explain the distinctiveness of numerous physiological fluids. Literature survey reveals many important analytic studies going on non-Newtonian fluids with peristalsis flowing inside a porous space influenced by compliant walls: Raju and Devanathan [5], Srivastava and Srivastava [6], Srinivas and Kothandapani [7], Sankad and Radhakrishnamacharya [8], Alsaedi et al. [9], and Sankad and Asha [10].

In the classical problems of fluid flow past bodies, it is usual to apply no-slip on the surface of the bodies considered. However, with the advent of miniature devices and investigations with rarified gases, it is watched that fluid slips on the surface of the bodies in the following situations: when fluid flow occurs over the surface of a

G. Sankad · P. S. Nagathan (✉)

Department of Mathematics, B.L.D.E.A.S.V.P.Dr.P.G. Halakatti College of Engineering and Technology, (Affiliated to Visvesvaraya Technological University, Belagavi, India), Vijayapur 586103, Karnataka, India

e-mail: math.gurunath@bldeacet.ac.in; math.pratima@bldeacet.ac.in

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41, https://doi.org/10.1007/978-3-030-02487-1_26

415

porous medium (Beavers and Joseph [11]); at the point when the fluid flow happens in rarified gases at low density and low pressure (Sreekanth. A. K. [12]). Shehawey et al. [13], Sobh [14], Ellahi [15], Sankad and Pratima [16], and Hina [17] studied peristaltic transport influenced by slip, in uniform as well as nonuniform channel.

Among numerous non-Newtonian models that define the physiological fluids, Jeffrey model is most probably considerable since Newtonian fluid model can be worked out from this by taking $\lambda_1 = 0$. Most of the researchers hypothesize that blood shows Newtonian and non-Newtonian behaviors while it circulates in human body. Also, this model is a simple linear model that makes use of time derivative as an alternative for convective derivative.

Hayat and Ahamed [18] have considered the consequences of magnetic effect and endoscope on the peristaltic transport involving Jeffrey fluid. Mahmouda et al. [19] looked into the motion of a Jeffrey fluid within a permeable space under peristalsis having magnetic effect. Sudhakar Reddy et al. [20] discussed the flow of a Jeffery fluid in a uniform tube travelling with the velocity of the peristaltic wave within a fixed frame of reference, having variable viscosity. Subba Reddy et al. [21] have examined and analyzed the magnetic field effects on the stream of a Jeffrey fluid flowing inside a permeable peristaltic channel with asymmetry and slip effects. Arun Kumar et al. [22] considered the impacts of elastic wall and heat transfer of a non-Newtonian Jeffrey fluid in a peristaltic conduit. Dheia and Ahmed [23] studied the impacts of wall and heat on the bolus, believed as Jeffrey fluid, moving in the esophagus. Also, surface of the esophagus is well thought-out as peristaltic wave with porosity. Their results depict that velocity enhances with rise in Jeffrey parameter, Darcy number, thermal conductivity, and Grashoff number. Bhatti and Ali [24] examined the impacts of slip condition and MHD for the peristaltic flow of blood considering as a Jeffrey fluid model along the permeable membrane. The present study aims in discussing the flow of a Jeffrey fluid considering a uniform conduit influenced by wall effects moving under peristalsis.

2 Mathematical Formulation

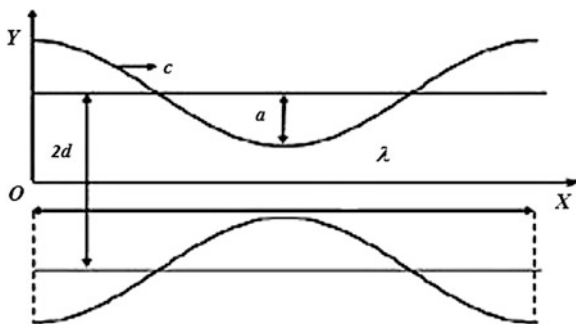
Consider an incompressible fluid, namely the Jeffrey fluid, moving within a uniform conduit in between the flexible peristaltic walls as described in Figure 1. Here, the channel thickness is considered $2d$, time t , wave amplitude a , and wave length λ . x is the direction of wave propagation and y is perpendicular to x axis. The motion is discussed considering only half the width of channel.

The peristaltic wave is propagating with speed c along the conduit wall. The infinite wavelenght of wall equation is

$$y = \pm\eta = \pm \left[d + a \sin \frac{2\pi}{\lambda} (x - ct) \right], \quad (1)$$

The elastic wall is governed by the equation:

Fig. 1 Physical model



$$L(\eta) = p - p_0, \tag{2}$$

The operator L used to represent stretched membrane movement accompanied by viscosity damping force is given by:

$$L = -T \frac{\partial^2}{\partial x^2} + m \frac{\partial^2}{\partial t^2} + C \frac{\partial}{\partial t}. \tag{3}$$

Here, T is elastic tension in the membrane, m is mass per unit area, and C is coefficient of viscous damping force.

The flow is governed by the equations:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{4}$$

$$\rho \left[\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} \right] = -\frac{\partial p}{\partial x} + \frac{\mu}{1 + \lambda_1} \nabla^2 u - \mu \frac{u}{K}, \tag{5}$$

$$\rho \left[\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} \right] = -\frac{\partial p}{\partial y} + \frac{\mu}{1 + \lambda_1} \nabla^2 v - \mu \frac{v}{K}. \tag{6}$$

The velocity in the direction of x and y are correspondingly u and v . Due to symmetrical plane, the normal velocity is zero. Experimentally, it is proved in several physiological situations that flow is accompanied with very small Reynolds number. Hence, infinite wavelength is assumed. The ratio of relaxation time to retardation time is λ_1 , porous medium permeability is K , fluid density is ρ , fluid viscosity coefficient is μ and $\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$, $\nabla^2 \nabla^2 = \nabla^4$, and pressure is P . Due to the tension in the muscles, pressure is exerted on the outer surface of the wall and is denoted by P_0 . Here, we assume $P_0 = 0$.

$$\frac{\partial u}{\partial y} = 0, \text{ at } y = 0, \text{ (theregularitycondition),} \tag{7}$$

$$u = -d \frac{\sqrt{D_a}}{\beta} \frac{\partial u}{\partial y} \quad \text{at} \quad y = \pm \eta(x, t), \text{ (theslipcondition)}. \tag{8}$$

where D_a is Darcy number and β is slip parameter.

With reference to [3], the peripheral conditions are

$$\frac{\partial}{\partial x} L(\eta) = -\rho \left[\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} \right] + \frac{\mu}{1 + \lambda_1} \nabla^2 u - \mu \frac{u}{K}, \quad \text{at} \quad y = \pm \eta(x, t), \tag{9}$$

where

$$\frac{\partial}{\partial x} L(\eta) = \frac{\partial p}{\partial x} = -T \frac{\partial^3 \eta}{\partial x^3} + m \frac{\partial^3 \eta}{\partial x \partial t^2} + C \frac{\partial^2 \eta}{\partial x \partial t}. \tag{10}$$

We introduce the stream function ψ given by:

$$u = \frac{\partial \psi}{\partial y}, \quad v = -\frac{\partial \psi}{\partial x} \tag{11}$$

Dimensionless variables are

$$x' = \frac{x}{\lambda}, \quad y' = \frac{y}{\lambda}, \quad \psi' = \frac{\psi}{cd}, \quad t' = \frac{ct}{\lambda}, \quad u' = \frac{u}{c}, \quad v' = \frac{\lambda v}{cd}, \quad P' = \frac{d^2 P}{\mu \lambda c}, \quad \eta' = \frac{\eta}{d}. \tag{12}$$

Introducing nondimensional variables in (4)–(10) and after deletion of primes, we get

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{13}$$

$$R_e \delta \left(\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} \right) = -\frac{\partial p}{\partial x} + \frac{1}{1 + \lambda_1} \left(\delta^2 \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) - \frac{u}{D_a}, \tag{14}$$

$$R_e \delta^3 \left(\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} \right) = -\frac{\partial p}{\partial y} + \frac{\delta^2}{1 + \lambda_1} \left(\delta^2 \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) - \delta^2 \frac{v}{D_a}, \tag{15}$$

$$\frac{\partial u}{\partial y} = 0, \quad \text{at} \quad y = 0. \tag{16}$$

$$u = -\frac{\sqrt{D_a}}{\beta} \frac{\partial u}{\partial y} \quad \text{at} \quad y = \pm \eta(x, t) = \pm (1 + \epsilon \sin 2\pi(x - t)), \tag{17}$$

and

$$\begin{aligned}
 & -R_e \delta \left(\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} \right) + \frac{1}{1 + \lambda_1} \left(\delta^2 \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right) - \frac{u}{D_a} \\
 & = E_1 \frac{\partial^3 \eta}{\partial x^3} + E_2 \frac{\partial^3 \eta}{\partial x \partial t^2} + E_3 \frac{\partial^2 \eta}{\partial x \partial t} \quad \text{at } y = \pm \eta(x, t). \quad (18)
 \end{aligned}$$

Here, $\epsilon = \frac{a}{d}$ the amplitude ratio, $\delta = \frac{d}{\lambda}$ the wall slope parameter, $R_e = \frac{\rho c d}{\mu}$ the Reynolds number, and $D_a = \frac{K}{d^2}$ Darcy number. The elastic parameters are defined as $E_1 = \frac{-T d^3}{c \mu \lambda^3}$, $E_2 = \frac{m c d^3}{\mu \lambda^3}$, and $E_3 = \frac{c d^3}{\mu \lambda^2}$. The parameter E_1 represents the rigidity, E_2 the stiffness, and E_3 the viscous damping force in the membrane.

3 Solution

Usually, the analytic solution of the governing equations is not possible in general; hence, we assume long wavelength approximation to solve Equations (13–18).

Equations (13)–(15) yield the compatibility equations as:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \quad (19)$$

$$0 = -\frac{\partial p}{\partial x} + \frac{1}{1 + \lambda_1} \frac{\partial^2 u}{\partial y^2} - \frac{u}{D_a}, \quad (20)$$

$$0 = -\frac{\partial p}{\partial y}. \quad (21)$$

The boundary conditions (16)–(18) become

$$\frac{\partial u}{\partial y} = 0, \quad \text{at } y = 0. \quad (22)$$

$$u = -\frac{\sqrt{D_a}}{\beta} \frac{\partial u}{\partial y} \quad \text{at } y = \pm \eta(x, t) = \pm (1 + \epsilon \sin 2\pi(x - t)), \quad (23)$$

$$\frac{1}{1 + \lambda_1} \frac{\partial^2 u}{\partial y^2} - \frac{u}{D_a} = E_1 \frac{\partial^3 \eta}{\partial x^3} + E_2 \frac{\partial^3 \eta}{\partial x \partial t^2} + E_3 \frac{\partial^2 \eta}{\partial x \partial t} \quad \text{at } y = \pm \eta(x, t). \quad (24)$$

Solving Equations (19) and (20) with boundary conditions (22)–(24), we get

$$u = \frac{E}{N^2} \left[-1 - \frac{\cosh(Ny)}{T_1} \right], \tag{25}$$

where

$$E = -\epsilon \left[(E_1 + E_2)(2\pi)^3 \cos(2\pi(x - t)) - E_3(2\pi)^2 \sin(2\pi(x - t)) \right],$$

$$T_1 = DN \sinh(N\eta) - \sinh(N\eta), \quad D = -\frac{\sqrt{D_a}}{\beta}, \quad N = \sqrt{\frac{1 + \lambda_1}{D_a}}. \tag{26}$$

The mean velocity \bar{u} is

$$\bar{u} = \int_0^1 u dt. \tag{27}$$

The stream function $u = \frac{\partial \psi}{\partial y}$ can be found by Equation (25) and using the condition $\psi = 0$ at $y = 0$ is

$$\psi = \frac{E}{N^2} \left[-y - \frac{\sinh(Ny)}{NT_1} \right]. \tag{28}$$

4 Numerical Results and Discussion

The effect of elastic wall properties is examined and the nonlinear governing equations are solved using small Reynolds number and large wave length approximation. The consequences of the parameters under consideration on the mean velocity profile $\bar{u}(y)$ are obtained. From Equation (25), graphs are plotted and depicted in Figures 2, 3, 4, 5 and 6 to observe the consequences of the physical parameters, say Jaffrey parameter λ_1 , slip parameter β 's, elastic parameters, and Darcy number D_a on mean velocity $\bar{u}(y)$. Graphs are plotted using the values: $\epsilon = 0.2$; $E_1 = 0.1$; $E_2 = 0.2$; $E_3 = 0.4$; $D_a = 0.5$; $\beta = 0.1$; $x = 0.5$; and $\lambda_1 = 1$.

Figures 2 and 3 show the variation in $\bar{u}(y)$ under the effect of viscous damping force (E_3) of the elastic wall in presence ($E_2 \neq 0$) and absence ($E_2 = 0$) of stiffness in the wall. It is concluded that the mean velocity reduces with enhancement in viscous damping force, the elastic parameter. The mean velocity $\bar{u}(y)$ reduces by gain in the slip parameter β and Darcy parameter D_a as depicted in Figures 4 and 5. Figure 6 presents the flow structure of mean velocity $\bar{u}(y)$ for different Jeffrey parameter (λ_1). This figure indicates that rise in λ_1 results gradual increase in the mean velocity $\bar{u}(y)$.

Families of curves representing the streamlines of the flow are instantaneously tangential to the velocity vector that explains the path of a fluid particle travelling at any instance. Trapping is an attractive phenomena in peristalsis. The creation of the bolus of the fluid circulating internally enclosed by streamline patterns for various values of Jeffrey parameter λ_1 , slip parameter β , Darcy number D_a , and elastic parameters are shown in Figures 7, 8, 9, 10, 11, 12, 13, 14, 15, and 16. Figures 7, 8, 9, and 10 reveal that the trapped bolus enhances in size as there is rise in viscous damping force in the presence (Figures 7 and 8) and absence (Figures 9 and 10) of stiffness, respectively. It is shown from Figures 11, 12, 13, 14, 15, and 16 that as the slip parameter (Figures 11 and 12) and Jeffrey parameter (Figures 15 and 16) are enhanced, the bolus size reduces. Size of the bolus enhances with rise in Darcy number as shown in Figures 13 and 14.

Fig. 2 Effect of mean velocity with y for various values of $E_2 \neq 0$ and $E_3 = 0.4$.

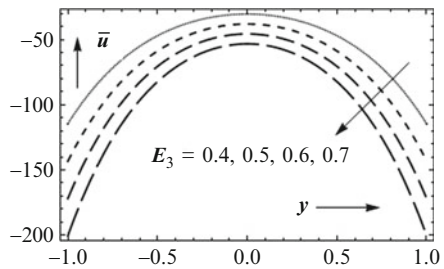


Fig. 3 Effect of mean velocity with y for various values of $E_2 = 0$ and $E_3 = 0.7$.

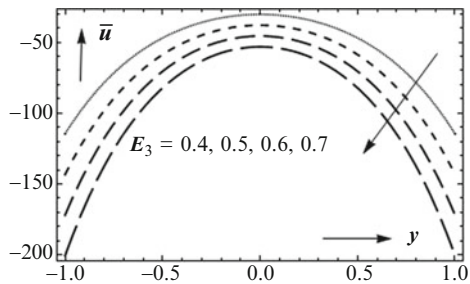


Fig. 4 Effect of mean velocity with y for various values of β .

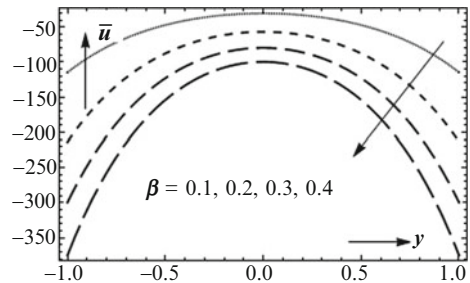


Fig. 5 Effect of mean velocity with y for various values of D_a .

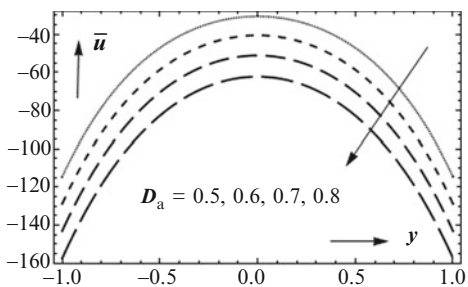


Fig. 6 Effect of mean velocity with y for various values of λ_1 .

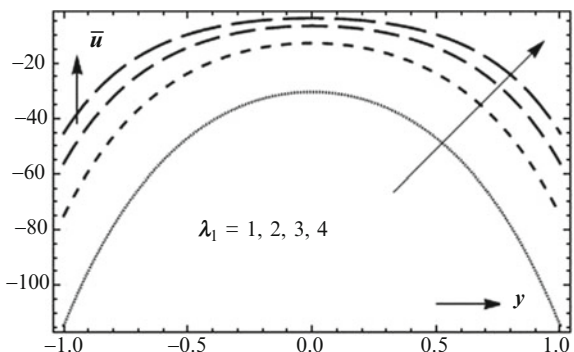


Fig. 7 Streamlines for $E_2 \neq 0$ and $E_3 = 0.4$.

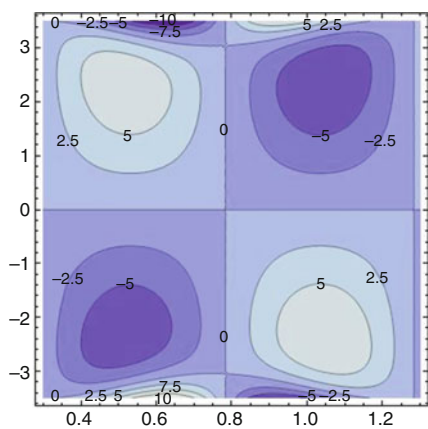


Fig. 8 Stream lines for $E_2 \neq 0$ and $E_3 = 0.7$.

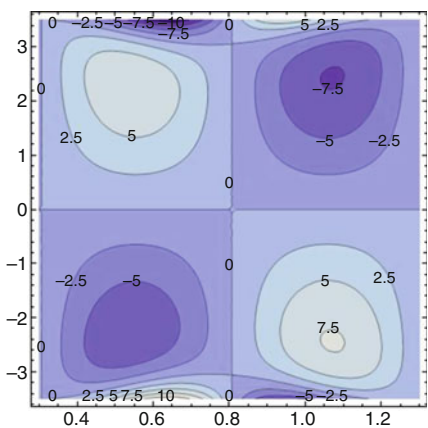


Fig. 9 Streamlines for $E_2 = 0$ and $E_3 = 0.4$.

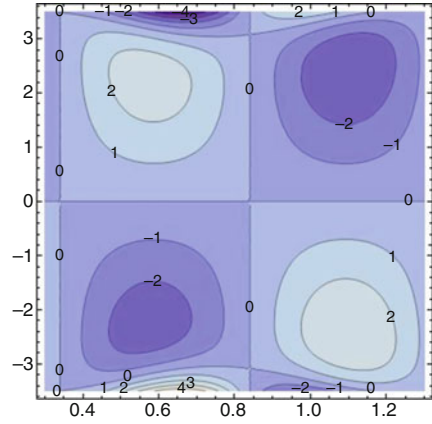


Fig. 10 Streamlines for $E_2 = 0$, and $E_3 = 0.7$.

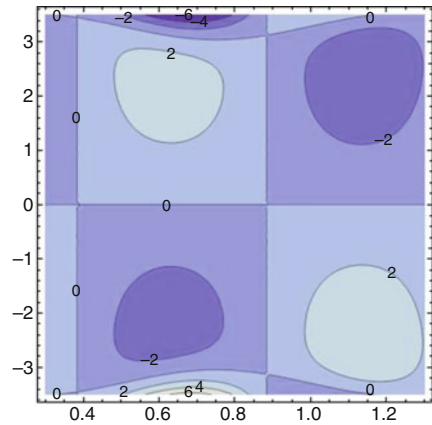


Fig. 11 Streamlines for $\beta = 0.1$.

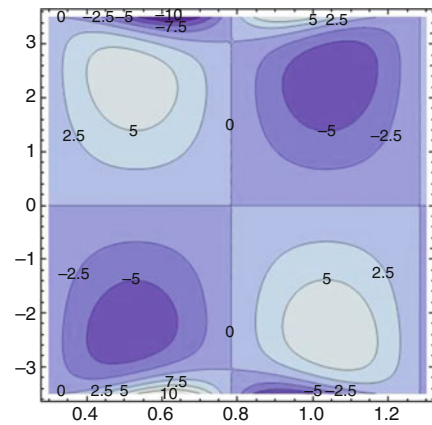


Fig. 12 Streamlines for $\beta = 0.2$.

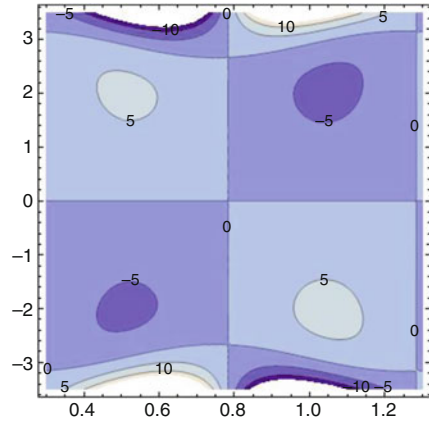


Fig. 13 Streamlines for $D_a = 0.5$.

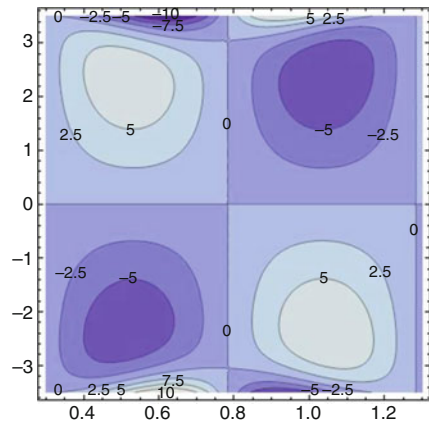


Fig. 14 Streamlines for $D_a = 0.6$.

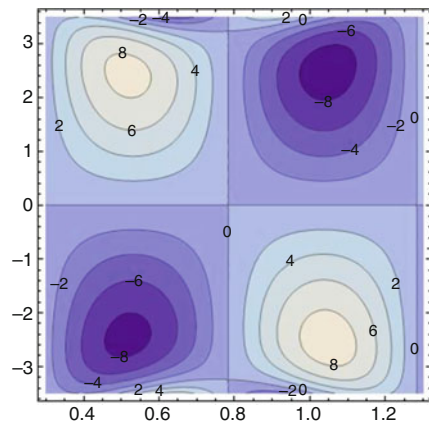


Fig. 15 Streamlines for $\lambda_1 = 1$.

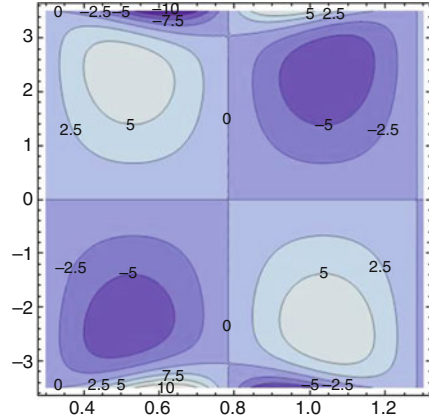
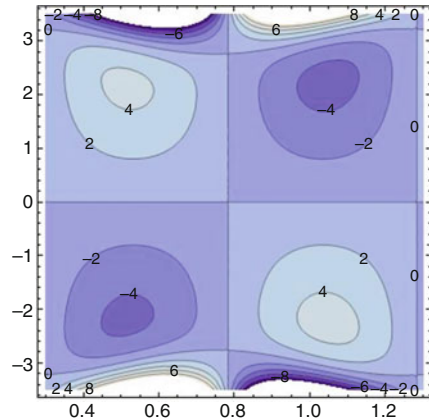


Fig. 16 Streamlines for $\lambda_1 = 2$.



5 Conclusion

Here, the results of various parameters of a Jeffrey fluid model in a symmetric conduit on the velocity profile are observed graphically and the trapping phenomenon is also shown by plotting streamlines for different applicable parameters. The analytical solutions are obtained for mean velocity and stream function. The major findings can be reviewed as:

- The mean velocity enhances with gain in Jeffrey parameter. This result is in agreement with that of Dheia and Ahmed [23].
- The mean velocity reduces as the viscous damping force E_3 , Darcy number Da , and slip parameter β increase.
- The size of trapped bolus increases with increase in E_3 and Da .
- Bolus decreases with increase in β and λ_1 .

References

1. Latham, T. W.: Fluid motions in a peristaltic pump. M.S. Thesis. MIT, Cambridge, MA. (1966).
2. Weinberg, S. L., Eckstein, E. C., Shapiro, A. H.: An experimental study of peristaltic pumping. *J. Fluid Mech.* **49**, 461–497(1971)
3. Mittra, T. K., Prasad, S. N.: On the influence of wall properties and Poiseuille flow in Peristalsis. *J. Biomech.* **6**, 681–693 (1973)
4. Misra, J. C., Pandey, S. K.: Peristaltic transport of blood in small vessels: Study of mathematical model. *Computers and Mathematics with Applications*.**43**,1183–1193 (2002)
5. Raju, K. K., Devanathan, R.: Peristaltic motion of a non-Newtonian fluid. *Rheol. Acta.* **11**, 170–178 (1972)
6. Srivastava, L. M., Srivastava, V. P.: Peristaltic transport of a non-Newtonian fluid: Applications to the vas deferens and small intestine. *Annals of Biomedical Engineering.* **13**, 137–153(1985)
7. Srinivas, S., Kothandapani, M.: The influence of heat and mass transfer on MHD peristaltic flow through a porous space with compliant walls. *Appl Math Comput.* **213**, 197–208(2009)
8. Sankad, G. C., Radhakrishnamacharya, G.: Effects of magnetic field on the peristaltic transport of couple stress fluid in a channel with wall properties. *Int. J. Biomath.* **4**, 365–378(2011)
9. Alsaedi, A., Ali, N., Tripathi, D., Hayat, T.: Peristaltic flow of couple stress fluid through uniform porous medium. *Applied Mathematics and Mechanics.* **35**, 469–480(2014)
10. Sankad, G. C., Patil, Asha.: Impact of permeable lining of the wall on the peristaltic flow of Herschel Bulkley fluid. *Applications and Applied Mathematics (AAM)*.**11**, 663–669(2016)
11. Beaver, G.S., Joseph D.D.: Boundary conditions at a naturally permeable wall. *J. Fluid Mech.* **30**, 197–207(1967)
12. Srekanth, A. K.: Slip flow through long circular tubes. In proceedings of the 6th international Symposium on Rarefied Gas Dynamics, L Trilling and H.Y. Wachman, Eds. Academic press, New York, NY,USA. 667–680 (1968)
13. El-Shehawey, E. F., El-Dabe, N. T., El-Desoky, I. M.: Slip effects on the peristaltic flow of a non-Newtonian Maxwellian fluid. *Acta Mechanica.* **186**, 141–159(2006)
14. Sobh, A. M.: Interaction of couple stresses and slip flow on peristaltic transport in a uniform and non-uniform channels. *Turkish J. Eng. Env. Sci.* **32**, 117–123(2008)
15. Ellahi, R.: Effects of the slip boundary condition on non-Newtonian flows in a channel. *Communications in Nonlinear Science and Numerical Simulation.* **14**, 1377–1384(2009)
16. Sankad, G. C., Pratima, S. N.: Influence of the wall properties on the peristaltic transport of a couple stress fluid with slip effects in porous medium. *Procedia Engineering.* **127**, 862–868(2015)
17. Hina, S.: MHD peristaltic transport of Eyring-Powell fluid with heat/mass transfer, wall properties and slip conditions. *J. Magnetism and Magnetic Materials.* **404**, 148–158 (2016)
18. Hayat, T., Ahamed, N., Ali, N.: Effects of an endoscope and magnetic field on the peristalsis involving Jeffrey fluid. *Communications in Nonlinear Science and Numerical Simulation.* **13**, 1581–1591(2008)
19. Mahmouda, S. R., Afifi, N. A. S., Al-Isede, H. M.: Effect of porous medium and magnetic field on peristaltic transport of a Jeffrey fluid. *International Journal of mathematical analysis.* **5(21)**, 1025–1034(2011)
20. Sudhakar Reddy, M., Subba Reddy, M. V., Jayarami Reddy, B. Ramakrishna, S.: Effect of variable viscosity on the peristaltic flow of a Jeffrey fluid in a uniform tube. *Pelagia Research Library, Advances in Applied Science Research.* **3(2)**, 900–908(2012)
21. Subba Reddy, M. V., Jayarami Reddy, B., Nagendra, N., Swaroopa, B.: Slip effects on the peristaltic motion of a Jeffrey fluid through a porous medium in an asymmetric channel under the effect magnetic field. *Journal of Applied Mathematics and Fluid Mechanics.* **4(1)**, 59–72(2012)

22. Arun Kumar, M., Sreenadh, S., Srinivas, A. N. S.: Effects of wall properties and heat transfer on the peristaltic transport of a Jeffrey fluid in a channel. Pelagia Research Library, *Advances in Applied Science Research*. **4(6)**, 159–172 (2013)
23. Dheia, G. Salih Al-Khafajy., Ahmed M. Abdulhadi.: Effects of wall properties and heat transfer on the peristaltic transport of a Jeffrey fluid through porous medium channel. *Mathematical theory and modeling*. **4(9)**, 86–99 (2014)
24. Bhatti, M.M., Ali Abbas, M.: Simultaneous effects slip and MHD on peristaltic blood flow of Jeffrey fluid model through porous medium. *Alexandria Engineering Journal*. **55(2)**, 1017–1023(2016)

Linear and Nonlinear Double Diffusive Convection in a Couple Stress Fluid Saturated Anisotropic Porous Layer with Soret Effect and Internal Heat Source



Kanchan Shakya

1 Introduction

Studies of double diffusive convection in porous media play a significant role in many areas, such as the petroleum industry, solidification of binary mixtures, and migration of solutes in water-saturated soils. Other examples include geophysics systems, crystal growth, electrochemistry, the migration of moisture through air contained in fibrous insulation, the Earth's oceans, and magma chambers. The problem of double diffusive convection in a porous media has been presented by Ingham and Pop [1], Nield and Bejan [2], Vafai [3, 4], and Vadasz [5]. The study was continued by Poulikakos [6], Trevison and Bejan [7], and Momou [8] among others. The first study of double diffusive convection in porous media was mainly concerned with linear stability analysis and was performed by Nield [9].

The growing importance of non-Newtonian fluids with suspended particles in modern technology and industries makes the investigation of such fluids desirable. These fluids are applied in the extrusion of polymer fluids in industry, exotic suspensions, fluid film lubrication, solidification of liquid crystals, cooling of metallic plates in baths, and colloidal and suspension solutions. Non-Newtonian stress fluids have specific features, such as the polar effect. The theory of polar fluids and related theories are models for fluids whose microstructure is mechanically significant. The theory for couple stress fluid was proposed by Stokes [10]; it is a simpler polar fluid theory, that shows all the important features and effects of such fluids that occur inside a deforming continuum. The stabilizing effect of the couple stress parameter is reported in the works of Sharma and Thakur [11], who

K. Shakya (✉)

Department of Applied Mathematics, School for Physical Sciences, Babasaheb Bhimrao Ambedkar University, Lucknow 226025, India
e-mail: kanchan_17mayraj@rediffmail.com

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_27

429

investigated thermal instability in an electrically conducting couple stress fluid with a magnetic field. Sunil et al. [12] studied the effect of suspended particles on double diffusive convection in a couple stress fluid-saturated porous medium, Sharma and Sharma [13] investigated the effect of suspended particles on couple stress fluid, heated from below, in the presence of rotation and a magnetic field. Malashetty et al. [14] performed an analytical study of linear and nonlinear double diffusive convection with the Soret effect in couple stress liquids. Gaikwad and Kamble [15] analyzed the linear stability of double diffusive convection in a horizontal, sparsely packed, rotating, anisotropic porous layer in the presence of the Soret effect. Malashetty and Kollur [16] investigated the onset of double diffusive convection in an anisotropic porous layer saturated with couple stress fluid. Shivakumara et al. [17] analyzed the linear and nonlinear stability of double diffusive convection in a couple stress fluid-saturated porous layer.

In the study of double diffusive convection in the Soret effect, in some of the important areas of application in engineering, including geophysics, oil reservoirs, and groundwater, researchers have developed a great interest in these type of flows. In the presence of cross diffusion two transport properties are produced: the Soret effect and the Dufour effect. The Soret effect describes the tendency of a solute to diffuse under the influence of a temperature gradient. There are only a few studies available on double diffusive convection in the presence of the Soret effect. The diffusion material is heated unevenly. A mixture of gases or a solution is caused by the presence of temperature gradient in the system. The effect was described by Swiss scientist J. Soret, who was the first to study thermodiffusion (1879). Hurler and Jakeman argue that the liquid mixture, the Dufour term, is indeed small, and thus the Dufour effect will be negligible when compared with the Soret effect. They conducted an experimental and theoretical study of Soret-driven thermosolutal convection in a binary fluid mixture [18]. Malashetty et al. [19] performed an analytical study of linear and nonlinear double diffusive convection with the Soret effect in couple stress liquids. Rudraiah and Malashetty [20] discussed double diffusive convection in a porous medium in the presence of the Soret and Dufour effects. Bahloul et al. [21] studied double diffusive convection and Soret-induced convection in a shallow horizontal porous layer analytically and numerically. Malashetty and Biradar [22] carried out an analytical study of linear and nonlinear double diffusive convection in a fluid-saturated porous layer with Soret and Dufour effects. Also in another study, Bhadauria and Hashim et al. [23] performed linear and nonlinear double diffusive convection in a saturated anisotropic porous layer with couple stress fluid. Hill [25] showed linear and nonlinear double diffusive convection in a saturated anisotropic porous layer with a Soret effect and an internal heat source. Bhadauria et al. [26] studied effect of internal heating on double diffusive convection in a couple stress fluid saturated anisotropic porous medium. A study concerning an internal heat source in porous media was provided by Tveitereid [24], who performed thermal convection in a horizontal porous layer with internal heat sources. Srivastava et al. [27] performed linear and nonlinear analyses of double diffusive convection in a porous layer with a concentration-based internal heat source. Bhadauria [28], Horton and Rogers [29], and Lapwood [30] studied

the effect of internal heating on double diffusive convection in a couple stress fluid-saturated anisotropic porous medium. Govender [31] showed that the Coriolis effect on the stability of centrifugally driven convection in a rotating anisotropic porous layer is subject to gravity. Kapil [32] performed at the onset of convection in a dusty couple stress fluid with variable gravity through a porous medium in hydromagnetics.

The aim of this chapter was to study the Soret effect and an internal heat source with a couple stress fluid. However, in the present study, stability analysis of the Soret and internal heating effect on double diffusive convection in an anisotropic porous layer with a couple stress fluid was performed.

1.1 Nomenclature

Table 1

Latin symbols	
a	wave number
C	Couple stress parameter $C = \frac{\mu_c}{\mu d^2}$
Le	Lewis number $Le = \frac{\kappa_T}{\kappa_s}$
d	height of porous layer
\bar{g}	acceleration due to gravity
D	Cross diffusion due to T component
Da	Darcy number $Da = \frac{\kappa_z}{d^2}$
Ra_T	thermal Rayleigh number $Ra_T = \frac{\beta_T g \Delta T K_z d}{\nu \kappa_T z}$
Ra_S	solotal Rayleigh number $Ra_S = \frac{\beta_S g \Delta S K_z d}{\nu \kappa_T z}$
K	permeability of porous medium $K_x(ii + jj) + K_z(kk)$
K_x	permeability in x-direction
K_z	permeability in z-direction
T	temperature
ΔT	temperature difference across the porous layer
t	time
p	reduced pressure
q	fluid velocity(u,v,w)
Pr	Prandtl number $Pr = \frac{\varepsilon \gamma \nu d^2}{\kappa_T K}$
R_i	Internal heat source parameter $R_i = \frac{Qd^2}{\kappa_T}$
V_a	Vadasz number $V_a = (\frac{Pr}{Da})$
S	solute concentration
N_u	Nusselt number
S_h	Sherwood number
ΔS	solute difference across the porous layer
(x,y,z)	space co-ordinates

(continued)

Table 1 (continued)

Greek symbols	
β_T	coefficient of thermal expansion
β_S	coefficient of solute expansion
ξ	mechanical anisotropic parameter
η	thermal anisotropic parameter
κ_s	effective concentration diffusivity
κ_{Tz}	effective thermal diffusivity
σ	dimensionless oscillatory frequency
μ	dynamic viscosity of the fluid
μ_c	couple stress viscosity of the fluid
k	porosity
γ	heat capacity ratio $\frac{(\rho c_p)_m}{(\rho c_p)_f}$
ν	kinematic viscosity ($\frac{\mu}{\rho_0}$)
ρ	fluid density
 Other symbols	
∇_1^2	$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$
∇^2	$\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$
 Subscripts	
b	basic state
c	critical
0	reference value
 Superscripts	
'	perturbed quantity
*	dimensionless quantity
osc	oscillatory
st	stationary

2 Mathematical Formulation

We consider an infinitely extended horizontal plane at $z=0$ and $z=d$ a fluid-saturated porous medium, which is heated from below and cooled from above. The Darcy model has been employed in the momentum equation. Further, an internal heat source term has been included in the energy equation. A cartesian frame of reference is chosen in such a way that the origin lies on the lower plane and the z -axis is

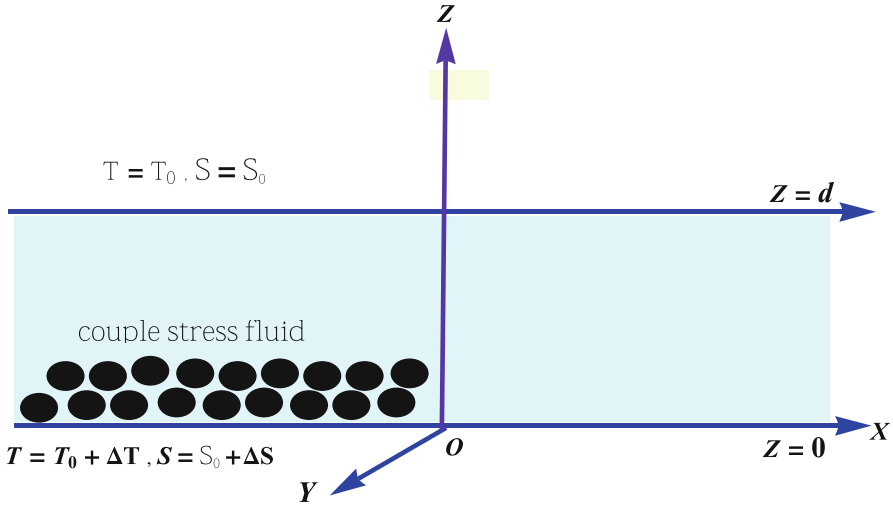


Fig. 1 Physical configuration of the problem

vertical upward. An adverse temperature gradient is applied across the porous layer and the lower and upper planes are kept at temperatures $T_0 + \Delta T$ and T_0 , with a concentration $S_0 + \Delta S$ and S_0 respectively. The physical configuration of the model is reported in the Figure 1. The governing equations are given below

$$\left\{ \begin{array}{l} \nabla \cdot \vec{q} = 0, \\ \frac{\rho_0}{\varepsilon} \left(\frac{\partial \vec{q}}{\partial t} \right) = -\nabla p + \rho g - \frac{1}{K} (\mu - \mu_c \nabla^2) \vec{q}, \\ \gamma \frac{\partial T}{\partial t} + (\vec{q} \cdot \nabla) T = \nabla (\kappa_{Tz} \cdot \nabla T) + Q(T - T_0), \\ \varepsilon \frac{\partial S}{\partial t} + (\vec{q} \cdot \nabla) S = \kappa_s \nabla^2 S + D \nabla^2 T, \\ \rho = \rho_0 [1 - \beta_T (T - T_0) + \beta_S (S - S_0)] \end{array} \right. \quad (1)$$

where the physical variables have their usual meanings as given in the nomenclature. The externally imposed thermal and solutal boundary conditions are given by

$$\left\{ \begin{array}{lll} T = T_0 + \Delta T, & \text{at } z = 0 \text{ and } T = T_0, & \text{at } z = d, \\ S = S_0 + \Delta S, & \text{at } z = 0 \text{ and } S = S_0, & \text{at } z = d, \end{array} \right. \quad (2)$$

3 Basic State

In this state, the velocity, pressure, temperature, and density profiles are given by

$$\vec{q}_b = 0, p = p_b(z), T = T_b(z), S = S_b(z), \rho = \rho_b(z). \quad (3)$$

Substituting Equation (3) in Equation (1), we obtain the following relations:

$$\frac{dp_b}{dz} = -\rho_b g, \quad (4)$$

$$\kappa_T \frac{d^2(T_b - T_0)}{dz^2} + Q(T_b - T_0) = 0, \quad (5)$$

$$K_s \frac{d^2 S_b}{dz^2} + D \frac{d^2 T_b}{dz^2} = 0, \quad (6)$$

$$\rho_b = \rho_0[1 - \beta_T(T_b - T_0) + \beta_S(S_b - T_0)]. \quad (7)$$

The solution of equation (5), subject to the boundary conditions (2), is given by

$$T_b = T_0 + \Delta T \frac{\sin\left(\left(\sqrt{\frac{Qd^2}{\kappa_T}}\right)\left(1 - \frac{z}{d}\right)\right)}{\sin\left(\sqrt{\frac{Qd^2}{\kappa_T}}\right)}. \quad (8)$$

The solution of equation (6), subject to the boundary conditions (2),

$$S_b = S_0 + (\Delta S + \frac{D\Delta T}{K_s})\left(1 - \frac{z}{d}\right) - \frac{D\Delta T}{K_s} \frac{\sin\left(\left(\sqrt{\frac{Qd^2}{\kappa_T}}\right)\left(1 - \frac{z}{d}\right)\right)}{\sin\left(\sqrt{\frac{Qd^2}{\kappa_T}}\right)} \quad (9)$$

Now, we superimpose finite amplitude perturbations on the basic state in the form:

$$\vec{q} = q_b + q', T = T_b + T', p = p_b + p', S = S_b + S', \rho = \rho_b + \rho', \quad (10)$$

Infinitesimal perturbation was applied to the basic state of the system and then the pressure term was eliminated by taking the curl twice of Equation (1). The resulting equations were nondimensional using the following transformations:

$$(x, y, z) = (x^*, y^*, z^*)d, \quad t = t^* \left(\frac{\gamma d^2}{\kappa_{Tz}}\right), \quad (11)$$

$$(u, v, w) = (u^*, v^*, w^*)\left(\frac{\kappa T z}{d}\right), \quad T = (\Delta T)T^*, \quad S = (\Delta S)S^*$$

T_b, S_b in dimensionless forms are given

$$T_b = \frac{\sin \sqrt{R_i}(1-z)}{\sin \sqrt{R_i}}, \tag{12}$$

$$S_b = \frac{S_r L_e R_a T \sin(\sqrt{R_i}(1-z))}{R_a S \sin \sqrt{R_i}} - \left(\frac{S_r L_e R_a T}{R_a S} + 1\right)(1-z)$$

to obtain nondimensional equation (on dropping the asterisks for simplicity), and use the stream function $u = \frac{\partial \psi}{\partial z}, w = -\frac{\partial \psi}{\partial x}$

$$\frac{1}{V_a} \frac{\partial}{\partial t} \nabla_1^2 \psi + \left(\frac{\partial^2}{\partial x^2} + \frac{1}{\xi} \frac{\partial^2}{\partial z^2}\right)(1 - C \nabla_1^2) \psi = Ra_T \frac{\partial T}{\partial x} - Ra_S \frac{\partial S}{\partial x} = 0 \tag{13}$$

$$\left[\frac{\partial}{\partial t} - \frac{\partial^2}{\partial z^2} - \eta \frac{\partial^2}{\partial x^2} - R_i\right]T - f(z) \frac{\partial \psi}{\partial x} - \frac{\partial(\psi, T)}{\partial(x, z)} = 0 \tag{14}$$

$$\left[\frac{\partial}{\partial t} - \frac{1}{L_e} \left(\frac{\partial^2}{\partial z^2} + \frac{\partial^2}{\partial x^2}\right)\right]S - S_r \frac{Ra_T}{Ra_S} \nabla^2 T - b(z) \frac{\partial \psi}{\partial x} - \frac{\partial(\psi, S)}{\partial(x, z)} = 0 \tag{15}$$

where $V_a = \frac{\varepsilon P_r}{D_a}$ is Vadasz number, $Ra_T = \frac{\beta_T g \Delta T K_z d}{\nu \kappa T_z}$ is the thermal Rayleigh number, $Ra_S = \frac{\beta_S g \Delta S K_z d}{\nu \kappa T_z}$ is the solute Rayleigh number, $R_i = \frac{Q d^2}{\kappa T_z}$ is the internal heat source parameter, $C = \frac{\mu C}{\mu d^2}$ is the couple stress fluid, $L_e = \frac{\kappa T_z}{\kappa S}$ is the Lewis number, and $\chi = \frac{\varepsilon}{\gamma}$ is normalized porosity. The above system will be solved by considering stress-free and isothermal boundary conditions as given below:

$$w = \frac{\partial^2 w}{\partial z^2} = T = S = 0 \quad \text{on} \quad z = 0, z = 1. \tag{16}$$

4 Linear Stability Analysis

To study linear stability analysis according to solving the eigenvalue problem defined by Equations (13)–(15) subject to the boundary condition by Equations (5), (6), using time-dependent periodic disturbance in the horizontal plane:

$$(w, T, S) = (W, \Theta, \phi) \exp[i(lx + my) + \sigma t] \tag{17}$$

where l, m are horizontal wave number and $\sigma = \sigma_r + i\sigma_j$ the growth rate. Substituting Equation (17) into the linearized equations (13)–(15), we obtain

$$\left[\frac{\sigma}{V_a} \delta^2 + \delta_1^2 (1 - C\delta^2) \right] W + aRa_T \Theta - aRa_S \phi = 0 \tag{18}$$

$$[\sigma + \eta_1 - R_i] \Theta - 2aFW = 0 \tag{19}$$

$$\left[\sigma + \frac{\delta^2}{L_e} \right] \phi - 2aBW + S_r \delta^2 \frac{Ra_T}{Ra_S} \Theta = 0. \tag{20}$$

Where $D = d/dz$ and $a^2 = l^2 + m^2$. The boundary conditions are (17). Now read

$W = D^2 W = \Theta = \phi = 0$ at $z = 0, 1$:

We assume that the solutions of equations (13)–(15) satisfying the boundary conditions (17),

$(W(z), \Theta(z), \phi(z)) = (W_0, \Theta_0, \phi_0) \sin n\pi z$ ($n = 1, 2, 3, \dots$)

in the form of the thermal Rayleigh number can be obtained as

$$Ra_T = \frac{Ri - (\sigma + \eta_1)}{2a^2 F} \left[\frac{(\delta^2 + L_e \sigma) \left(\frac{\sigma}{V_a} \delta^2 + \delta_1^2 (1 - C\delta^2) \right) - 2a^2 B L_e Ra_S}{\sigma + \delta^2 + \delta^2 S_r L_e} \right] \tag{21}$$

where $a^2 = l^2 + m^2$, $\delta^2 = \pi^2 + a^2$, $\delta_1^2 = \frac{\pi^2}{\xi} + a^2$, $\eta_1 = \pi^2 + \eta a^2$,

$F = \int_0^1 \frac{dT_b}{dz} \sin^2(\pi z) dz$, $B = \int_0^1 \frac{dS_b}{dz} \sin^2(\pi z) dz$, η is a representative viscosity of the fluid. The growth rate σ is in general a complex quantity such that $\sigma = \sigma_r + i\sigma_j$. The system with $\sigma_r < 0$ is always stable, whereas for $\sigma_r > 0$ it will become unstable. For the neutral stability state $\sigma_r = 0$.

4.1 Stationary State

The values of the thermal Rayleigh number and the corresponding wave number of the system for a stationary mode of convection are given below:

$$Ra_T^{st} = \frac{Ri - \eta_1}{2a^2 F} \left[\frac{\delta^2 \delta_1^2 (1 - C\delta^2) - 2a^2 B Ra_S L_e}{\delta^2 (1 + L_e S_r)} \right], \tag{22}$$

It is important to note the critical wave number $a = a_c^{st}$, which is the result given by Malashetty et al. [19]. For single component fluid, $Ra_S = 0$, i.e., in the absence of a solute Rayleigh number, Equation (22) gives

$$Ra_T^{st} = \frac{(Ri - \eta_1) \delta_1^2 (1 - C\delta^2)}{2a^2 F (1 + L_e S_r)}. \tag{23}$$

For the system without internal heating, i.e., $R_i = 0, F = -1/2$, we get

$$Ra_T^{st} = \frac{(\eta_1)\delta_1^2(1 - C\delta^2)}{a^2(1 + L_e S_r)} \tag{24}$$

which is the one obtained by Shivakumara et al. [17]. When $C = 0$ (i.e., Newtonian fluid case), Eq. (3.11) reduces to

$$Ra_T^{st} = \frac{(\pi^2 + \eta^2 a^2)(a^2 + \frac{\pi^2}{\xi})}{a^2(1 + L_e S_r)} \tag{25}$$

In the case of no Soret effect

$$Ra_T^{st} = \frac{(\pi^2 + \eta^2 a^2)(a^2 + \frac{\pi^2}{\xi})}{a^2} \tag{26}$$

Lastly, in the case of isotropic porous medium, put $\eta = \xi = 1$

$$Ra_T^{st} = \left(\frac{\pi^2 + a^2}{a}\right)^2 \tag{27}$$

which has the critical value $Ra_c^{St} = 4\pi^2$ for $a_c^{St} = \pi^2$ and which are the classical results obtained by Horton and Rogers [29] and Lapwood [30].

4.2 Oscillatory State

For the corresponding wave number of the system for the oscillatory mode of convection, we now set $\sigma = i\sigma_i$ in Equation (21) and clear the complex quantities from the denominator, to obtain

$$Ra_T^{osc} = \Delta_1 + i\sigma_i \Delta_2.$$

$$\Delta_1 = \frac{1}{2a^2 F} \frac{A_1 B_1 + \sigma^2 A_2 B_2}{B_1^2 + \sigma^2 B_2^2} \tag{28}$$

$$\Delta_2 = \frac{1}{2a^2 F} \frac{A_2 B_1 - A_1 B_2}{B_1^2 + \sigma^2 B_2^2}, \tag{29}$$

where, $A_1 = (R_i - \eta_1)(\delta^2 \delta_1^2 (1 - C\delta^2) - \frac{\sigma^2}{V_a} L_e \delta^2) + \sigma^2 (L_e \delta_1^2 (1 - C\delta^2) + \frac{\delta^4}{V_a}) - (R_i - \eta_1) Ra_S 2a^2 B L_e,$

$A_2 = (R_i - \eta_1)(L_e \delta_1^2 (1 - C\delta^2) + \frac{\delta^4}{V_a}) - \delta^2 \delta_1^2 (1 - C\delta^2) + \frac{\sigma^2}{V_a} L_e \delta^2 + Ra_S 2a^2 B L_e$

$B_1 = \delta^2 (1 + S_r L_e)$

$B_2 = 1$

For oscillatory onset $\Delta_2 = 0$ and ($\sigma_i \neq 0$), where σ is the oscillatory frequency, which is not given for brevity.

We have the necessary expression for the oscillatory Rayleigh number as:

$$Ra_T^{osc} = \Delta_1. \quad (30)$$

5 Nonlinear Stability Analysis

In this section, we study the nonlinear stability analysis using a minimal truncated Fourier series. For simplicity, we consider only two-dimensional rolls, so that all the physical quantities are independent of y . Consider the stream function ψ such that $u = \frac{\partial \psi}{\partial z}$, $w = -\frac{\partial \psi}{\partial x}$, then taking curl to eliminate the pressure term from Equation (1) and then the resulting nondimensional equations by using transformation given by Equation (11) and the following equation

$$\left(\frac{1}{Va} \frac{\partial}{\partial t} \nabla^2 \psi + \left(\frac{\partial^2}{\partial x^2} + \frac{1}{\chi} \frac{\partial^2}{\partial z^2} \right) (1 - C \nabla^2) \psi \right) + Ra_T \frac{\partial T}{\partial x} - Ra_S \frac{\partial S}{\partial x} = 0, \quad (31)$$

$$\left(\frac{\partial}{\partial t} - \frac{\partial^2}{\partial z^2} - \eta \frac{\partial^2}{\partial x^2} - R_i \right) T - f(z) \frac{\partial \psi}{\partial x} - \frac{\partial(\psi, T)}{\partial(x, z)} = 0, \quad (32)$$

$$\left[\frac{\partial}{\partial t} - \frac{1}{Le} \left(\frac{\partial^2}{\partial z^2} + \frac{\partial^2}{\partial x^2} \right) \right] S - \frac{\partial \psi}{\partial x} b(z) - \frac{\partial(\psi, S)}{\partial(x, z)} - S_r \frac{Ra_T}{Ra_S} \nabla^2 T = 0 \quad (33)$$

It should be noted that the effect of nonlinearity is to distort the temperature and concentration fields through the interaction of ψ and T , ψ , and S . As a result, a component of the form $\sin(2\pi z)$ will be generated, where V is zonal velocity induced by rotation. A minimal Fourier series that describes the finite amplitude convection is given by

$$\psi = A_1(t) \sin(ax) \sin(\pi z), \quad (34)$$

$$T = B_1(t) \cos(ax) \sin(\pi z) + B_2(t) \sin(2\pi z), \quad (35)$$

$$S = C_1(t) \cos(ax) \sin(\pi z) + C_2(t) \sin(2\pi z), \quad (36)$$

where the amplitudes $A_1(t)$, $B_1(t)$, $B_2(t)$, $C_1(t)$, $C_2(t)$ are functions of time and are to be determined. Substituting the above expressions in Equations (31)–(33) and equating the like terms, the following set of nonlinear autonomous differential equations were obtained

$$\frac{dA_1(t)}{dt} = \frac{-V_a}{\delta^2}(\delta^2(1 + C\delta^2)A_1 + aRa_T B_1 - aRa_S C_1) \tag{37}$$

$$\frac{dB_1(t)}{dt} = 2aFA_1 - \pi aA_1 B_2 + (R_i - \eta_1)B_1 \tag{38}$$

$$\frac{dB_2(t)}{dt} = \frac{\pi a}{2}A_1 B_1 + (R_i - 4\pi^2)B_2 \tag{39}$$

$$\frac{dC_1(t)}{dt} = 2aBA_1 - \delta^2 S_r \frac{Ra_T}{Ra_S} B_1 - \delta^2 \frac{1}{L_e} C_1 - \pi aA_1 C_2 \tag{40}$$

$$\frac{dC_2(t)}{dt} = \pi \frac{a}{2} A_1 C_1 - 4\pi^2 S_r \frac{Ra_T}{Ra_S} B_2 - \frac{4\pi^2}{L_e} C_2 \tag{41}$$

where $A = 1 + 4c\pi^2$. The numerical method was used to solve the above nonlinear differential equation to find the amplitudes.

5.1 Steady Finite Amplitude Convection

For steady-state finite amplitude convection we have to set the left-hand side of the Equations (37)–(41) to zero.

$$(\delta^2(1 + C\delta^2)A_1 + aRa_T B_1 - aRa_S C_1) = 0 \tag{42}$$

$$2aFA_1 - \pi aA_1 B_2 + (R_i - \eta_1)B_1 = 0 \tag{43}$$

$$\frac{\pi a}{2}A_1 B_1 + (R_i - 4\pi^2)B_2 = 0 \tag{44}$$

$$2aBA_1 - \delta^2 S_r \frac{Ra_T}{Ra_S} B_1 - \delta^2 \frac{1}{L_e} C_1 - \pi aA_1 C_2 = 0 \tag{45}$$

$$\pi \frac{a}{2} A_1 C_1 - 4\pi^2 S_r \frac{Ra_T}{Ra_S} B_2 - \frac{4\pi^2}{L_e} C_2 = 0 \tag{46}$$

on solving for the amplitudes in terms of A_1 , we obtain

$$B_1 = \frac{4aF(z)(4\pi^2 - R_i)A_1}{a^2 A_1^2 \pi^2 - 8\pi^2 R_i + 2R_i^2 + 8\pi^2 \eta_1 - 2R_i \eta}$$

$$B_2 = \frac{2a^2 F(z)\pi A_1^2}{a^2 A_1^2 \pi^2 - 8\pi^2 R_i + 2R_i^2 + 8\pi^2 \eta_1 - 2R_i \eta_1},$$

$$C_1 = \frac{16(8A_1 B L_e \pi^2 R_a S R_i a + 2A_1 B L_e R_a S R_i^2 a + A_1^3 B L_e \pi^2 R_a S a^3 + A_1^3 F L_e^2 \pi^2 R_a T S a^3 - 8A_1 F L_e \pi^2 R_a T S_r a \delta^2 + 2A_1 F L_e R_a T R_i S_r a \delta^2 + 8A_1 B L_e \pi^2 R_a S a \eta - 2A_1 B L_e R_a S R_i a \eta_1)}{R_a S (A_1^2 L_e^2 a^2 + 8\delta^2)(-8\pi^2 R_i + 2R_i^2 + A_1^2 \pi^2 a^2 + 8\pi^2 \eta_1 - 2R_i \eta_1)},$$

$$C_2 = \frac{2(8A_1^2 B L_e^2 \pi^2 R_a S R_i a^2 + 2A_1^2 B L_e^2 R_a S R_i^2 a^2 + A_1^4 B L_e^2 \pi^2 R_a S a^4 - 8A_1^2 F L_e \pi^2 R_a T S a^2 \delta^2 - 8A_1^2 F L_e^2 \pi^2 R_a T S a^2 \delta^2 + 2A_1^2 F L_e^2 R_a T R_i S a^2 \delta^2 + 8A_1^2 B L_e^2 \pi^2 R_a S a^2 \eta_1 - 2A_1^2 B L_e^2 R_a S R_i a^2 \eta_1)}{R_a S \pi (A_1^2 L_e^2 a^2 + 8\delta^2)(-8\pi^2 R_i + 2R_i^2 + A_1^2 \pi^2 a^2 + 8\pi^2 \eta_1 - 2R_i \eta_1)}.$$

To solve the above equation, a quadratic equation in $\frac{A_1^2}{8}$ is given by

$$a_0 x^2 + a_1 x + a_2 = 0$$

$$\text{where } x = \frac{A_1^2}{8},$$

$$a_0 = L_e^2 a^4 \pi^2 \delta_1^2 R_a S (1 + C \delta^2)$$

$$a_1 = \frac{1}{4} \delta_1^2 R_a S (1 + C \delta^2) (R_i - \eta_1) L_e^2 a^2 (R_i - 4\pi^2) - \frac{1}{2} (R_i - 4\pi^2) F R_a T R_a S L_e^2 a^4 - 2L_e a^4 \pi^2 R_a S (B + L_e F S_r) + a^2 \pi^2 \delta^2 \delta_1^2 R_a S (1 + C \delta^2)$$

$$a_2 = \frac{(R_i - 4\pi^2)}{4} (\delta^2 \delta_1^2 R_a S (1 + C \delta^2) (R_i - \eta_1) - 2L_e a^2 B R_a S (R_i - \eta_1) - 2a^2 \delta^2 F R_a S (L_e S_r + R_a T))$$

The required root of the above equation is

$$x = \frac{-a_1 + \sqrt{a_1^2 - 4a_0 a_2}}{2a_0}$$

5.2 Steady Heat and Mass Transport

In the study of this type of problem, quantification of heat and mass transport is very important in porous media. Let Nu and Sh be noted as the rate of heat and mass transport per unit for the fluid phase.

The Nusselt number and Sherwood number are defined by

$$Nu = 1 + \left[\frac{\int_0^{2\pi/a} \frac{\partial T}{\partial z} dx,}{\int_0^{2\pi/a} \frac{\partial T_b}{\partial z} dx,} \right]_{z=0} \tag{47}$$

$$Sh = 1 + \left[\frac{\int_0^{2\pi/a} \frac{\partial S}{\partial z} dx,}{\int_0^{2\pi/a} \frac{\partial S_b}{\partial z} dx,} \right]_{z=0} \tag{48}$$

substituting the value of $T, T_b, S,$ and S_b in Equations (47)–(48),

$$Nu = 1 - \frac{2\pi B_2}{\sqrt{R_i} \cot \sqrt{R_i}}, \tag{49}$$

$$Sh = 1 - \frac{2\pi C_2 Ra_S \sin \sqrt{R_i}}{-S_r Ra_T \cos \sqrt{R_i} \sqrt{R_i} + \sin \sqrt{R_i} Ra_S + \sin \sqrt{R_i} S_r Ra_T}$$

substituting B_2, C_2 of Equations (5.1) into (49) gives

$$Nu, Sh \tag{50}$$

6 Results and Discussion

This chapter investigates the combined effect of internal heating and the Soret effect on stationary and oscillatory convection in a anisotropic porous medium with couple stress fluid. In this section, we discuss the effects of the parameters in the governing equations on the onset of double diffusive convection numerically and graphically. The stationary and oscillatory expressions for different values of the parameters such as the Vadasz number, the couple stress parameter, the solute Rayleigh number, the mechanical anisotropic parameter, and the thermal anisotropic parameter are computed, and the results are depicted in the figures. The neutral stability curves in the (Ra_T, a) plane for various parameter values are shown in Figure 2a–e. We fixed the values for the parameters as $Va = 5, C = 2, Ra_S = 100, L_e = 20, \xi = .5, \eta = .5, S_r = .05,$ and $R_i = 2,$ except for the varying parameter. The effect of the Vadasz number Ta on the neutral curves is shown in Figure 2. We find that for fixed values of all other parameters, the minimum value of the Rayleigh number for the oscillatory mode increases as a function of increasing $Va,$ indicating that the effect of the Vadasz number is to stabilize the system. In addition, the critical wave number increases with increasing $Va.$ We observed that by increasing the value of internal heat source $R_i,$ the mechanical anisotropic parameter ξ decreased the stationary and oscillatory Rayleigh number, which means that the internal heat source $R_i,$ mechanical anisotropic parameter ξ destabilized. Figure 2 depicts the effect of the couple stress parameter C on the neutral stability curves. We find that with an increase in the value of the couple stress parameter, the value of the Rayleigh number for both stationary and oscillatory mode is enhanced, indicating that it stabilizes the onset of double diffusive convection and depicts the effect of the solute Rayleigh number Ra_S on the stability curve for stationary and oscillatory convection. We show that the effect of increasing Ra_S is to decrease the value of the Rayleigh number for stationary and oscillatory convection and the corresponding wave number. Thus, the solute Rayleigh number becomes unstable. We also show that the effect of an increasing Lewis number L_e and the thermal anisotropic

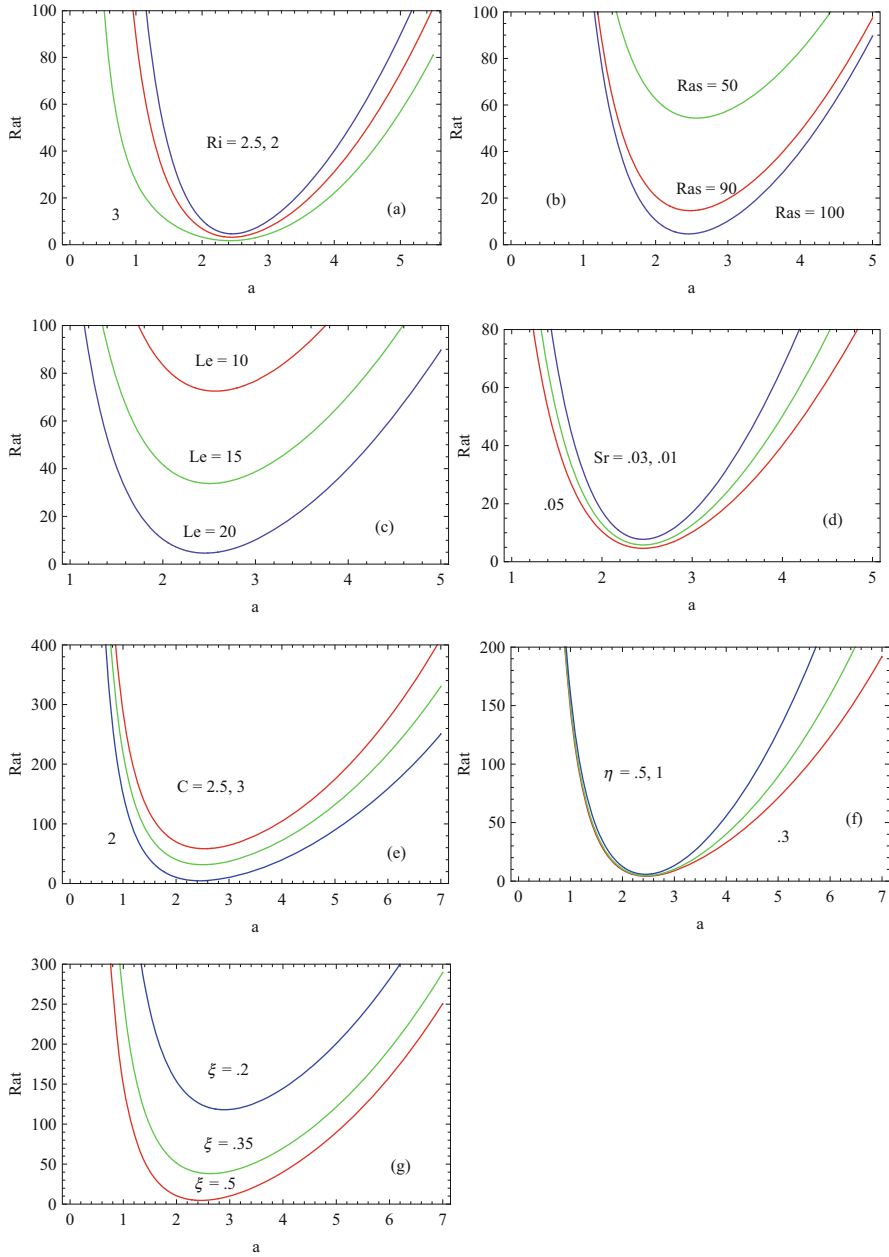


Fig. 2 Stationary neutral stability curves for the different values of (a), (b), (c), (d), (e), (f), (g)

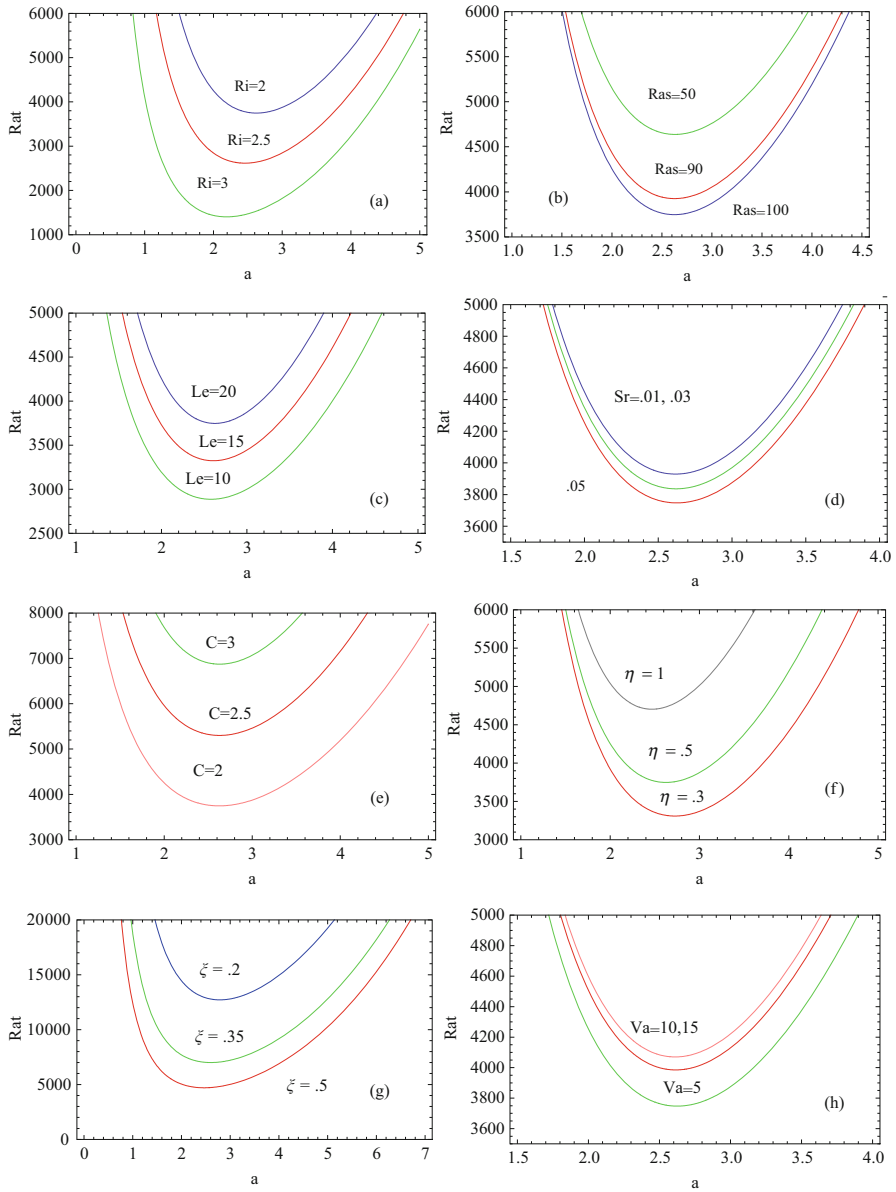


Fig. 3 Oscillatory neutral stability curves for the different values of (a), (b), (c), (d), (e), (f), (g), (h)

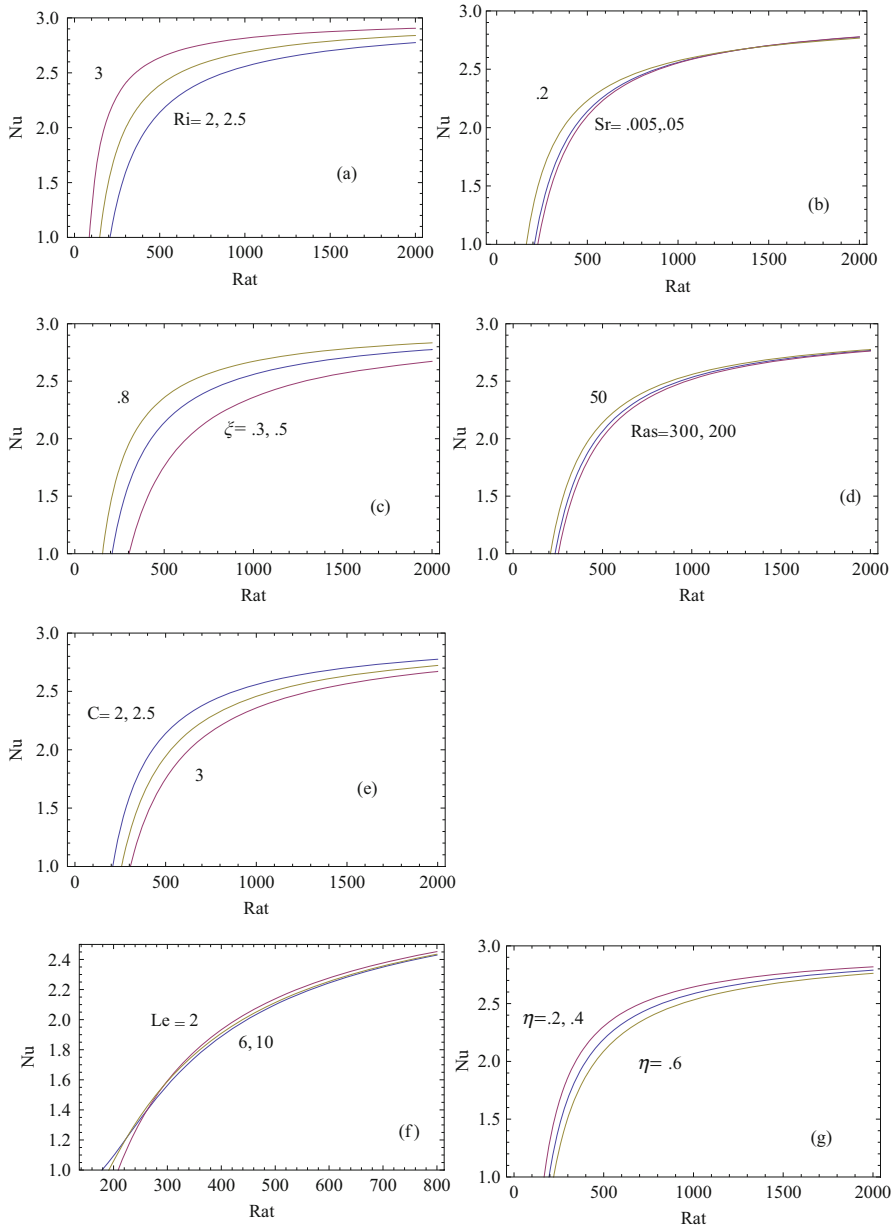


Fig. 4 Variation of Nu with Rat for different values of parameters

parameter η is to increase the value of the Rayleigh number for stationary convection and decrease the value of oscillatory convection. With regard to the corresponding wave number, we found it unstable for the stationary and stable for the oscillatory

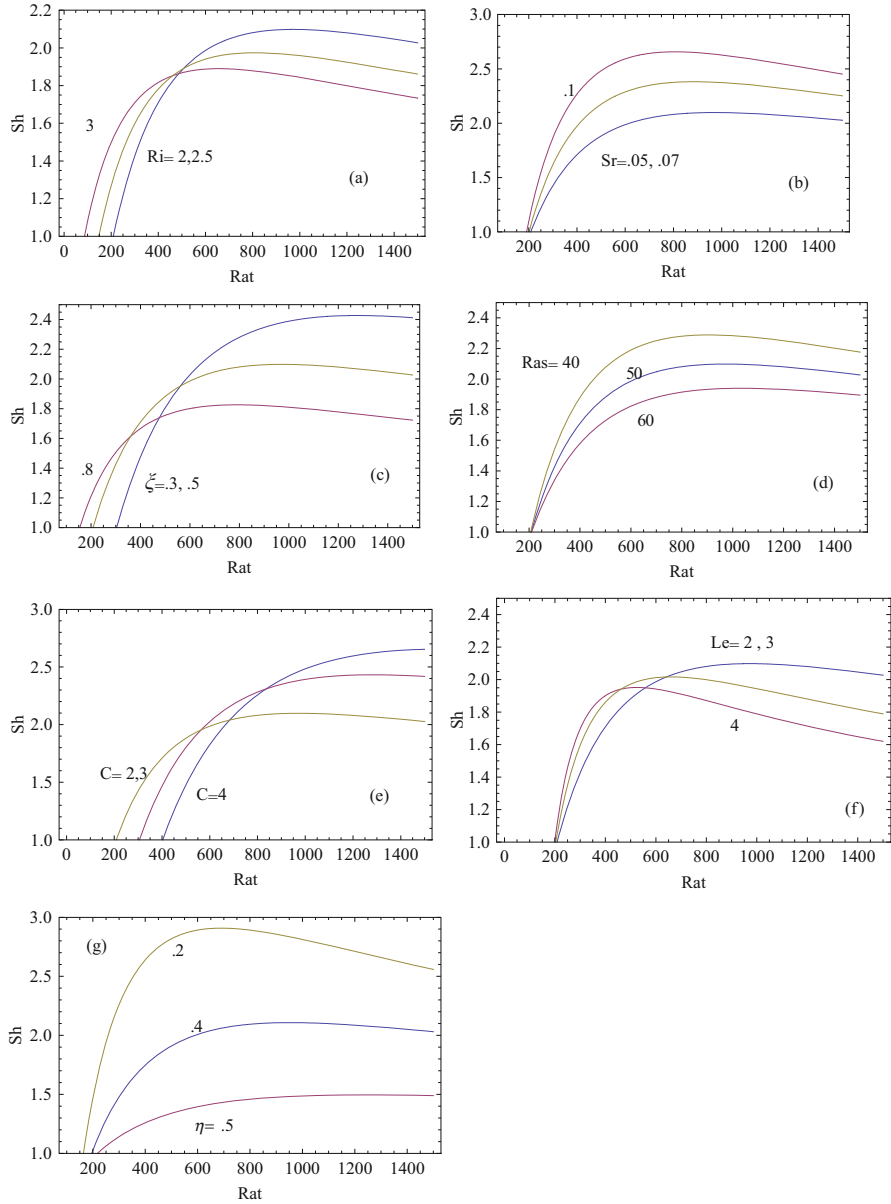


Fig. 5 Variation of Sh with Rat for different values of parameters

modes. We find that Figures 2 and 3 show that an increase in the value of the Soret parameter S_r decreases the Rayleigh number, indicating that the Soret parameter destabilizes the onset of stationary and oscillatory convection.

We use the parameter in a graph of the Nusselt and Sherwood number $C = 2$, $Ra_S = 20$, $L_e = 2$, $\xi = .5$, $\eta = .5$, $S_r = .05$, and $R_i = 2$, and Figures 4a and 5c show that an increase in the value of the internal Rayleigh number R_i decreases the rate of heat and increases mass transfer. We note that the effect of increasing the solute Rayleigh number Ra_S and the thermal anisotropic parameter η is to increase the value of the Nusselt number N_u and the Sherwood number S_h , thus reducing the heat and mass transfer. In Figures 4b and 5a, it can be found that with an increase in the value of the Soret parameter S_r , the mechanical anisotropic parameter ξ and then the value of the Nusselt number N_u and the Sherwood number S_h decrease; thus, the heat and mass transfer across the porous layer also decrease.

7 Conclusions

The Soret effect and the internal heating effect on double diffusive convection in a anisotropic porous medium saturated with a couple stress fluid that is heated and salted from below was investigated using linear and nonlinear stability analysis. The linear analysis is carried out using the normal mode technique. The following conclusions were drawn:

- 1) The Vadasz number Va has a stabilizing effect on oscillatory convection.
- 2) The internal heat parameter R_i , the solute Rayleigh number Ra_S , the Soret parameter S_r , and the mechanical anisotropic parameter ξ destabilize the system in the stationary and oscillatory modes.
- 3) The couple stress fluid C has a stabilizing effect on both the stationary and the oscillatory convection.
- 4) The normalized porosity parameter η and the Lewis number L_e have a destabilizing effect in the case of stationary and opposite oscillatory convection.
- 5) With the increasing value of the mechanical anisotropic parameter ξ , the Soret parameter S_r then increases the value of the Nusselt number N_u , i.e., increasing heat transfer, but increasing the value of the internal Rayleigh number R_i , and the normalized porosity parameter η and the solutal Rayleigh number Ra_S decrease the value of the Nusselt number N_u .
- 6) Mass transfer S_h increases with the increasing value of the internal Rayleigh number R_i , the mechanical anisotropic parameter ξ , the Soret parameter S_r , and decreases with the normalized porosity parameter η and the solutal Rayleigh number Ra_S .

Acknowledgements The author Kanchan Shakya gratefully acknowledges the financial assistance from Babasaheb Bhimrao Ambedkar University, Lucknow, India as UGC research fellowships.

References

1. Ingham D. B, Pop I. eds. *Transport Phenomena in Porous Media*, vol. III, 1st edn. Elsevier, Oxford 2005.
2. Nield D. A., Bejan, A. *Convection in Porous Media*. 3rd edn. Springer, New York 2013.
3. Vafai K. ed. *Handbook of Porous Media*. Marcel Dekker, New York 2000.
4. Vafai K. ed. *Handbook of Porous Media*. Taylor and Francis (CRC), Boca Raton 2005.
5. Vadasz P. ed. *Emerging Topics in Heat and Mass Transfer in Porous Media*. Springer, New York 2008.
6. Poulikakos D., "Double diffusive convection in a horizontally sparsely packed porous layer," *Int. Commun. Heat Mass Transf.*, vol.13, pp. 587–598, 1986.
7. Trevisan O. V. and Bejan A., "Mass and heat transfer by natural convection in a vertical slot filled with porous medium," *Int. J. Heat Mass Transf.*, vol. 29, pp. 403–415, 1986.
8. Mamou M., "Stability analysis of double-diffusive convection in porous enclosures," *Transport Phenomena in Porous Media II* ed D B Ingham and I Pop (Oxford: Elsevier), pp. 113–54, 2002.
9. Nield D. A., "Onset of thermohaline convection in a porous medium," *Water Resour. Res.*, vol. 4, iss. 4, pp. 553–560, 1968.
10. Stokes V. K., "Couple stresses in fluids," *Phys. Fluids*, vol. 9, pp. 1709–1716, 1966.
11. Sharma R. C. and Thakur K. D., "On couple stress fluid heated from below in porous medium in hydrodynamics," *Czechoslov. J. Phys.*, vol. 50, iss. 6, pp. 753–758, 2000.
12. Sunil, Sharma R. C., Chandel R. S., "Effect of Suspended Particles on Couple-Stress Fluid Heated and Solute from Below in Porous Medium," *J. Porous Media*, vol. 7, iss. 1, pp. 9–18, 2004.
13. Sharma R. C., and Sharma M., "Effect of suspended particles on couple-stress fluid heated from below in the presence of rotation and magnetic field," *Indian J. Pure Appl. Math.*, vol. 35, pp. 973–989, 2004.
14. Malashetty M. S., Gaikwad S. N., Swamy M., "An analytical study of linear and non-linear double diffusive convection with Soret effect in couple stress liquids," *Int. J. Therm. Sci.*, vol. 45, iss. 9, pp. 897–907, 2006.
15. Gaikwad S. N. and Kamble S. S., "Linear stability analysis of double diffusive convection in a horizontal sparsely packed rotating anisotropic porous layer in the presence of Soret effect." *J. Applied Fluid Mech.*, vol. 7, pp. 459–471, 2014.
16. Malashetty M. S. and Kollur P., "The onset of Double Diffusive convection in a Couple stress fluid saturated anisotropic porous layer," *Transp. Porous Med.*, vol. 86, pp. 435–459, 2011.
17. Shivakumara I. S., Lee J., Suresh Kumar S., "Linear and nonlinear stability of double diffusive convection in a couple stress fluid-saturated porous layer," *Arch Appl Mech.*, vol. 81, pp. 1697–1715, 2011.
18. Hurle D. T., Jakeman E., "Soret driven thermosolutal convection", *J. Fluid Mech.* 47 667–687, 1971.
19. Malashetty M. S., Gaikwad S. N., Swamy M., "An analytical study of linear and nonlinear double diffusive convection with Soret effect in couple stress liquid," *Int. J. Thermal Sci.* 45 897–907, 2016.
20. Rudraiah N., Malashetty M. S., "The influences of coupled molecular diffusive on double diffusive convection in a porous medium," *ASME, J. Heat Transfer* 108 872–878, 1986.
21. Bahloul A., Boutana N., Vasseur P., "Double diffusive convection and Soret induced convection in a shallow horizontal porous layer," *Fluid Mech.* 491 325–352, 2003
22. Malashetty M. S., Biradar B. S., "Linear and nonlinear double diffusive convection in a fluid saturated porous layer with cross diffusion effect," *Transp. Porous Media* 91 649–670, 2012.
23. Altawallbeh A. A., Bhadauria B. S., Hashim I., "Linear and nonlinear effect of rotation on the onset of double diffusive convection in a Darcy porous medium saturated with a couple stress fluid," *Int. J. Heat Mass Transf.*, 59 103–111, 2013.
24. Tveitereid M., "Thermal convection in a horizontal porous layer with internal heat sources," *Int. J. Heat Mass Transf.*, vol. 20, pp. 1045–1050, 1977.

25. Hill A. A., "Double-diffusive convection in a porous medium with a concentration based internal heat source," *Proc. R. Soc.*, vol. A 461, pp. 561–574, 2005.
26. Bhadauria B. S., Kumar A., Kumar J., Sacheti N. C., Chandran P., "Natural convection in a rotating anisotropic porous layer with internal heat," *Transp. Porous Medium*, vol. 90, iss. 2, pp. 687–705, 2011.
27. Srivastava A., Bhadauria B. S., Hashim I., "Effect of internal heating on double diffusive convection in a couple stress fluid saturated anisotropic porous medium," *Adv. Mater. Sci. Appl.*, vol. 3, iss. 1, pp. 24–45, 2014.
28. Bhadauria B. S., "Double diffusive convection in a saturated anisotropic porous layer with internal heat source," *Transp. Porous Med.*, vol. 9, pp. 299–320, 2012.
29. Horton C. W., and Rogers F. T., "Convection currents in a porous medium", *J. Appl. Phys.*, vol. 16, pp. 367–370, 1945.
30. Lapwood E. R., "Convection of a fluid in a porous medium", *Proc. Camb. Philol. Soc.*, vol. 44, pp. 508–521, 1948.
31. Govender, S., Coriolis effect on the stability of centrifugally driven convection in a rotating anisotropic porous layer subject to gravity. *Transp. Porous Media* 69, 55–66, 2007.
32. Kapil, C., On the onset of convection in a dusty couple stress fluid with variable gravity through a porous medium in hydromagnetics. *J. Appl. Fluid Mech.* 8, 55–63, 2015.

Modeling of Wave-Induced Oscillation in Pohang New Harbor by Using Hybrid Finite Element Model



Prashant Kumar, Rupali, and Rajni

2010 Mathematics Subject Classification 65M60, 65L60, 76B15

1 Introduction

In coastal engineering, major task is to construct a new harbor or redesign an existing harbor in such a way that to obstruct the incoming sea waves and protect the offshore structures. In harbor planning, harbor resonance is the biggest problem induced by long waves that causes losses such as damaging ships, delaying in loading and unloading of cargo, breaking of harbor boundaries, and many more. To ensure safe environment for the mooring system, it is required to analyze the behavior of wave field inside and outside the harbor. In fluid dynamics, the mild-slope equation describes the combined effect of refraction and diffraction for water waves propagating over mild slopes of sea bed. The mild-slope equation is treated as an efficient model to study the refraction and diffraction of the linear and

This research work is supported by the Science Engineering Research Board (SERB), Department of Science and Technology (DST), Government of India, under the Project grant no. “ECR/2016/001680” at National Institute of Technology, Delhi.

P. Kumar (✉) · Rupali

Department of Applied Sciences, National Institute of Technology Delhi, Delhi, India
e-mail: prashantkumar@nitdelhi.ac.in; rupali@nitdelhi.ac.in

Rajni

Jindal Global Business School, O P Jindal Global University, Sonapat, Delhi, India
e-mail: rajni@jgu.edu.in

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_28

449

nonlinear waves in water of variable depth. The mild-slope equation was initially derived by Berkhoff [1]. After that, many modified versions of mild-slope equation were presented by several investigators to study the wave problem [2–4]. In the past few years, many researchers have investigated the wave behavior in a harbor using various numerical schemes [5–7, 11] and predicted the wave field under the resonance conditions. The finite element method (FEM) was used by Demirbilek [7], Xing [8], Ham [9], Woo [10], and Zelt Raichlen [11] to investigate wave behavior in irregular-shaped harbor with variable bathymetry. Further, a combined method termed as HFEM is formulated, which was applied on offshore structures to study wave scattering problems [12–14]. In 2012, Bellotti [15] presented a finite element model based on linear shallow water equations and converts the time-dependent problem into an eigenvalue one by applying radiation condition at the interface boundary instead of imposing zero surface elevation. Another approach based on FEM was investigated by Jung [16] in which domain is subdivided into three regions and Galerkin approach is utilized in the varying bathymetry region. A numerical investigation of transverse oscillations within the harbor with small bottom slope was examined by Wang et al. [17]. In this paper, the hybrid method based on mild-slope equation has been utilized in which finite element has been taken over the bounded region of variable depth and Fourier Bessel solution is taken for the constant depth unbounded region. Further, hybrid triangular mesh has been considered for discretization of domain. The numerical scheme is implemented on rectangular domain with fully reflecting boundaries, to observe the convergence behavior for different discretization. After that, the amplification factor has been obtained for a realistic PNH with fully reflecting boundaries under the resonance conditions. Moreover, the present numerical scheme provides an efficient tool to investigate the strong and weak wave field regions in an irregular-shaped geometry.

2 Mathematical Formulation

The model geometry of the computational domain is shown in Figure 1. The geometry is divided into two regions, i.e., bounded and unbounded region. The bounded region consists of harbor boundary, and interior of the harbor and open sea region consists of ocean area outside the bounded region. The region Ω_1 denotes the bounded region with variable depth and Ω_R is the unbounded region with constant depth.

The mild-slope equation is obtained from energy conservation principle where c and c_g represent the phase and group velocity, ω is angular frequency, and ϕ is the unknown potential function. In the unbounded region, MSE reduces to the Helmholtz equation $(\Delta^2 + k^2)\phi = 0$, where k is the wave number defined by the dispersion relation $\omega^2 = gk \tanh kh$. The potential function in the outer region is given by the equation $\phi = \phi_{inc} + \phi_{sct}$, where ϕ_{inc} denotes the incident velocity potential and ϕ_{sct} refers to the scattered velocity potential.

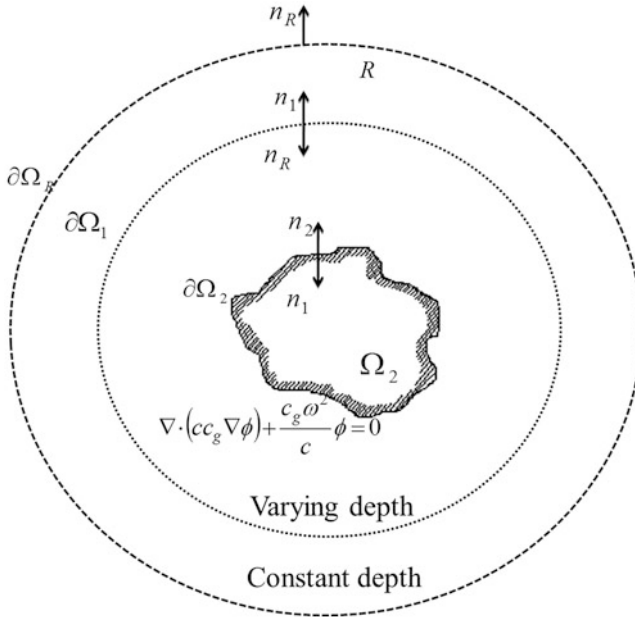


Fig. 1 Model geometry of the harbor problem is shown, the bounded and unbounded region are denoted by Ω_1 and Ω_R , respectively with pseudo boundary $\partial\Omega_1$. Further, n_1, n_2, n_R represent the outward normal vector.

The scattered wave satisfies Helmholtz equation in the unbounded region as well as the radiation condition $\lim_{r \rightarrow \infty} \sqrt{r} \left(\frac{\partial}{\partial r} - ik \right) \phi_{sct} = 0$. Hence, its solution is given in series form as:

$$\phi_{sct} = \sum_{n=0}^{\infty} H_n^1(kr) (\alpha_n \cos n\theta + \beta_n \sin n\theta) \tag{1}$$

where $H_n^1(kr)$ is the Hankel function of n^{th} order of first kind. The potential function is obtained using the matching boundary conditions at the boundary $\partial\Omega_1$, that is:

$$\phi_1 = \phi_R, \quad \frac{\partial \phi_1}{\partial n_1} = \frac{\partial \phi_R}{\partial n_1} \tag{2}$$

The variational principle technique of the finite element method is applied in the bounded domain Ω_1 . The functional obtained by applying the energy conservation approach in bounded as well as unbounded regions is given as follows:

$$F_H(\phi) = \frac{1}{2} \int \int_{\Omega_1} \left[cc_g (\nabla \phi_1)^2 - \frac{c_g \omega^2}{c} \phi_1^2 \right] dx dy - \int_{\partial\Omega_1} cc_g \phi_1 \frac{\partial \phi_1}{\partial n_1} ds - \frac{1}{2} \int_{\partial\Omega_2} i\beta \omega c_g \phi_1^2 ds - \frac{1}{2} \int_{\partial\Omega_1} cc_g \phi_{sct} \frac{\partial(\phi_R L - \phi_{inc})}{\partial n_R} ds - \int_{\partial\Omega_1} cc_g \phi_{inc} \frac{\partial \phi_{sct}}{\partial n_R} ds = C \tag{3}$$

By taking the basis function as linear triangular element, the potential function for each element is written as:

$$\phi^e = \sum_{j=1}^3 N_j(x, y)\phi_j = N^T \phi, \tag{4}$$

where the matrices $N^T = [N_1 N_2 N_3]$ and $\phi^T = [\phi_1 \phi_2 \phi_3]$ denote the basis vector and unknown potential vector. Therefore, $\nabla \phi^e = \sum_{j=1}^3 \nabla N_j(x, y)\phi_j = D^T \phi$ and is the vector of gradient of basic function:

$$\phi_s^e = \sum_{j=1}^3 N_j \phi_{sj} = N^T \phi_s, \quad \frac{\partial \phi_s}{\partial r} = \sum_{j=1}^3 \frac{\partial N_j}{\partial r} \phi_{sj} = \sum_{j=1}^3 Q_{sj} \phi_{sj} = Q_s^T \phi_s. \tag{5}$$

Using the finite element discretization and taking first variation of the functional to be zero, the equation reduces to the matrix equation:

$$[K][\psi] + [U] = [0] \tag{6}$$

Here, $[\psi]$ is the unknown matrix and $[K]$ is stiffness matrix defined as:

$$[K] = \begin{bmatrix} [S_1] & [S_2] \\ [S_2]^T & [S_3] \end{bmatrix}, \tag{7}$$

where the matrix components are described as:

$$S_1 = \int \int_{\Omega_1} \left[cc_g D D^T - \frac{c_g \omega^2}{c} N N^T \right] dx dy - \int_{\partial \Omega_2} i c_g \beta \omega N N^T ds \tag{8}$$

$$S_2 = \int_{\partial \Omega_1} cc_g Q_s N_s^T ds \tag{9}$$

$$S_3 = - \int_{\partial \Omega_1} cc_g N Q_s^T ds \tag{10}$$

And, the matrix $[U_1]$ and $[U_2]$ is given by following integral:

$$U_1 = - \int_{\partial \Omega_1} cc_g N \frac{\partial \phi_{inc}}{\partial n_1} ds \tag{11}$$

$$U_2 = \int_{\partial \Omega_1} cc_g Q_s \phi_{inc} ds \tag{12}$$

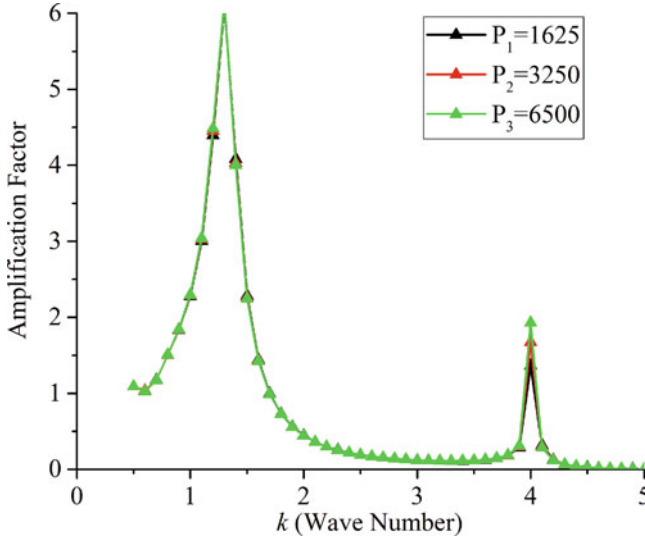


Fig. 2 Amplification factor for $P_1 = 1625$, $P_2 = 3250$, and $P_3 = 6500$ discrete hybrid elements. The abscissa is considered as nondimensional wave number.

3 Convergence of the Numerical Method

The convergence of the numerical model for the rectangular harbor is estimated. The model domain is discretized into P_1 , P_2 , and P_3 number of hybrid elements. In this case, $P_1 = 1625$, $P_2 = 3250$, and $P_3 = 6500$ discrete elements with 70, 140, and 210 elements are lying on the pseudo boundary that separate the computational domain from the open unbounded region. The amplification factor with respect to nondimensional wave number k ($=KL$, where K is wave number and L is characteristic length of the harbor) is estimated for different number of discretization (see Figure 2) and are represented by black, red, and green lines, respectively. The numerical scheme shows high accuracy and the convergence of the solution as the discretization increases.

4 Numerical Simulation Results

The numerical scheme is applied on realistic PNH which is located in South Korea. Firstly, the PNH is discretized into a number of discrete triangular elements and then amplification factor is computed at two key locations WS-01 and WS-02 with respect to nondimensional wave numbers for the incoming wave coming with an incident angle $\alpha = \pi/8$ and $\alpha = \pi/2$.

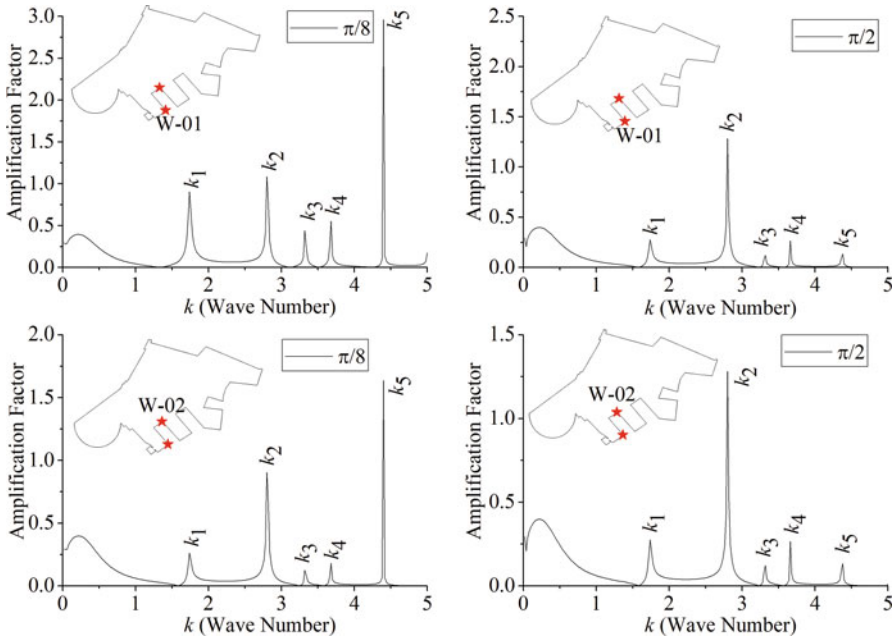


Fig. 3 Amplification factor graph with respect to nondimensional wave number k shows the resonance peaks at wave numbers $k_1 = 1.74, k_2 = 2.80, k_3 = 3.30, k_4 = 3.70,$ and $k_5 = 4.40$.

The resonance peaks are obtained at the five nondimensional wave numbers $k_1 = 1.74, k_2 = 2.80, k_3 = 3.30, k_4 = 3.70,$ and $k_5 = 4.40$ with frequency difference of 0.001. The frequency corresponding to these peaks are resonance frequency, and the incident waves coming with these resonance frequencies could cause many disasters such as damaging of harbor boundary, damaging coastal regions, breaking of mooring lines, and many more. Figure 3 represents the changes in the amplification by taking the incident angle variation and it is observed that the amplification for the incident angle $\alpha = \pi/8$ is high for the second ($k_2 = 2.80$) and fifth ($k_5 = 4.40$) resonance modes whereas for the incident angle $\alpha = \pi/2$ amplification is maximum at the second resonance mode ($k_2 = 2.80$) in both record stations. The wave amplification for the incident angle $\alpha = \pi/8$ is higher as compared to the incident wave arriving with angle $\alpha = \pi/2$. The second and fifth resonance peaks show high amplification on the boundary of PNH. The resonance modes obtained by amplification curve are significant to observe the wave behavior in the PNH.

Wave field contours are shown in Figure 4 corresponding to the incident angle $\alpha = \pi/2$, which indicates the region of strong and weak wave field. Therefore, the present numerical scheme can be applied to any irregular-shaped harbor to analyze the wave behavior.

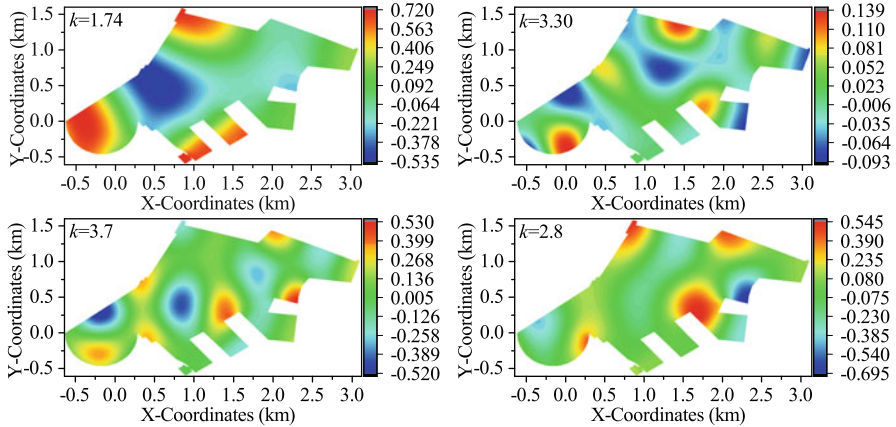


Fig. 4 Wave field inside the PNH for incident angle $\alpha = \pi/2$ corresponding to wave frequency $k_1 = 1.74, k_2 = 2.80, k_3 = 3.30,$ and $k_4 = 3.70$.

5 Conclusion

The numerical scheme based on mild-slope equation is obtained by utilizing the hybrid finite element method. Further, the scheme is applied to analyze the convergence behavior of rectangular harbor by employing different discrete discretization. Moreover, realistic Pohang New Harbor is modeled to predict the amplification factor at various key locations for different directional incident waves. It is observed that the direction of the incident wave drastically affects the wave oscillation in a harbor. The analysis of wave field is significantly important to determine the safe location for the mooring vessels, cargo, and many more inside the harbor. So, the present numerical scheme provides an efficient tool to predict the wave behavior and it can be applied to any irregular geometry domain.

References

1. JCW. Berkhoff, *Computation of combined refraction diffraction*, Proc. 13th Conf. Coast. Eng. Vancouver, Canada, ASCE (1972), 471–90.
2. P. Chamberlain, D. Porter, *The modified mild-slope equation*, J Fluid Mech.291 (1995), 393–407.
3. Y. Toledo, T W. Hsu, A. Roland, *Extended time-dependent mild-slope and wave-action equations for wave-bottom and wave-current interactions*, Proc R Soc A Math Phys Eng Sci.468 (2012), 184–205.
4. P. Kumar, H. Zhang, IK. Kim, Y. Shi, DA. Yuen, *Wave spectral modeling of multidirectional random waves in a harbor through combination of boundary integral of Helmholtz equation with Chebyshev point discretization*, Comput Fluids 108 (2015),13–24.

5. A. Cerrato, JA. Gonzalez, L. Rodriguez-Tembleque, *Boundary element formulation of the Mild-Slope equation for harmonic water waves propagating over unidirectional variable bathymetries*, Eng Anal Bound Elem. 62 (2016), 22–34.
6. P. Kumar, H. Zhang, DA. Yuen, IK. Kim, *Wave field analysis in a harbor with irregular geometry through boundary integral of Helmholtz equation with corner contributions*, Comput Fluids.88 (2013), 287–97.
7. Z. Demirebilek, V. Panchang, *CGWAVE: a coastal surface water wave model of mild slope equation*, US army Engr., waterways experiment station, Vicksburg (1998) Tech rep. CHL-98-26.
8. X. Xing, *Computer Modeling for Wave Oscillation Problems in Harbors and Coastal Regions*, Ph.D. thesis, (2009).
9. S. Ham, K. Bathe, *A finite element method enriched for wave propagation problems*, Comput Struct.94 (2012),1–12.
10. S. Woo, P.L. Liu, *Finite-element model for modified Boussinesq equations. II: Applications to nonlinear harbor oscillations*, J Waterw PORT, Coast Ocean Eng.130 (2004),17–28.
11. J.A. Zelt, F. Raichlen, *A Lagrangian model for wave-induced harbour oscillations*, J Fluid Mech. 213 (1990), 203–25.
12. JR. Houston, *Combined refraction and diffraction of short waves using the dual-reciprocity boundary-element method*, Appl Ocean Res.3 (1981), 163–70.
13. TK. Tsay, PLF. Liu, *A finite element model for wave refraction, diffraction, reflection and dissipation*, Appl Ocean Res.11 (1989), 33–8.
14. CC. Mei, HS. Chen, *A hybrid element method for steady linearized free-surface flows*, Int J Numer Methods Eng. 10 (1976), 1153–75.
15. G. Bellotti, R. Briganti, GM. Beltrami, L. Franco, *Modal analysis of semi-enclosed basins*, Coast Eng. 64 (2012),16–25.
16. TH. Jung, S. Son, Y. Ryu, *Finite element solution of linear Waves on a sloping bottom boundary*. J Coast Res. 33 (2016), 731–37.
17. G. Wang, JH. Zheng, JPY. Maa, JS. Zhang, AF. Tao, *Numerical experiments on transverse oscillations induced by normal-incident waves in a rectangular harbor of constant slope*, Ocean Eng. 57 (2013), 1–10.

Similarity Solution of Hydromagnetic Flow Near Stagnation Point Over a Stretching Surface Subjected to Newtonian Heating and Convective Condition



KM Kanika, Santosh Chaudhary, and Mohan Kumar Choudhary

1 Introduction

In the presence of magnetic field, introducing current through an electrically conducting fluid represents the theory of magnetohydrodynamics (MHD). Study of MHD flow of an electrically conducting fluid has considerable interest in the applications of modern metallurgical and engineering processes. Particularly, such applications are involved in cooling of nuclear reactors, purifications of molten metals, MHD generator, power generators, gas turbines, and crystal growth. MHD effects of the boundary layer on a semi-infinite plate in the presence of uniform transverse magnetic field was firstly investigated by Rossow [1]. Furthermore, Chaudhary and Kumar [2], Singh et al. [3], Daniel and Daniel [4], Kiyasfar and Pourmahmoud [5], and Nayak [6] have done an analytical work on the electrically conducting fluid in the presence of magnetic field.

The stagnation point flow interprets the motion of fluid near the stagnation region, which exists for both cases of a fixed or a moving body towards fluid. This region conflicts the highest pressure, the highest rate of heat transfer, and the mass deposition. It is a fundamental topic in fluid dynamics, which attracted the attention of many researchers. Stagnation point flow has wide applications in industrial and technical areas, such as solar central receivers exposed to wind currents, hydrodynamic processes, cooling of electronic devices by fans, and heat exchangers placed in a low-velocity environment. Some earlier pub-

KM Kanika (✉) · S. Chaudhary · M. K. Choudhary
Department of Mathematics, Malaviya National Institute of Technology, Jaipur 302017, India
e-mail: kanikatomar94@gmail.com; d11.santosh@yahoo.com; mkc2212@gmail.com

© Springer Nature Switzerland AG 2019
V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41,
https://doi.org/10.1007/978-3-030-02487-1_29

457

lished works of distinguished researchers are mentioned in the studies by Jat and Chaudhary [7], Mabood and Khan [8], Borrelli et al. [9], and Chaudhary and Choudhary [10].

Stretching surface has been an immense application in industrial and engineering processes, including annealing and tinning of copper wire, crystal growing, extraction of polymer sheet, glass fiber production, drawing of plastic films, manufacture of food, thermal energy storage, and many of others. Crane [11] was the first to study the flow of incompressible fluid over a linearly stretching sheet. Some other studies have been carried out by Mohamed et al. [12], Chaudhary et al. [13], Chaudhary and Choudhary [14], Daniel [15], and Hsiao [16] with a stretching surface under the different conditions.

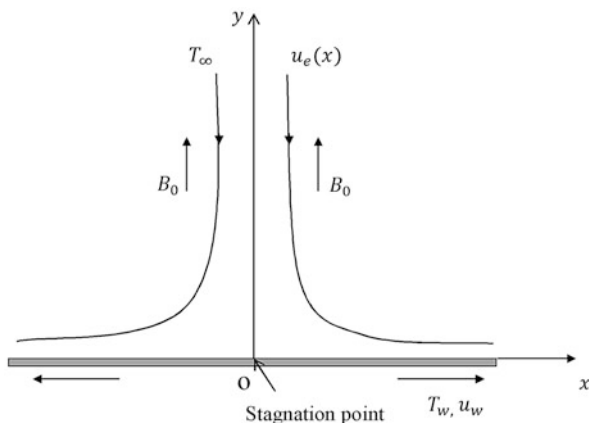
Newtonian heating (NH) is a way of heat exchange between boundary surface and ambient fluid, where the rate of heat exchange with a finite heat capacity is proportional to the temperature of local surface. Application of NH can be found in many engineering devices like cooling mechanism for nuclear reactors, solar power collectors, and heat exchangers. Many researchers like Merkin [17], Salleh et al. [18], and Akbar and Khan [19] have introduced the problem of NH in different ways under the distinct physical effects. Moreover, convective boundary condition (CBC) plays an important role in diverse technologies and industrial operations such as material drying, laser pulse heating, transpiration cooling process, and textile drying. Zhang and Zheng [20], Srinivasacharya and Bindu [21], and Shahzadi and Nadeem [22] presented an excellent review of boundary layer flow with CBC and some related applications.

Motivated by the previous mentioned research, main objective of the present study is to extend the research of Mohamed et al. [12] with the consideration of an electrically conducting fluid in the presence of uniform transverse magnetic field. Problem is solved numerically with the help of perturbation technique.

2 Formulation of the Problem

Consider a steady laminar two-dimensional stagnation point flow of an electrically conducting fluid past a stretching sheet with constant temperature T_w . Rectangular coordinates (x, y) are surmised in a manner that x -axis is measured along the stretching wall and y -axis is taken normal to the wall as shown in Figure 1. Uniform magnetic field with strength B_0 is applied normal to the stretching wall and induced magnetic field comes out to be negligible if the Reynolds number is very small. Moreover, it is also assumed that the external velocity $u_e(x) = ax$ and the stretching velocity $u_w(x) = bx$, where a and b are positive constants. The ambient fluid for above the sheet is kept at a constant temperature T_∞ . Under the above assumptions and neglecting viscous dissipation, the governing boundary layer equations are

Fig. 1 Flow configuration and geometrical coordinates



$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0 \tag{1}$$

$$u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} = u_e \frac{du_e}{dx} + v \frac{\partial^2 u}{\partial y^2} - \frac{\sigma_e B_0^2}{\rho} (u - u_e) \tag{2}$$

$$u \frac{\partial T}{\partial x} + v \frac{\partial T}{\partial y} = \alpha \frac{\partial^2 T}{\partial y^2} \tag{3}$$

with the related boundary conditions:

$$\begin{aligned}
 y = 0 : \quad & u = u_w(x), \quad v = 0, \quad \frac{\partial T}{\partial y} = -h_w T \text{ (NH)}, \quad -\kappa \frac{\partial T}{\partial y} = h_w (T_w - T) \text{ (CBC)} \\
 y \rightarrow \infty : \quad & u \rightarrow u_e(x), \quad T \rightarrow T_\infty
 \end{aligned} \tag{4}$$

where u and v are the velocity components along the x - and y -axes, respectively. $\nu = \frac{\mu}{\rho}$ is the kinematic viscosity, μ is the coefficient of viscosity, ρ is the fluid density, σ_e is the electrical conductivity, T is temperature of the fluid, α is the thermal diffusivity, h_w is the heat transfer coefficient, and κ is the thermal conductivity.

Following transformation variables are introduced (Mohamed et al. [12]):

$$\psi(x, y) = \sqrt{\nu u_e x} f(\eta) \tag{5}$$

$$\eta = \sqrt{\frac{u_e}{\nu x}} y \tag{6}$$

$$\theta(\eta) = \frac{T - T_\infty}{T_\infty} \text{ (NH)}, \quad \theta(\eta) = \frac{T - T_\infty}{T_w - T_\infty} \text{ (CBC)} \tag{7}$$

where $\psi(x, y)$ is the stream function, by the definition of stream function $u = \frac{\partial\psi}{\partial y}$ and $v = -\frac{\partial\psi}{\partial x}$, which identically satisfies the continuity equation (1), $f(\eta)$ is the dimensionless stream function, η is the similarity variable, and $\theta(\eta)$ is the dimensionless temperature. The momentum and energy equations (2) and (3) with the boundary conditions (4) are reduced into the following nonlinear differential equations:

$$f''' + f f'' - f'^2 - M(f' - 1) + 1 = 0 \tag{8}$$

$$\theta'' + Pr f \theta' = 0 \tag{9}$$

subject to the boundary conditions:

$$\begin{aligned} \eta = 0 : \quad f = 0, \quad f' = \varepsilon, \quad \theta' = -\gamma(1 + \theta) \quad (\text{NH}), \quad \theta' = -\gamma(1 - \theta) \quad (\text{CBC}) \\ \eta \rightarrow \infty : \quad f' \rightarrow 1, \quad \theta \rightarrow 0 \end{aligned} \tag{10}$$

where primes denote differentiation with respect to η , $M = \frac{\sigma_e B_0^2}{\rho a}$ is the magnetic parameter, $Pr = \frac{\nu}{\alpha}$ is the Prandtl number, $\varepsilon = \frac{b}{a}$ is the stretching parameter, and $\gamma = \sqrt{\frac{\nu}{a}} h_w$ (NH) or $\gamma = \sqrt{\frac{\nu}{a}} \frac{h_w}{\kappa}$ (CBC) is the conjugate parameter.

Physical quantities of interest are the local skin-friction coefficient C_f and the local Nusselt number Nu_x , defined as follows:

$$C_f = \frac{\tau_w}{\frac{\rho u_e^2}{2}} \tag{11}$$

$$Nu_x = \frac{x q_w}{\kappa (T_w - T_\infty)} \tag{12}$$

where $\tau_w = \mu \left(\frac{\partial u}{\partial y} \right)_{y=0}$ is the surface shear stress and $q_w = -\kappa \left(\frac{\partial T}{\partial y} \right)_{y=0}$ is the surface heat flux.

Therefore, after using the similarity transformations (5) to (7), the equations (11) and (12) can be defined in the following form:

$$C_f = \frac{2}{\sqrt{Re_x}} f''(0) \tag{13}$$

$$Nu_x = -\frac{\sqrt{Re_x}}{\theta(0)} \theta'(0) \quad (\text{NH}), \quad Nu_x = -\sqrt{Re_x} \theta'(0) \quad (\text{CBC}) \tag{14}$$

where $Re_x = \frac{u_e x}{\nu}$ is the local Reynolds number.

3 Computational Technique

Perturbation technique is employed for numerical solution of equations (8) and (9) along with boundary conditions (10). Applying the power series in a term of small magnetic parameter M as:

$$f(\eta) = \sum_{i=0}^{\infty} (M)^i f_i(\eta) \tag{15}$$

$$\theta(\eta) = \sum_{j=0}^{\infty} (M)^j \theta_j(\eta) \tag{16}$$

Substituting equations (15) and (16) and its derivative in the equations (8) to (10) and then equating the coefficients of like power of M :

$$f_0''' + f_0 f_0'' - f_0'^2 = -1 \tag{17}$$

$$\theta_0'' + Pr f_0 \theta_0' = 0 \tag{18}$$

$$f_1''' + f_0 f_1'' - 2f_0' f_1' + f_0'' f_1 = f_0' - 1 \tag{19}$$

$$\theta_1'' + Pr f_0 \theta_1' = -Pr f_1 \theta_0' \tag{20}$$

$$f_2''' + f_0 f_2'' - 2f_0' f_2' + f_0'' f_2 = -f_1 f_1'' + f_1'^2 + f_1' \tag{21}$$

$$\theta_2'' + Pr f_0 \theta_2' = -Pr (f_1 \theta_1' + f_2 \theta_0') \tag{22}$$

with the boundary conditions:

$$\begin{aligned} \eta = 0 : \quad & f_i = 0, \quad f_0' = \varepsilon, \quad f_j' = 0, \quad \theta_0' = -\gamma(1 + \theta_0) \quad (\text{NH}), \quad \theta_j' = -\gamma\theta_j \quad (\text{NH}) \\ & \theta_0' = -\gamma(1 - \theta_0) \quad (\text{CBC}), \quad \theta_j' = \gamma\theta_j \quad (\text{CBC}) \\ \eta \rightarrow \infty : \quad & f_0' \rightarrow 1, \quad f_j' \rightarrow 0, \quad \theta_i \rightarrow 0, \quad i \geq 0, \quad j > 0 \end{aligned} \tag{23}$$

Equations (17) and (18) for the nonmagnetic case was obtained by Mohamed et al. [12], and the remaining ordinary differential equations are solved numerically by using shooting technique with fourth order Runge Kutta method. Step size is taken 0.001, and the above process is repeated until the results correct up to six places of decimal.

4 Validation

To validate the numerical method, the results of the local skin friction coefficient $f''(0)$ for different values of the stretching parameter ϵ are compared with the earlier results of Mohamed et al. [12] in the absence of magnetic parameter M , which are presented in Table 1. In this table, it can be claimed that the numerical results obtained by using the perturbation technique are very close to the previous published results. Finally, it can also be concluded that the demonstrated results are reliable and effective.

5 Discussion of the Results

Influence of various physical parameters such as stretching parameter ϵ , magnetic parameter M , conjugate parameter γ , and Prandtl number Pr on velocity profile $f'(\eta)$ and temperature profile $\theta(\eta)$ are discussed through graphs in this section. Furthermore, the computational results of shear stress $f''(0)$ and heat transfer $\theta'(0)$ regarding the above mentioned physical parameters are shown in tabular form.

The effects of the stretching parameter ϵ and the magnetic parameter M on the fluid flow $f'(\eta)$ are demonstrated in Figures 2 and 3 respectively, keeping other parameters constant. From these figures, it is analyzed that the velocity increases with an increment in the stretching parameter ϵ and the magnetic parameter M .

Figures 4, 5, 6, 7, 8, 9, 10 and 11 illustrate the impact of stretching parameter ϵ , magnetic parameter M , conjugate parameter γ , and Prandtl number Pr on temperature distribution $\theta(\eta)$ for the NH and CBC cases, respectively, while the other parameters are constant. It is clear from Figures 4 and 5 that the dimensionless temperature is decreased with the increasing value of stretching parameter ϵ but the opposite phenomenon occurs in the NH case for $\eta > 2$. For the higher value of stretching parameter, the retarding force decreases in the motion of flow. So, fluid velocity increases and temperature decreases due to incitement in the stretching parameter. Figures 6 and 7 describe that the temperature reduces for NH case by increasing the magnetic parameter M but opposite is true for $\eta > 1.5$. Whereas, for the case of CBC reverse behavior is observed as compared to NH case. This is due to the reason that increase in the magnetic parameter tends to the enhancement of Lorentz force leading to the development of resistance to fluid flow which in turn

Table 1 Comparison of the results for $f''(0)$ with the earlier published results when $M = 0.0$

ϵ	$f''(0)$	
	Mohamed et al. [12]	Present results
0.0	1.2325877	1.232588
0.5	0.7132949	0.713295
1.0	0.0000000	0.000000
2.0	-1.8873066	-1.887304

generates more heat resulting in increment of fluid temperature. From Figures 8 and 9, it can be seen that the temperature of a flow field is an increasing function of the conjugate parameter γ . This phenomenon occurs because heat transfer coefficient is equivalent to the thermal resistance on warm fluid side. So, warm fluid side convection reduces and surface temperature rises along booming quantity in conjugate parameter. On the other hand, the temperature decreases with increasing the Prandtl number Pr and reverse behavior is noted in the NH case for $\eta > 1.5$, and these observations are presented by Figures 10 and 11. As Prandtl number is inversely proportional to the thermal conductivity, the increasing value of Prandtl

Fig. 2 Effects of ε on velocity profiles with $M=0.1$

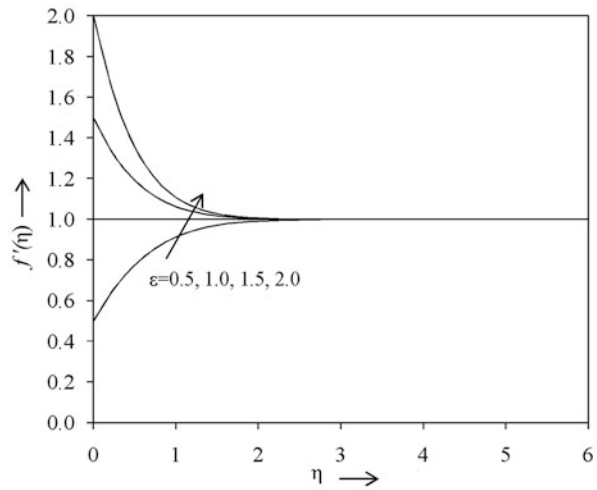


Fig. 3 Effects of M on velocity profiles with $\varepsilon = 0.1$

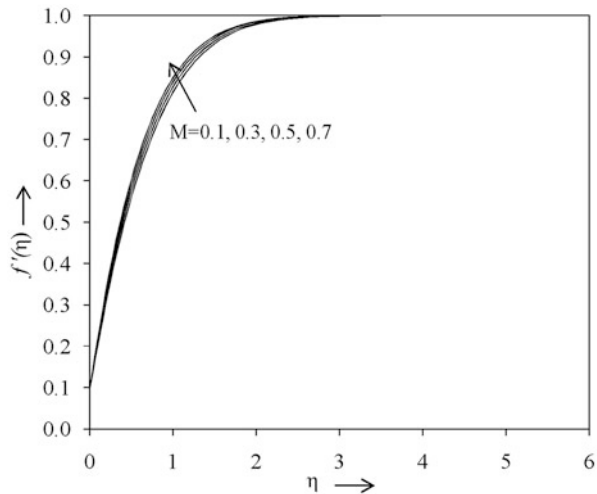


Fig. 4 Effects of ε on temperature profiles for NH case with $M = 0.1$, $\gamma = 1.0$, and $Pr = 0.72$

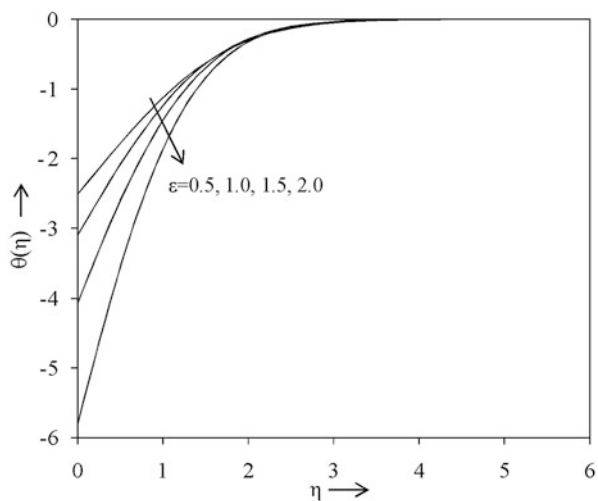


Fig. 5 Effects of ε on temperature profiles for CBC case with $M = 0.1$, $\gamma = 1.0$, and $Pr = 0.72$

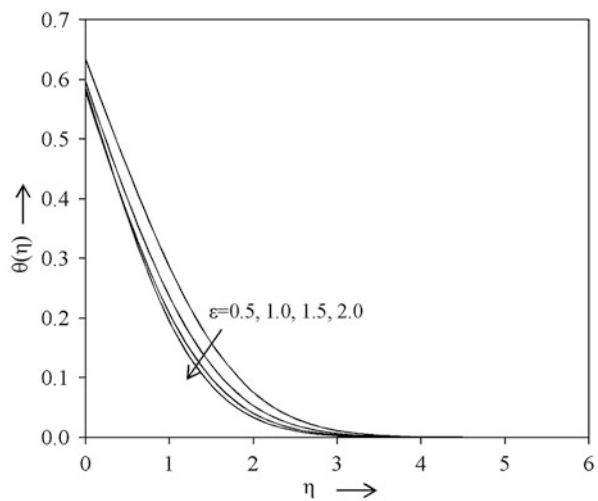


Fig. 6 Effects of M on temperature profiles for NH case with $\varepsilon = 0.1$, $\gamma = 1.0$, and $Pr = 0.72$

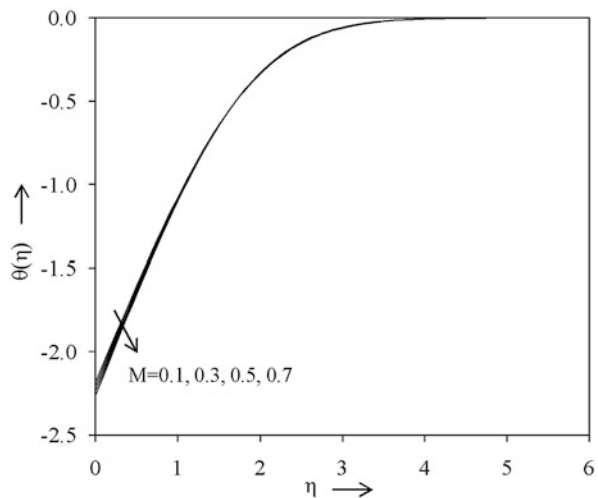


Fig. 7 Effects of M on temperature profiles for CBC case with $\varepsilon = 0.1$, $\gamma = 1.0$, and $Pr = 0.72$

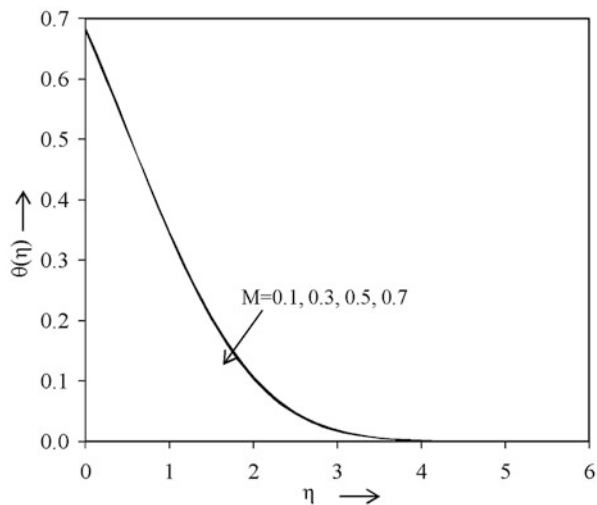


Fig. 8 Effects of γ on temperature profiles for NH case with $\varepsilon = 0.1$, $M = 0.1$, and $Pr = 0.72$

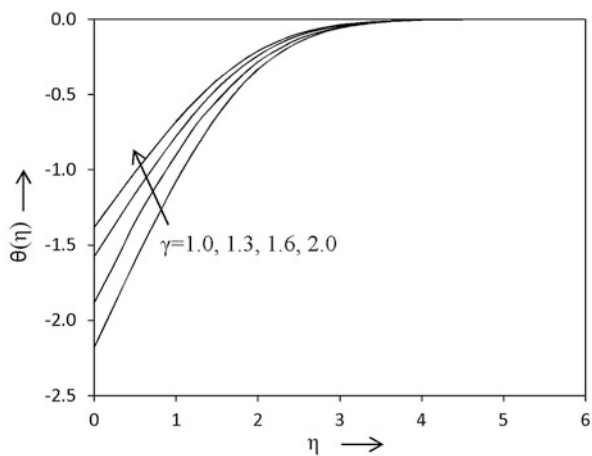


Fig. 9 Effects of γ on temperature profiles for CBC case with $\varepsilon = 0.1$, $M = 0.1$, and $Pr = 0.72$

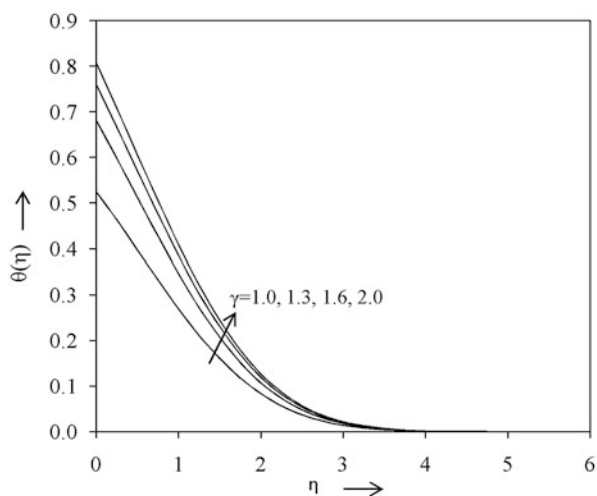


Fig. 10 Effects of Pr on temperature profiles for NH case with $\varepsilon = 0.1$, $M = 0.1$, and $\gamma = 1.0$

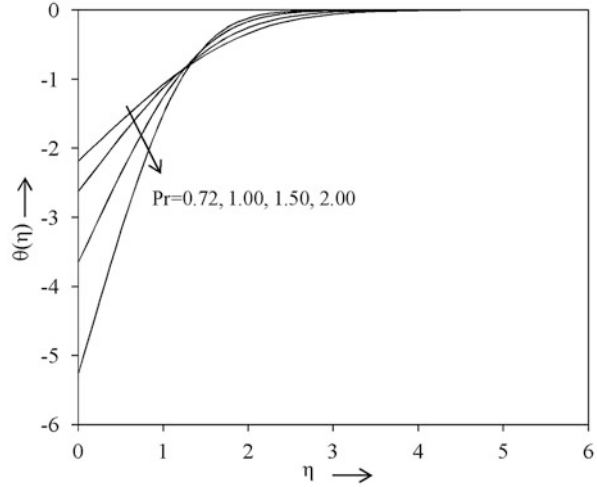
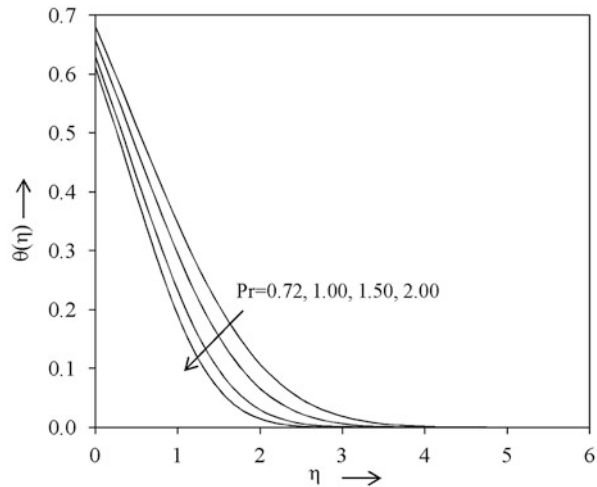


Fig. 11 Effects of Pr on temperature profiles for CBC case with $\varepsilon = 0.1$, $M = 0.1$, and $\gamma = 1.0$



number reduces diffusion of energy and hence temperature of the fluid decreases strongly.

Finally, the effects of stretching parameter ε , magnetic parameter M , conjugate parameter γ , and Prandtl number Pr on the wall shear stress $f''(0)$ and heat transfer rate $\theta'(0)$ for all cases are presented in the form of numerical data in Table 2. As well the local skin friction C_f and the local Nusselt number Nu_x are proportional to $f''(0)$ and $\theta'(0)$, respectively. From this table, it is observed that the wall shear stress $f''(0)$ decreases with the increasing value of the stretching parameter ε but reverse is true for the magnetic parameter M . Moreover, the positive values of shear stress for all values of the physical parameters are denotative of the fact that fluid utilizes a drag force on the surface and negative shear stress implies that fluid exerts a drag

Table 2 Values of $f''(0)$ and $\theta'(0)$ with different values of physical parameters

ε	M	γ	Pr	$f''(0)$	$\theta'(0)$ for NH	$-\theta'(0)$ for CBC
0.5	0.1	1.0	0.72	0.730376	1.517189	0.3753258
1.0				0.000000	2.105341	0.4173960
1.5				-0.887893	3.065240	0.4289890
2.0				-1.913175	4.805400	0.4250723
0.1	0.1			1.181020	1.187295	0.3284310
	0.3			1.247159	1.305741	0.3280625
	0.5			1.310040	1.332417	0.3252443
	0.7			1.370097	1.369421	0.3232542
	0.1	1.3			0.925670	0.3767543
		1.6			0.872321	0.3802064
		2.0			0.770342	0.3948707
		1.0	1.00		1.631282	0.3786723
			1.50		2.658976	0.3813452
			2.00		4.272536	0.3976832

force by the surface. It is also evident that rising values of stretching parameter ε , magnetic parameter M , and Prandtl number Pr leads to the increment in the heat transfer rate $\theta'(0)$ for all cases, while opposite phenomenon arises in the CBC case for stretching parameter ε , when $\varepsilon < 1.5$, and Prandtl number Pr . Further, the local Nusselt number is reduced with the increment of the values of conjugate parameter γ for all cases. From the practical point of view, positive sign of the heat transfer rate implies that there is a heat flow from the wall and vice versa.

6 Conclusions

A numerical model is developed to examine the viscous, incompressible, laminar stagnation point flow of electrically conducting fluid towards a stretching sheet with NH and CBC. Similarity transformations are used to convert the system of partial differential equations into the set of nonlinear ordinary differential equations with the associated boundary conditions. Transformed equations are solved numerically by using the perturbation technique. From the results of the problem, it was concluded that an accretion in the value of stretching parameter increases the thickness of momentum boundary layer and the heat transfer rate for all analyzed cases but opposite is true for heat transfer rate in CBC case for $\varepsilon < 1.5$. Subsequently, thermal boundary layer thickness and shear stress decrease with the rising values of stretching parameter for all cases, while opposite phenomenon occurs in thermal boundary layer for the case of NH when $\eta > 2$. For all considered cases, fluid velocity, skin friction, and Nusselt number rise with the increasing values of magnetic parameter, and consequently, temperature of fluid is reduced

for NH case but raised for $\eta > 1.5$. Moreover, the reverse trend is observed in the CBC case, as compared to the NH case. An increment in the conjugate parameter increases the thermal boundary layer thickness and decreases the surface heat flux for all considered cases. Further, increasing values of the Prandtl number reduces thermal boundary layer thickness, whereas reverse behavior occurs in the NH case for $\eta > 1.5$. Finally, the surface heat flux is increased in case of NH and decreased in case of CBC.

References

1. Rossow, V.J.: On flow of electrically conducting fluids over a flat plate in the presence of a transverse magnetic field. *NACA TN* 1, 489–508 (1957)
2. Chaudhary, S., Kumar, P.: MHD forced convection boundary layer flow with a flat plate and porous substrate. *Meccanica* 49, 69–77 (2014)
3. Singh, H., Ram, P., Kumar, V.: Unsteady MHD free convection past an impulsively started isothermal vertical plate with radiation and viscous dissipation. *Fluid Dynamics & Material Processing* 10, 1–30 (2014)
4. Daniel, Y.S., Daniel, S.K.: Effects of buoyancy and thermal radiation on MHD flow over a stretching porous sheet using homotopy analysis method. *Alexandria Engineering Journal* 54, 705–12 (2015)
5. Kiyasatfar, M., Pourmahmoud, N.: Laminar MHD flow and heat transfer of power-law fluids in square microchannels. *International Journal of Thermal Sciences* 99, 26–35 (2016)
6. Nayak, M.K.: MHD 3D flow and heat transfer analysis of nanofluid by shrinking surface inspired by thermal radiation and viscous dissipation. *International Journal of Mechanical Sciences* 124–125, 185–193 (2017)
7. Jat, R.N., Chaudhary, S.: Radiation effects on the MHD flow near the stagnation point of a stretching sheet. *Z Angew Math Phys* 61, 1151–1154 (2010)
8. Mabood, F., Khan, W.A.: Approximate analytic solutions for influence of heat transfer on MHD stagnation point flow in porous medium. *Computers & Fluids* 100, 72–78 (2014)
9. Borrelli, A., Giantesio, G., Patria, M.C.: An exact solution for the 3D MHD stagnation-point flow of a micropolar fluid. *Communications in Nonlinear Science and Numerical Simulation* 20, 121–135 (2015)
10. Chaudhary, S., Choudhary, M.K.: Heat and mass transfer by MHD flow near the stagnation point over a stretching or shrinking sheet in a porous medium. *Indian Journal of Pure & Applied Physics* 54, 209–217 (2016)
11. Crane, L.J.: Flow past a stretching plate. *Z Angew Math Phys* 21, 645–647 (1970)
12. Mohamed, M.K.A., Salleh, M.Z., Nazar, R., Ishak, A.: Numerical investigation of stagnation point flow over a stretching sheet with convective boundary conditions. *Boundary Value Problems* 2013, 4 (2013)
13. Chaudhary, S., Choudhary, M.K., Sharma, R.: Effects of thermal radiation on hydromagnetic flow over an unsteady stretching sheet embedded in a porous medium in the presence of heat source or sink. *Meccanica* 50, 1977–1987 (2015)
14. Chaudhary, S., Choudhary, M.K.: Partial slip and thermal radiation effects on hydromagnetic flow over an exponentially stretching surface with suction or blowing. *Thermal Science* <https://doi.org/10.2298/TSCI160127150C>, (2016)
15. Daniel, Y.S.: MHD laminar flows and heat transfer adjacent to permeable stretching sheets with partial slip condition. *Journal of Advanced Mechanical Engineering* 4, 1–5 (2017)

16. Hsiao, K.L.: Micropolar nanofluid flow with MHD and viscous dissipation effects towards a stretching sheet with multimedia feature. *International Journal of Heat and Mass Transfer* 112, 983–990 (2017)
17. Merkin, J.H.: Natural-convection boundary-layer flow on a vertical surface with Newtonian heating. *International Journal of Heat and Fluid Flow* 15, 392–398 (1994)
18. Salleh, M.Z., Nazar, R., Pop, I.: Boundary layer flow and heat transfer over a stretching sheet with Newtonian heating. *Journal of Taiwan Institute of Chemical Engineers* 41, 651–655 (2010)
19. Akbar, N.S., Khan, Z.H.: Influence of magnetic field for metachronal beating of cilia for nanofluid with Newtonian heating. *Journal of Magnetism and Magnetic Materials* 381, 235–242 (2015)
20. Zhang, Y., Zheng, L.: Similarity solutions of Marangoni convection boundary layer flow with gravity and external pressure. *Chinese Journal of Chemical Engineering* 22, 365–369 (2014)
21. Srinivasacharya, D., Bindu, K.H.: Entropy generation in a micropolar fluid flow through an inclined channel with slip and convective boundary conditions. *Energy* 91, 72–83 (2015)
22. Shahzadi, I., Nadeem, S.: Inclined magnetic field analysis for metallic nanoparticles submerged in blood with convective boundary condition. *Journal of Molecular Liquids* 230, 61–73 (2017)

Modelling Corrosion Phenomenon of Magnesium Alloy AZ91 in Simulated Body Fluids



Ramalingam Vaira Vignesh and Ramasamy Padmanaban

1 Introduction

Biodegradable bioimplants is a major thrust area in the field of orthopedic implants. The biodegradable bioimplant materials are characterized based on the elastic modulus, biocompatibility, and biodegradability of the materials. The elastic modulus of Magnesium (Mg) and Mg alloys is similar to that of bone, which diminishes the stress shielding effect. Mg is one of the key elements found in the physiological fluids and responsible for functions such as osteoconduction, excitability of nerves and bones, etc., and hence it is biocompatible. Mg and its alloys are highly susceptible to corrosion in chloride containing physiological environments [1], which attests the biodegradable nature of these alloys. The biodegradable property is advantageous as it eliminates a secondary surgery to remove the temporary implants. The amalgamation of mechanical properties, biocompatibility, and biodegradability makes Mg and its alloys to be one of the best biodegradable biomaterials for bone implants.

One of the widely used biodegradable Mg alloy is AZ91. The microstructure of AZ91 alloy consists of primary phase (α -Mg) and secondary phase (β -Mg₁₇Al₁₂) in the matrix [2] as shown in Figure 1. The mechanical, tribological, and corrosion properties of the alloy are dependent on the size, distribution, and dispersion of the secondary phase in the alloy matrix [3–10]. The high activity of Mg and its alloys accelerates the corrosion in physiological environments. The potential difference between the primary and secondary phases induces intergranular and micro galvanic

R. Vaira Vignesh · R. Padmanaban (✉)

Department of Mechanical Engineering, Amrita School of Engineering, Coimbatore, India

Amrita Vishwa Vidyapeetham, Amrita University, Coimbatore, India

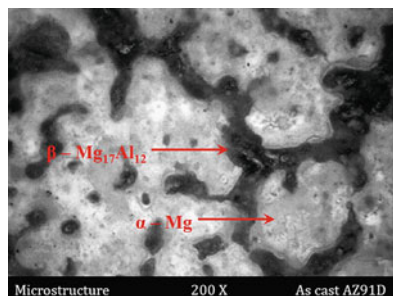
e-mail: dr_padmanaban@cb.amrita.edu

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41, https://doi.org/10.1007/978-3-030-02487-1_30

471

Fig. 1 Microstructure of Mg alloy AZ91



corrosion [11]. Hence the corrosion of the alloy could be controlled by altering the microstructure, which influences the intergranular and micro galvanic corrosion.

Alloying, compositing, and coating are some of the notable ways to engineer the corrosion rate of AZ91. Anik et al. [11] found that the eutectic phase of AZ91 acted as anodic barriers in the absence of Cl^- ions and pH of 7. Wang et al. [12] found that passive current density increased and the passive film became unstable in AZ91 alloy with addition of Ce to the matrix. This was attributed to the formation of CeO_2 , which increased the donor concentration in the passive film. Hwang et al. [13] reported the formation of MgO and Mn_2O_3 oxide layer during plasma electrolytic oxidation of AZ91 alloy using $\text{KOH} + \text{KF} + \text{Na}_2\text{SiO}_3$ solution in the presence of KMnO_4 , a strong oxidizing agent. The Mn_2O_3 and MgO layers acted as potential diffusion barriers and hence enhanced the corrosion resistance of the alloy. Luo et al. [14] reported that 0.3 wt.% of Y was the optimum quantity for increasing the corrosion resistance of AZ91 alloy. The formation of new intermetallic phase MgAl_4Y reduced the amount and continuity of β phase, which reduced the galvanic contact between β phase and α phase. This leads to an increase in the corrosion resistance of the alloy.

Zhan et al. [15] found that addition of alloying elements Zn, Sn, and In to twin rolled continuous cast AZ91 resulted in the formation of Mg-In, Mg-Al, and Mg-Sn intermetallics. The formation of globular α phase increased the tensile strength, while intermetallics enhanced micro galvanic corrosion. Ko et al. [16] found that Zr incorporation along with annealing process enhanced the corrosion resistance of AZ91 alloy in NaCl solution. Annealing formed MgO on the surface which reduced the surface activity. Ghayad et al. [17] investigated the corrosion properties of AZ91 and ZM60 with graded composition of Ca, Sr, Misch Metal (MM) and Cu, MM, respectively, in 3.5 wt.% of NaCl solution. They found that addition of less than 2% Cu to ZM60 alloy enhances corrosion resistance. The addition of 0.6% Ca to the AZ91 matrix increased the polarization resistance. The formation of Al_4MM and β phase precipitation in AZ91–0.4Ca–0.14Sr–1.2MM improved the corrosion resistance.

The degradation kinetics of implants during implantation period explicates the biocompatible nature of the implant material. Witte et al. [18] reported that the absence of subcutaneous gas formation adjacent to implants during in-vivo studies on AZ91 attested its biocompatibility and corrosion resistance. Choudhary et al.

[19] observed transgranular cracks and reported that AZ91D alloy is susceptible to stress corrosion cracking in simulated body fluid (SBF) solution in the low strain rate range. Xue et al. [20] found that the in-vivo corrosion rate of AZ91 alloy was lower than the in-vitro corrosion rate. Walter et al. [21] reported low polarization resistance and high localized degradation in SBF for AZ91D with rough surface. The severity of localized degradation of the rough surface specimen was higher than that of the smooth polished specimen. Tahmasebifar et al. [22] reported that AZ91 alloy plates with high porosity and rough texture increased the cell adhesion and proliferation. Wen et al. [23] investigated the biodegradability and surface chemistry of AZ31D and AZ91. They found that the corrosion rate of AZ91 decreased with increase in the immersion period.

All the experimental works described so far suggested that the corrosion rate of AZ91 is controlled by grain size, distribution, and dispersion of β phase in the matrix. Modelling of corrosion phenomenon could elaborate the influence of AZ91's microstructure on corrosion rate. Jia et al. [24] studied the influence of geometric parameters on the galvanic current distribution in Mg alloy AZ91D coupled to steel using a boundary element method. Deshpande et al. [25] validated the numerical model for galvanic corrosion of Mg alloy AE44 coupled with aluminium alloy AA6061. Grogan et al. [26] developed a numerical model to predict the effect of corrosion on the mechanical integrity of bioabsorbable metallic stents. Bakhsheshi-Rad et al. [27] used gene expression program to model corrosion behavior of biodegradable Mg alloy Mg-Zn-RE-Ca and Mg-Zn-RE and observed that Mg-Zn-RE-Ca had lower corrosion current density than Mg-Zn-RE alloy. Most of the works reported in literature are restricted to stationary cathode and anode surfaces.

In this study, Arbitrary Lagrangian Eulerian (ALE) formulation was used to effectively trace the interfaces (electrolyte and material surface). The corrosion behavior of Mg alloy AZ91 with continuous and dispersed secondary phase particles was modelled using Comsol Multiphysics[®] software. The polarization data required for building the model was obtained from the literature [1]. The corrosion current density of the alloy with continuous and dispersed secondary phase particles was evaluated. The dependence of anodic current density on the structure and dispersion of secondary phase of the alloy was also investigated.

1.1 Nomenclature

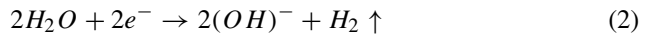
N_i	Flux
D_i	Diffusion coefficient
c_i	Concentration
z_i	Charge
F	Faraday's constant
u_i	Mobility
ϕ	Potential
U	Solvent velocity

$f_a(\phi)$	Current density of anodic species
$f_c(\phi)$	Current density of cathodic species
(X, Y)	Reference frame co-ordinates
(x, y)	Spatial frame co-ordinates
M	Atomic mass
Z	Electron number
ϕ_L	Level set function
i_l	Electrolyte current density vector
σ_l	Electrolyte conductivity
n	Normal vector pointing out of the domain
$i_{loc,m}$	Local individual electrode reaction current density
M_i	Molar mass of the corroding species
ρ_i	Density of the corroding species
$u_{dep,i,m}$	Stoichiometric coefficient
n_m	Number of electrons participating in the electrode reaction
$micro(x, y)$	Microstructure function

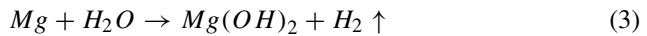
2 Computational Methodology

2.1 Corrosion of Magnesium Alloy AZ91 in SBF

The high activity of magnesium increases its degradation in corrosive environments. The corrosion rate is influenced by the presence of chloride ions. Owing to the biocompatible nature of magnesium alloy AZ91, it is used as bioimplants. The human physiological environment has chloride ions and temperature of 37 °C, which facilitates the corrosion of AZ91 alloy. The anodic and cathodic reaction of Mg alloy during corrosion is given by Equations (1) and (2), respectively.



Being highly active, magnesium readily loses its electron and becomes positive ion in the presence of an electrolyte medium (SBF). The cathodic reaction is the reduction of water molecule forming hydroxyl ions and hydrogen gas. The overall reaction is given by Equation (3), which indicates the formation of magnesium hydroxide layer as a corrosion product. In the presence of chloride ions, magnesium hydroxide will transform to magnesium chloride as given by Equation (4).



Continuous and rapid corrosion of magnesium alloy AZ91 results in the accumulation of H_2 gas in the neighboring tissues, which is a lethal phenomenon. So the corrosion rate of the Mg alloy should be engineered to enable its use as biodegradable bioimplant. One of the engineering techniques to reduce the corrosion rate of the alloy is disintegration and dispersion of β phase in the alloy. This reduces the potential gradient between α phase and β phase of the alloy, which also reduces the intergranular corrosion.

2.2 Governing Equations

During corrosion process, local ion concentration change affects the mass transport and distribution of various ion species. Measuring these local changes using conventional methods of measurements is challenging and difficult. Numerical methods help us to find critical parameters affecting local changes and understanding corrosion at the local level. One of the widely utilized numerical methods to investigate the transport phenomenon is finite element method (FEM). FEM is developed to solve the governing differential equation and predict the concentration, potential and current distribution, which is used to study the controllable factors of corrosion phenomenon.

$$\frac{\partial c_i}{\partial t} = -\nabla \cdot N_i + R_i \quad (5)$$

$$N_i = -D_i \nabla c_i - z_i F u_i c_i \nabla \phi + c_i U \quad (6)$$

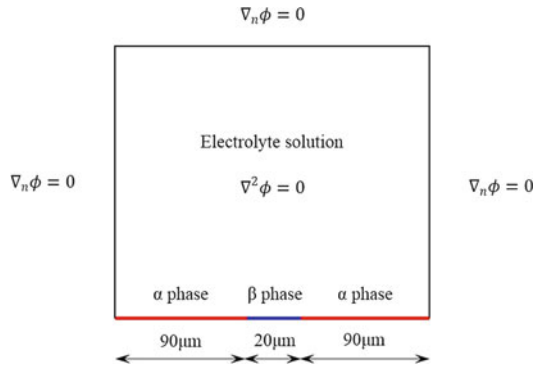
$$\frac{\partial c_i}{\partial t} = -\nabla \cdot N_i = D_i \nabla^2 c_i + z_i F u_i \nabla \cdot (c_i \nabla \phi) - \nabla \cdot (c_i U) \quad (7)$$

Nernst-Planck equation is used to describe the movement of the species in an ionic solution as given by Equation (5). The term on the left-hand side denotes the accumulation of species i . The first term ($-\nabla \cdot N_i$) denotes the transport of species i due to diffusion, migration, and convection and it is given in Equation (6). The second term (R_i) denotes a source or sink. The three additive fluxes, namely diffusion flux, migration flux, and convection flux were associated with the species flux. The conservation of species flux is given by Equation (7).

The following assumptions were made in the current study.

1. Electrolyte solution is incompressible.
2. Electrolyte solution is well mixed.
3. Absence of concentration gradient in the electrolyte.
4. Primary phase of the alloy acts as anode and secondary phase of the alloy acts as cathode.
5. Dissolution of primary phase is the anodic reaction and evolution of H_2 is the cathodic reaction.

Fig. 2 Layout of the developed model



From the above assumptions, the transport of species associated with diffusion and convection is neglected. The simplified form of Equation (7) is given by Equation (8), which takes the form of Laplace equation for the potential and represents the upper bound for the corrosion rate.

$$\nabla^2 \phi = 0 \tag{8}$$

Figure 2 shows the electrolyte domain over which Equation (3) is solved. As observed from the figure, β phase is sandwiched between α phases. In α phase, the coring effect results in variation of Al content from 1% in the center to 9% in the grain boundary. In this study, α phase is assumed to have a homogeneous composition of 3% Al and 97% Mg content. The boundary conditions at the surface of anode and cathode are given by Equations (9) and (10), respectively [28].

$$\nabla_n \phi = -\frac{f_a(\phi)}{\sigma} \tag{9}$$

$$\nabla_n \phi = -\frac{f_c(\phi)}{\sigma} \tag{10}$$

$f_a(\phi)$ and $f_c(\phi)$ are the linear interpolation functions for the polarization curves of α phase and β phase, respectively, with SBF solution as electrolyte. The data for interpolating the polarization curves were obtained from the literature [1]. The potential gradient was obtained by dividing the current density (corresponding to the potential at anodic and cathodic surface) by conductivity of the electrolyte solution. Equation (11) is the insulation boundary condition, which is applied to the boundaries of electrolyte as shown in Figure 2.

$$\nabla_n \phi = 0 \tag{11}$$

2.3 Arbitrary Lagrangian Eulerian Method

ALE is a moving mesh technique, which combines the Lagrangian and Eulerian frames of reference and capable of capturing large deformations with high accuracy. ALE method was incorporated using Comsol Multiphysics[®]. Equations (12) and (13) were used to solve the mesh displacement method, which resulted in smooth deformation of mesh subjected to the boundary constraints.

$$\frac{\partial^2}{\partial X^2} \left(\frac{\partial x}{\partial t} \right) + \frac{\partial^2}{\partial Y^2} \left(\frac{\partial y}{\partial t} \right) = 0 \quad (12)$$

$$\frac{\partial^2}{\partial X^2} \left(\frac{\partial y}{\partial t} \right) + \frac{\partial^2}{\partial Y^2} \left(\frac{\partial x}{\partial t} \right) = 0 \quad (13)$$

Faraday's law was used to calculate the normal velocity of the anode surface from the current density and the normal velocity was calculated using Equation (14). The normal velocity at the cathode surface was considered to be zero and all other boundaries were considered to have zero displacement.

$$\vec{n} \cdot \vec{v} = \frac{M}{ZF\rho} j = \frac{M}{ZF\rho} f_a(\phi) \quad (14)$$

2.4 Model Development

The corrosion phenomenon of the alloy was modelled using a two-dimensional model. *Corrosion, Secondary Interface* module in Comsol was used to solve for the electrolytic potential over the electrolyte domain. Equations (15) and (16) were used to solve for the electrolytic potential.

$$i_l = -\sigma_l \nabla \phi_l \quad (15)$$

$$\nabla \cdot i_l = 0 \quad (16)$$

External corroding electrode boundary condition was used at the electrode surface. The boundary condition for the electrolyte potential is given by Equation (17). The dissolution of the primary phase with a velocity in the normal direction is given by Equation (18). $R_{dep,i,m}$ was calculated using Equation (19).

$$\vec{n} \cdot i_l = \sum_m i_{loc,m} + i_{dl} \quad (17)$$

$$n \cdot \frac{\partial x}{\partial t} = \sum_i \frac{R_{dep,i,m} \times M_i}{\rho_i} \quad (18)$$

$$R_{dep,i,m} = - \frac{u_{dep,i,m} \times i_{loc,m}}{n_m \times F} \quad (19)$$

The electrode reaction in α phase on the surface of the alloy was modelled using user-defined electrode kinetics. The local current density for α phase and β phase is given by Equations (20) and (21), respectively. The expression $(1 - micro(x, y))$ in Equation (20) ensures that the local current density is applied only at α phase in the surface. Similarly, $(micro(x, y))$ in Equation (21) ensures that the local current density is applied only at β phase in the surface.

$$i_\alpha = f(\phi_{s,ext} - \phi_l) \times (1 - micro(x, y)) \quad (20)$$

$$i_\beta = f(\phi_{s,ext} - \phi_l) \times (micro(x, y)) \quad (21)$$

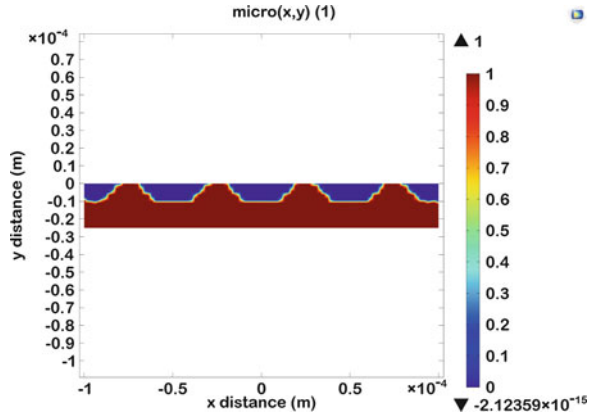
The experimental data for the polarization curves were obtained from the literature and a piecewise cubic interpolation function was developed for relating the corresponding local current density with the electrolyte potential. The electrolyte was meshed using free triangular mesh with coarse elements. The interface of electrolyte and the microstructural surface of the specimen were meshed using free triangular mesh with extremely fine elements.

3 Results and Discussion

In order to investigate the effect of microstructure and β phase distribution on the corrosion behavior of the alloy, two representative microstructure configurations were considered in the study. The first configuration of microstructure was represented with a continuous β phase network around α phase and the second configuration with a dispersed-discrete β phase.

The authentic effect of the phases on the corrosion behavior of the alloy could be analyzed, if the microstructural morphology along the entire depth of the alloy is considered. However, in 2D computation, the microstructural phases could be accounted only on the surface of the alloy. This is because the governing equations were solved over a computational domain representing the electrolyte solution and the microstructural phases were represented as the boundary of the computational domain.

Fig. 3 Representative image showing the continuous network of β phase surrounding α phase



3.1 Continuous Network of β Phase

Cathodic and anodic surface were represented based on the predefined microstructure, as described by Deshpande et al. [28].

The level set function (ϕ_L) is a function of two variables (x, y). The ϕ_L was used to represent the anodic (if $\phi_L > 0$) and cathodic surface (if $\phi_L \leq 0$). As shown in Figure 3, a portion of the alloy’s microstructure with dimensions of $200 \mu\text{m}$ length and $25 \mu\text{m}$ breadth was considered. The region above y distance = 0 represents the electrolyte solution. The maximum depth of α phase in the microstructure was considered as $10 \mu\text{m}$. In the representative microstructure shown in Figure 3, the blue region indicates α phase and the brown region indicates β phase.

As given in Equation (22), the boundary conditions are applied at the interface of electrolyte with microstructure, as both α phase and β phase are represented by a single boundary. Similarly, the normal velocity of the anodic surface in the ALE formulation is given by Equation (23). The average anodic current density was calculated using Equation (24), which was performed over the anodic surface.

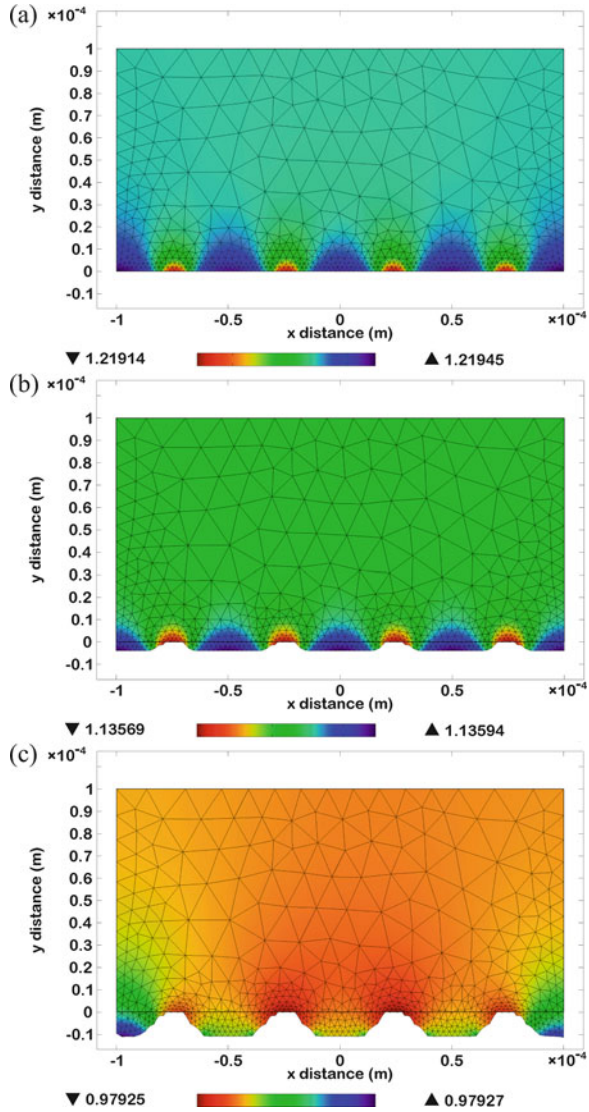
$$\nabla_n \phi = -\frac{f_a(\phi)}{\sigma} \times (\phi_L > 0) - \frac{f_c(\phi)}{\sigma} (\phi_L \leq 0) \tag{22}$$

$$\vec{n} \cdot \vec{v} = \frac{M}{ZF\rho} f_a(\phi) \times (\phi_L > 0) \tag{23}$$

$$\text{Average anodic current density} = \frac{\int_{\phi_L > 0} f_a(\phi) d\Omega}{\int_{\phi_L > 0} d\Omega} \tag{24}$$

The model predictions (electrolyte potential) at different intervals of time are shown in Figure 4. It is observed that the electrochemically active α phase is

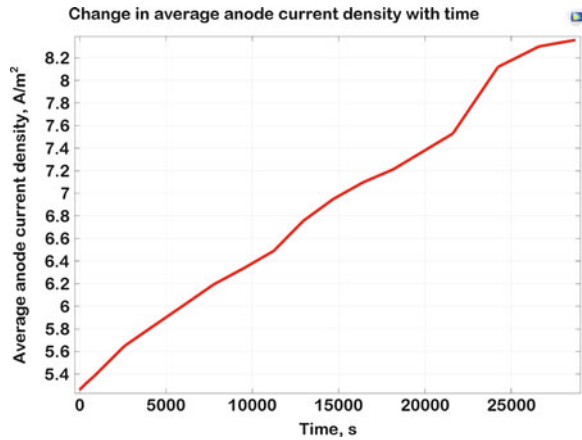
Fig. 4 Predicted electrolyte potential (V) at time (a) $T = 0$ s; (b) $T = 9553$ s; (c) $T = 28,314$ s



preferentially dissolved by the electrolyte, while the nobler β phase was intact in the microstructure. Hence the surface β phase fraction increased with progress of time. The computations were terminated just before the β phase fraction reached 100%.

As shown in Figure 4a, a high electrolyte potential was observed between the material and the electrolyte at the instance of immersion (time = 0 s). However the electrolyte potential decreased with progress of time, as evidenced from Figure 4b. The reduction in ratio of surface area of α phase to surface area of β phase reduced

Fig. 5 Average anodic current density vs Time for continuous network of β phase microstructure



the concentration gradient between these two phases. The electrolyte potential after complete dissolution of α phase in the electrolyte is shown in Figure 4c.

The average anodic current density is plotted against time as shown in Figure 5. Since α phase has more negative corrosive potential than β phase, the enrichment of β phase will accelerate the corrosion of the material. Hence the average anodic current density increased with increase in time and was varying between 5 A m^{-2} and 9 A m^{-2} . The β phase acted as barrier to corrosion process and therefore retarded the corrosion rate of the material. The corrosion result was found to be consistent with the experimental results available in the literature [23].

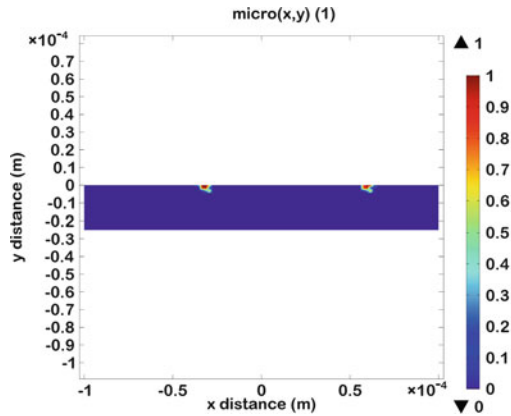
3.2 Discrete β Phase

The second microstructural configuration considered in the study had discrete and dispersed β phase in the matrix. Microstructural refinement techniques result in submicron and nano-sized β phase in the AZ91 alloy matrix [29–33]. In this study, β phase with irregular geometry and a depth of $2.5 \mu\text{m}$ was considered. The microstructure was depicted using the level set function and the representative microstructural image is shown in Figure 6.

The model predictions of electrolyte potential at different intervals of time are shown in Figure 7. Figure 7a shows the electrolyte potential at the instance of immersion of the material in the electrolyte medium. The model predicted that the electrolyte potential at time $t = 0$, $t = 9614$ and $t = 50,134$ s is 1.424 V , 1.425 V , 1.396 V , respectively. This is shown in Figure 8a–c, respectively.

It is observed that the electrolyte potential is fairly constant for dispersed β phase network, when compared with the continuous β phase network. Similar to the above results, electrochemically active α phase is preferentially dissolved by the electrolyte, while the nobler β phase was intact with the microstructure. This resulted in an increase in the fraction of β phase on the surface. However, the

Fig. 6 Representative image showing the discrete β phase microstructure



dissolution of α phase spattered off the discrete β phase into the electrolyte solution. Thereby decreasing the surface β phase fraction.

The average anodic current density is plotted against time as shown in Figure 8. A significant decrease in average anodic current density was observed with a variation between 2 A m^{-2} and 3 A m^{-2} . The average anodic current density of discrete β phase structure was six magnitudes lesser than the continuous β phase network. The rise and fall of the average anodic current density was attributed to the variation in surface β phase fraction. The average time taken by electrolyte to completely dissolve the material was found to be higher for the discrete β phase network configuration than the continuous β phase network configuration.

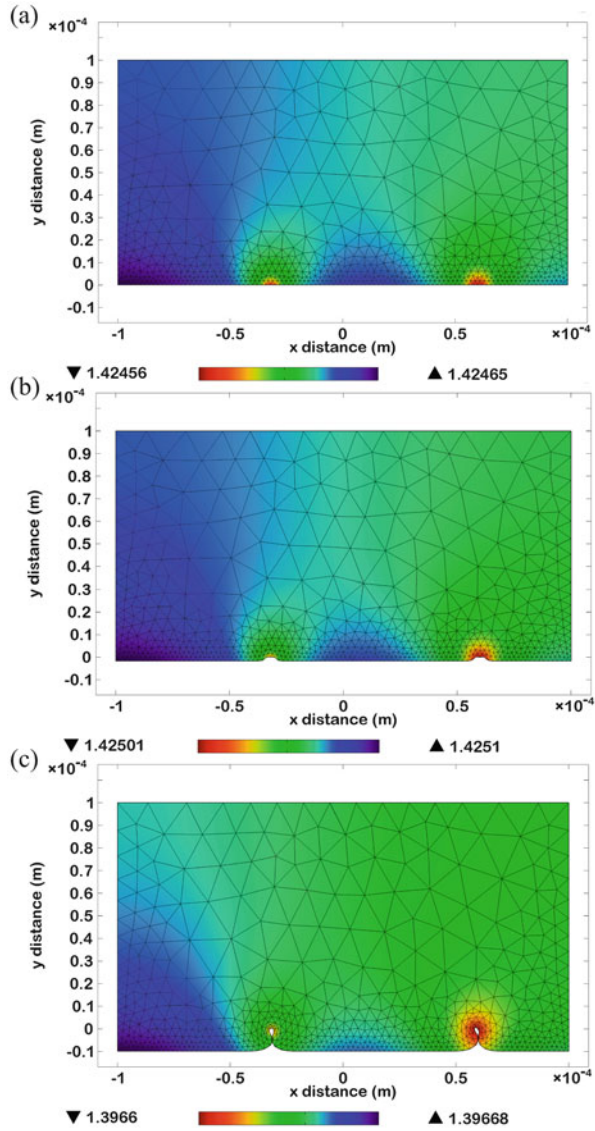
From the graphs Figures 6 and 8, it is observed that the time for dissolution of $10 \mu\text{m}$ in the continuous β phase structure is 28,314 s and the discrete β phase network is 50,134 s. Hence a lesser corrosion rate is observed in discrete β phase network than continuous β phase structure.

4 Conclusion

A numerical model was developed to study the effect of secondary phase distribution on the corrosion behavior of AZ91 Mg alloy using ALE method in Comsol Multiphysics[®]. The model efficiently tracks the moving boundary of the corroding primary phase of the alloy. The results demonstrated the following.

1. The anodic current density increases with increase in surface fraction of secondary phase.
2. The average anodic current density was found to be higher for AZ91 alloy with continuous network of β phase surrounding α phase.
3. The presence of discrete and dispersed β phase in the matrix resulted in low average anodic current.

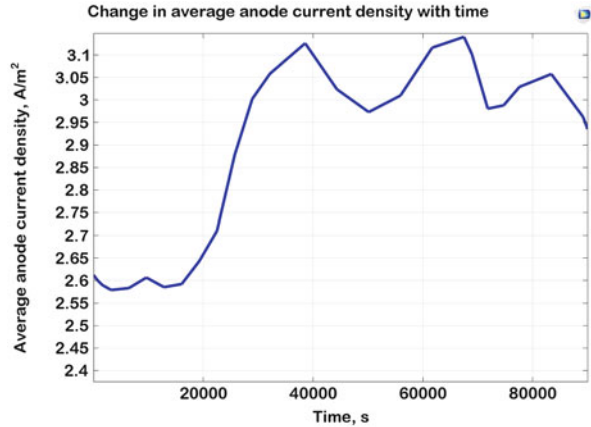
Fig. 7 Predicted electrolyte potential (V) at time (a) $T = 0$ s; (b) $T = 9641$ s; (c) $T = 50,134$ s



4. The average time taken by electrolyte to completely dissolve the material was found to be higher in discrete β phase network configuration than continuous β phase network configuration.

Acknowledgement The authors are grateful to Amrita Vishwa Vidyapeetham, Amrita University, India for their financial support to carry out this investigation through an internally funded research project no. AMRITA/IFRP-20/2016-2017.

Fig. 8 Average anodic current density vs Time discrete β phase microstructure



References

1. Liu, C., Yang, H., Wan, P., Wang, K., Tan, L., Yang, K.: Study on biodegradation of the second phase Mg₁₇Al₁₂ in Mg–Al–Zn Alloys: In vitro experiment and thermodynamic calculation. *Materials Science and Engineering: C* 35, 1-7 (2014)
2. Committee, A.I.H.: *ASM Handbook: Metallography and microstructures*. ASM International (2000)
3. Gobara, M., Shamekh, M., Akid, R.: Improving the corrosion resistance of AZ91D magnesium alloy through reinforcement with titanium carbides and borides. *Journal of Magnesium and Alloys* 3, 112-120 (2015)
4. Hamed Rahimi, R.M., Najafabadi, A.H.: Corrosion and Wear Resistance Characterization of Environmentally Friendly Sol-gel Hybrid Nanocomposite Coating on AA5083. *J. Mater. Sci. Technol.* 29, 603-608 (2013)
5. Kot, I., Krawiec, H.: The use of a multiscale approach in electrochemistry to study the corrosion behaviour of as-cast AZ91 magnesium alloy. *Journal of Solid State Electrochemistry* 19, 2379-2390 (2015)
6. Luo, T.J., Yang, Y.S.: Corrosion properties and corrosion evolution of as-cast AZ91 alloy with rare earth yttrium. *Materials & Design* 32, 5043-5048 (2011)
7. Salman, S., Ichino, R., Okido, M.: A comparative electrochemical study of AZ31 and AZ91 magnesium alloy. *International Journal of Corrosion* 2010, (2010)
8. Wang, L., Zhang, B.-P., Shinohara, T.: Corrosion behavior of AZ91 magnesium alloy in dilute NaCl solutions. *Materials & Design* 31, 857-863 (2010)
9. Ramalingam, V.V., Ramasamy, P.: Modelling Corrosion Behavior of Friction Stir Processed Aluminium Alloy 5083 Using Polynomial: Radial Basis Function. *Trans. Indian Inst. Met.* 1 (<https://doi.org/10.1007/s12666-12017-11110-12661>) (2017)
10. Vignesh, R.V., Padmanaban, R., Arivarasu, M., Thirumalini, S., Gokulachandran, J., Ram, M.S.S.S.: Numerical modelling of thermal phenomenon in friction stir welding of aluminum plates. In: *IOP Conference Series: Materials Science and Engineering*, pp. 012208. IOP Publishing, (2016)
11. Anik, M., Avci, P., Tanverdi, A., Celikyurek, I., Baksan, B., Gurler, R.: Effect of the eutectic phase mixture on the anodic behavior of alloy AZ91. *Materials & Design* 27, 347-355 (2006)
12. Wang, H., Li, Y., Wang, F.: Influence of cerium on passivity behavior of wrought AZ91 alloy. *Electrochimica Acta* 54, 706-713 (2008)

13. Hwang, D.Y., Kim, Y.M., Park, D.-Y., Yoo, B., Shin, D.H.: Corrosion resistance of oxide layers formed on AZ91 Mg alloy in KMnO₄ electrolyte by plasma electrolytic oxidation. *Electrochimica Acta* 54, 5479-5485 (2009)
14. Luo, T.J., Yang, Y.S., Li, Y.J., Dong, X.G.: Influence of rare earth Y on the corrosion behavior of as-cast AZ91 alloy. *Electrochimica Acta* 54, 6433-6437 (2009)
15. Zhan, Y., Zhao, H.-Y., Hu, X.-D., Ju, D.-Y.: Effect of elements Zn, Sn and In on microstructures and performances of AZ91 alloy. *Transactions of Nonferrous Metals Society of China* 20, s318-s323 (2010)
16. Ko, Y.G., Lee, K.M., Shin, D.H.: Electrochemical corrosion properties of AZ91 Mg alloy via plasma electrolytic oxidation and subsequent annealing. *Materials Transactions* 52, 1697-1700 (2011)
17. Ghayad, I., Girgis, N., Azim, A.: Effect of some alloying elements and heat treatment on the corrosion behavior of AZ91 and ZM60 magnesium alloys. *Int. J. Metall. Mater. Sci. Eng* 3, 21-32 (2013)
18. Witte, F., Kaese, V., Haferkamp, H., Switzer, E., Meyer-Lindenberg, A., Wirth, C.J., Windhagen, H.: In vivo corrosion of four magnesium alloys and the associated bone response. *Biomaterials* 26, 3557-3563 (2005)
19. Choudhary, L., Szmerling, J., Goldwasser, R., Raman, R.K.S.: Investigations into stress corrosion cracking behaviour of AZ91D magnesium alloy in physiological environment. *Procedia Engineering* 10, 518-523 (2011)
20. Xue, D., Yun, Y., Tan, Z., Dong, Z., Schulz, M.J.: In Vivo and In Vitro Degradation Behavior of Magnesium Alloys as Biomaterials. *Journal of Materials Science & Technology* 28, 261-267 (2012)
21. Walter, R., Kannan, M.B., He, Y., Sandham, A.: Effect of surface roughness on the in vitro degradation behaviour of a biodegradable magnesium-based alloy. *Applied Surface Science* 279, 343-348 (2013)
22. Tahmasebifar, A., Kayhan, S.M., Evis, Z., Tezcaner, A., Çinici, H., Koç, M.: Mechanical, electrochemical and biocompatibility evaluation of AZ91D magnesium alloy as a biomaterial. *Journal of Alloys and Compounds* 687, 906-919 (2016)
23. Wen, Z., Duan, S., Dai, C., Yang, F., Zhang, F.: Biodegradability and Surface Chemistry of AZ31D Compared with AZ91 Magnesium Alloy in a Modified Simulated Body Fluid. *Int. J. Electrochem. Sci* 9, 7846-7864 (2014)
24. Jia, J.X., Song, G., Atrens, A.: Experimental Measurement and Computer Simulation of Galvanic Corrosion of Magnesium Coupled to Steel. *Advanced Engineering Materials* 9, 65-74 (2007)
25. Deshpande, K.B.: Validated numerical modelling of galvanic corrosion for couples: Magnesium alloy (AE44)-mild steel and AE44-aluminium alloy (AA6063) in brine solution. *Corrosion Science* 52, 3514-3522 (2010)
26. Grogan, J.A., O'Brien, B.J., Leen, S.B., McHugh, P.E.: A corrosion model for bioabsorbable metallic stents. *Acta Biomaterialia* 7, 3523-3533 (2011)
27. Bakhsheshi-Rad, H.R., Abdellahi, M., Hamzah, E., Ismail, A.F., Bahmanpour, M.: Modelling corrosion rate of biodegradable magnesium-based alloys: The case study of Mg-Zn-RE-xCa (x = 0, 0.5, 1.5, 3 and 6 wt%) alloys. *Journal of Alloys and Compounds* 687, 630-642 (2016)
28. Deshpande, K.B.: Numerical modeling of micro-galvanic corrosion. *Electrochimica Acta* 56, 1737-1745 (2011)
29. Jain, V., Mishra, R.S., Gupta, A.K., Gouthama: Study of β -precipitates and their effect on the directional yield asymmetry of friction stir processed and aged AZ91C alloy. *Materials Science and Engineering: A* 560, 500-509 (2013)
30. Mahmudi, R., Kabirian, F., Nematollahi, Z.: Microstructural stability and high-temperature mechanical properties of AZ91 and AZ91 + 2RE magnesium alloys. *Materials & Design* 32, 2583-2589 (2011)

31. Rey, P., Gesto, D., del Valle, J., Verdera, D., Ruano, O.A.: Fine And Ultra-Fine Grained AZ61 And AZ91 Magnesium Alloys Obtained By Friction Stir Processing. In: Materials Science Forum, pp. 1002-1007. Trans Tech Publ, (2012)
32. Shanthi, M., Lim, C.Y.H., Lu, L.: Effects of grain size on the wear of recycled AZ91 Mg. Tribology International 40, 335-338 (2007)
33. Zhang, D., Wang, S., Qiu, C., Zhang, W.: Superplastic tensile behavior of a fine-grained AZ91 magnesium alloy prepared by friction stir processing. Materials Science and Engineering: A 556, 100-106 (2012)

Approximate and Analytic Solution of Some Nonlinear Diffusive Equations



Amitha Manmohan Rao and Arundhati Suresh Warke

1 Introduction

In the recent past, analytical and numerical approximation of nonlinear PDEs describing wave phenomena has been the active area of research in the fields of mathematics, science and engineering. The study of wave behaviour, propagation or nature of wave spreading has real life applications in diverse fields from aerospace engineering to social, biomedical and agricultural sciences as most of wave phenomena in nature is essentially nonlinear. Computation of the numerical or analytic solution of nonlinear diffusive PDEs is important, interesting but quite difficult due to nonlinearity, viscosity effects, discontinuous solutions, conditional stability and convergence to non-realistic solutions. This paper focuses on solving analytically, two nonlinear diffusive equations, which are Newell–Whitehead–Segel (NWS) and Burgers' equations.

The NWS equation is of the following form:

$$u_t(x, t) = ku_{xx}(x, t) + au(x, t) - bu^n(x, t), u = u(x, t), \quad (1)$$

A. M. Rao (✉)

Symbiosis International (Deemed University), Pune, Maharashtra, India

N.S.S. College of Commerce & Economics, Affiliated to University of Mumbai, Mumbai, Maharashtra, India

A. S. Warke

Symbiosis Institute of Technology, Pune, Maharashtra, India

e-mail: arundhatiw@sitpune.edu.in

© Springer Nature Switzerland AG 2019

V. K. Singh et al. (eds.), *Advances in Mathematical Methods and High Performance Computing*, Advances in Mechanics and Mathematics 41, https://doi.org/10.1007/978-3-030-02487-1_31

487

a , b , real constants and k , n , positive integers, appears in modelling various phenomena arising in fluid mechanics and has wide range of applications in various problems arising in mathematical sciences, mechanical, chemical and bio-engineering.

Viscid Burgers' equation is a nonlinear parabolic PDE of the following form [1]:

$$u_t + f(u)_x = \nu u_{xx} \quad u = u(x, t), \quad x \in \mathbb{R} \text{ and } t > 0, \quad (2)$$

where f is the flux function associated with u and ν , the viscosity parameter. Equation (2) is the balance between time evolution, nonlinear wave propagation and dissipative effect and is used as the simplest model equation for nonlinear wave propagation and fluid flow. Equation (2) appears in several areas of applied mathematics and fluid dynamics and is the lowest order approximation of one-dimensional weak shock waves [2]. When $\nu = 0$, Equation (2) becoming hyperbolic in nature, may develop some discontinuities at finite time even for smooth initial data.

Laplace Decomposition Method (LDM), a combination of Adomian Decomposition Method (ADM) and Laplace transform, is one of the efficient analytical techniques to solve nonlinear PDEs. The method has been developed with continuous modifications and implemented to solve diverse types of nonlinear PDEs [3–9]. Solving Equation (1) subject to $u(x, 0) = \lambda$ for arbitrary constant λ for exact solution has been discussed by various researchers by new iterative method [10], combined form of Elzaki transform method and Adomian decomposition method [11] and LDM [12].

Numerical approximation of the PDE given by Equation (2) has been discussed by many researchers. For example, PDE of the type Equation (2) with $u(x, 0) = 0$ was solved by LDM [13]. Although the PDEs considered in this paper are reduced diffusive model equations, it is well known that the wave propagation, wave profile and the nature of the final solution of dissipative PDE depend upon flux function, source term, initial condition and dissipative parameters. For example, Equation (1) has exact solution for $u(x, 0) = \lambda$ for constant parameter λ and the exact solution is given by $u(x, t) = \frac{2\lambda e^{2t}}{2 - \lambda(e^{2t} - 1)}$. But for $u(x, 0) = x$ Equation (1) has only approximate solution. Similarly, Equation (2) may have exact solution but a different source term may result in drastic change in the nature of PDE. Equation (2) can be solved exactly by using Cole-Hopf transformation but adding a source term will alter the physical system. Moreover initial condition with nonlinearity contributes for the nature of the solution.

In view of this, we have considered the numerical approximation of Equation (1) subject to $u(x, 0) = x$ by LDM and $u_t + f(u)_x + u = \nu u_{xx}$, subject to $u(x, 0) = e^{-x^2}$ and $u(x, 0) = e^{-x}$ and for various dissipative parameters by FDM and LDM. The focus is to study the impact of initial condition, source term on the wave propagation and the nature of the final solution as LDM highly depends upon the initial approximation arising out of initial condition and the source term. The error estimation was calculated in all the cases and the convergence of the method is

ensured through plotting of errors for specific values of x and t . The methodology, results and conclusion of the study are discussed in Sections 2, 3 and 4, respectively.

2 Methodology

2.1 Brief Description of LDM

This section describes the general procedure of LDM to solve NWS equation. The same procedure can be applied to solve Equation (2). Following nonlinear PDE is considered.

$$Du(x, t) + Ru(x, t) + Nu(x, t) = g(x, t) \tag{3}$$

subject to the initial condition

$$u(x, 0) = h(x) \tag{4}$$

where $D = \frac{\partial}{\partial t}$, R is a linear operator, N is nonlinear differential operator and $g(x, t)$ is the source term.

The corresponding NWS equation is

$$u_t(x, t) = ku_{xx}(x, t) + au(x, t) - bu^n(x, t), u(x, 0) = h(x)$$

Operating Laplace transform L with respect to t on both sides of Equation (3), we obtain

$$L[Du(x, t)] + L[Ru(x, t)] + L[Nu(x, t)] = L[g(x, t)] \tag{5}$$

By the differential property of Laplace transform and Equations (4) on (5), we get

$$sL[u(x, t)] - h(x) + L[Ru(x, t)] + L[Nu(x, t)] = L[g(x, t)] \tag{6}$$

$$su(x, s) - u(x, 0) = kL(u_{xx}) + au(x, s) - bL(u^n)$$

$$L[u(x, t)] = \frac{h(x)}{s} + \frac{1}{s}L[g(x, t)] - \frac{1}{s}L[Ru(x, t)] - \frac{1}{s}L[Nu(x, t)] \tag{7}$$

As per the procedure of LDM, the solution $u(x, t)$ is expressed as the infinite series of the form

$$u(x, t) = \sum_{n=0}^{\infty} u_n(x, t) \tag{8}$$

And also the nonlinear term $Nu(x, t)$ of Equation (8) is decomposed as

$$Nu(x, t) = \sum_{n=0}^{\infty} A_n(u) \tag{9}$$

where the components $u_n(x, t)$ of Equation (8) will be determined recursively. The Adomian polynomials A_n for arbitrary parameter λ of Equation (9) are computed from the following relation

$$A_n(u) = \frac{1}{n!} \left[\frac{d^n}{d\lambda^n} N \left(\sum_{i=0}^{\infty} \lambda^i u_i \right) \right]_{\lambda=0}, \quad n = 0, 1, 2, 3, \dots \tag{10}$$

By applying inverse Laplace transform to Equations (7), (8), (9) and (10), Adomian polynomials $A_n(u)$, the components of $u_n(x, t)$ and the solution $u(x, t)$ are obtained. The procedure of computation is explained below.

$$u(x, t) = L^{-1} \left(\frac{h(x)}{s} \right) + L^{-1} \left(\frac{1}{s} L [g(x, t)] \right) - L^{-1} \left(\frac{1}{s} L [Ru(x, t)] \right) - L^{-1} \left(\frac{1}{s} L [Nu(x, t)] \right),$$

$$u(x, t) = L^{-1} \left(\frac{h(x)}{s-a} \right) + L^{-1} \left[\frac{k}{s-a} L \left(\frac{\partial^2}{\partial x^2} \left(\sum_{n=0}^{\infty} u_n(x, t) \right) \right) \right] - L^{-1} \left[\frac{b}{s-a} L \sum_{n=0}^{\infty} A_n(u), \right]$$

where $\sum_{n=0}^{\infty} A_n(u) = u^n$ and $\sum_{n=0}^{\infty} u_n(x, t) = u(x, t)$ The recursive relation for the components of $u(x, t)$ is obtained by

$$u_0(x, t) = h(x)e^{at}$$

Here $u_0(x, t)$, the initial approximation, represents the term arising out of source term and given initial condition.

$$u_{(n+1)}(x, t) = L^{-1} \left[\frac{k}{s-a} L \left(\frac{\partial^2}{\partial x^2} \left(\sum_{n=0}^{\infty} u_n(x, t) \right) \right) \right] - L^{-1} \left[\frac{b}{s-a} L \sum_{n=0}^{\infty} A_n(u) \right], \quad n = 0, 1, 2, \dots \tag{11}$$

The algorithm presented in [14] has been used to calculate Adomian polynomials.

2.2 Finite Difference Method (FDM)

Finite difference method (FDM) is a powerful numerical technique used for finding approximate solutions for linear and nonlinear PDE. The basic principle of FDM is the approximation of derivatives in PDE by finite differences. In this paper we have used first order forward difference operator to t and central difference operator to x given by the following relations.

$$\begin{aligned} \text{Forward difference : } \left(\frac{\partial u}{\partial t}\right)_j &\cong \frac{u_{i,j+1}-u_{i,j}}{\Delta t}, \\ \text{Central Difference : } \left(\frac{\partial u}{\partial x}\right)_i &\cong \frac{u_{i+1,j}-u_{i-1,j}}{2\Delta x} \end{aligned}$$

3 Results and Discussion

3.1 Illustrative Examples of Equation (1) by LDM

For $k = 5, a = 2, b = -1, n = 2, h(x) = x$, Equation (1) becomes

$$u_t = 5u_{xx} + 2u + u^2, u(x, 0) = x \tag{12}$$

Various components of the numerical approximation of Equation (12) obtained by applying LDM are given as below.

$$u_0(x, t) = xe^{2t} \tag{13}$$

$$A_0(u) = u_0^2 = x^2e^{4t}$$

$$u_1(x, t) = L^{-1} \left[\frac{5}{s-2} L \left(\frac{\partial^2}{\partial x^2} (u_0) \right) \right] + L^{-1} \left[\frac{1}{s-2} L (A_0) \right] = \frac{x^2}{2} (e^{4t} - e^{2t})$$

$$A_1(u) = 2u_0u_1 = x^3 (e^{6t} - e^{4t})$$

$$u_2(x, t) = \frac{x^3e^{2t}}{4} (1 - 2e^{2t} + e^{4t}) + \frac{5e^{2t}}{2} (e^{2t} - 1 - 2t)$$

Substituting $u_0(x, t), u_1(x, t), u_2(x, t)$ in $\sum_{n=0}^{\infty} u_n(x, t) = u(x, t)$, we get the series solution. We have calculated first five terms and first ten terms approximations of $u(x, t)$ of Equation (12) for error estimation and understanding wave propagation.

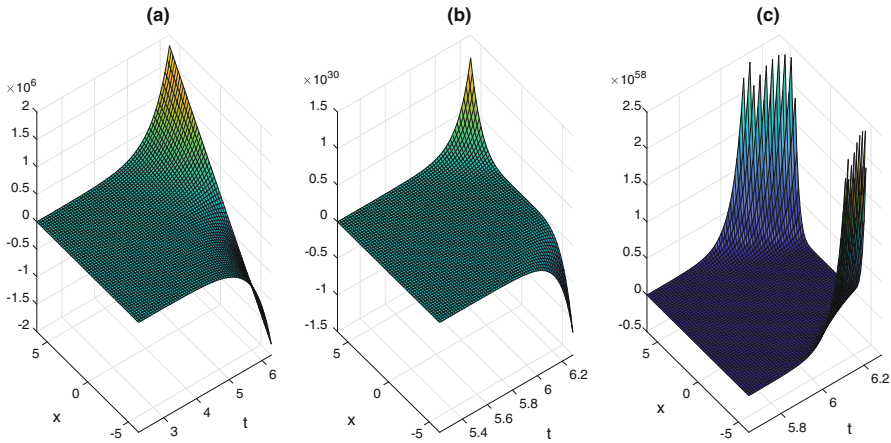


Fig. 1 (a) Initial approximation of Equation (11). (b) Five-term. (c) Ten-term approximation

Table 1 Absolute errors of ten-term approximation solution of Equation (12) by LDM

$x \backslash t$	$\left \sum_{n=0}^9 u_n - u_0 \right $			$\left \sum_{n=0}^9 u_n - \sum_{n=0}^4 u_n \right $		
	$t = 0.1$	$t = 0.2$	$t = 0.3$	$t = 0.1$	$t = 0.2$	$t = 0.3$
0.2	0.0758	0.4648	2.4297	9.0000e-04	0.0584	1.1337
0.4	0.0973	0.5874	3.3524	0.0010	0.0749	1.6296
0.6	0.1317	0.7666	4.7043	0.0013	0.1001	2.3874
0.8	0.1798	1.0147	6.6933	0.0017	0.1391	3.5815
1.0	0.2426	1.3475	9.6258	0.0022	0.1989	5.4741

$$\text{Five-term : } u(x, t) \cong u_0 + u_1 + u_2 + u_3 + u_4 \tag{14}$$

$$\text{Ten-term : } u(x, t) \cong u_0 + u_1 + u_2 + u_3 + \dots + u_9 \tag{15}$$

The graphs of Equations (13), (14) and (15) have been depicted as in Figure 1 through which the wave propagation at the initial stage at five-term approximation and ten-term approximation can be understood. The error estimation for different values of t and x in $(0,1]$ is presented in Table 1. From Table 1, it is clear that absolute error is small for values of x nearer to 0 and it increases towards 1. The error between ten-term approximation and initial approximation is compared with the error between ten-term approximation and five-term approximation and rapid convergence of LDM is achieved.

For $k = 1, a = 2, b = 3, n = 2, h(x) = x$, Equation (1) becomes

$$u_t = u_{xx} + 2u - 3u^2, u(x, 0) = x \tag{16}$$

$$u(x, t) = L^{-1} \left(\frac{x}{s-2} \right) + L^{-1} \left[\frac{1}{s-2} \left(L \left(\frac{\partial^2}{\partial x^2} \sum_{n=0}^{\infty} u_n(x, t) \right) \right) \right] - L^{-1} \left[\frac{3}{s-2} L \left(\sum_{n=0}^{\infty} A_n(u) \right) \right]$$

$$u_0(x, t) = x e^{2t} \tag{17}$$

$$u_1(x, t) = \frac{3x^2}{2} (e^{2t} - e^{4t})$$

Five-term and ten-term series approximations of $u(x, t)$ of Equation (16) are calculated here.

$$\text{Five-term : } u(x, t) \cong u_0 + u_1 + u_2 + u_3 + u_4 \tag{18}$$

$$\text{Ten-term : } u(x, t) \cong u_0 + u_1 + u_2 + u_3 + \dots + u_9 \tag{19}$$

The graphs of Equations (13), (18) and (19) have been depicted in Figure 2. The effect of dissipative parameter and nonlinear terms of Equations (12) and (16) can be seen by comparing (c) of Figures 1 and 2. The graphs in (a) and (b) of Figures 1 and 2 indicate that the wave profiles of both the solutions are almost same in the initial stages. The effect can be seen for better approximations. The tabulation of absolute

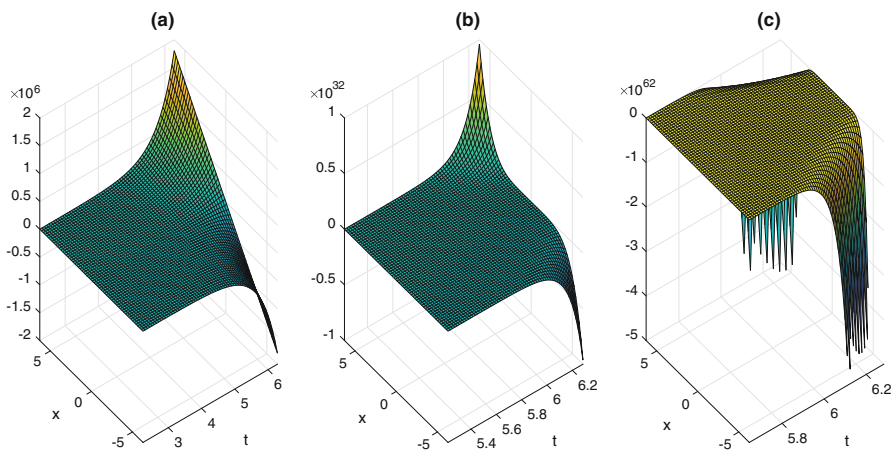


Fig. 2 (a) Initial profile of Equation (16). (b) Five-term. (c) Ten-term approximation

Table 2 Absolute errors of ten-term approximation solution of Equation (16) by LDM

$x \backslash t$	$\left \sum_{n=0}^9 u_n - u_0 \right $			$\left \sum_{n=0}^9 u_n - \sum_{n=0}^4 u_n \right $		
	$t = 0.01$	$t = 0.02$	$t = 0.03$	$t = 0.01$	$t = 0.02$	$t = 0.03$
0.2	0.0015	0.0037	0.0067	5.3563e-09	1.7378e-07	1.3379e-06
0.4	0.0052	0.0111	0.0179	1.8268e-09	5.2021e-08	3.4458e-07
0.6	0.0112	0.0233	0.0361	6.5823e-09	2.2807e-07	1.8641e-06
0.8	0.0196	0.0400	0.0611	1.9298e-08	6.4012e-07	5.0340e-06
1.0	0.0303	0.0611	0.0924	3.0189e-08	9.7614e-07	7.4772e-06

errors for different values of t and x in $(0,1]$ is shown in Table 2. A comparison of the errors between initial and five-term approximation indicates the efficiency of LDM. From Table 1 and Table 2, it has been observed that, as diffusive parameter decreases, there is significant decrease in errors.

3.2 Illustrative Examples of Burgers' Equation by FDM

Consider

$$u_t + f(u)_x + u = v u_{xx}, f(u) = \frac{1}{2}u^2, u(x, 0) = e^{-x^2}$$

$$u_t + uu_x + u = v u_{xx}, u(x, 0) = e^{-x^2} \tag{20}$$

Applying FDM to Equation (20), forward and central difference operator to t and x , respectively, $\frac{u_{i,j+1} - u_{i,j}}{\Delta t} + u_{i,j} \left(\frac{u_{i+1,j} - u_{i-1,j}}{2\Delta x} \right) + u_{i,j} = v \left(\frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j}}{(\Delta x)^2} \right)$
 Simplification of the above expression gives the expression for $u_{i,j+1}$ as

$$u_{i,j+1} = u_{i,j} \left[1 - \Delta t - \frac{\Delta t}{2(\Delta x)^2} - \frac{\Delta t}{2\Delta x} (u_{i+1,j} - u_{i-1,j}) \right] + \frac{v\Delta t}{2(\Delta x)^2} (u_{i+1,j} + u_{i-1,j})$$

By using initial condition, $u(x, 0) = e^{-x^2}$, we get the successive approximations. Here we have calculated first seven approximations for $\Delta t = 0.01, \Delta x = 0.25$, which are depicted in Figure 3 and the comparison of the same with solution by LDM (15) is shown in Figure 4. The graph of absolute errors of the Equation (20) by LDM for different dissipative values is depicted in Figure 5. The numerical results of $u(x, t)$ and the relative errors for various values of x and t are given in Table 3. The absolute errors of FDM approximations of the equation (20) for

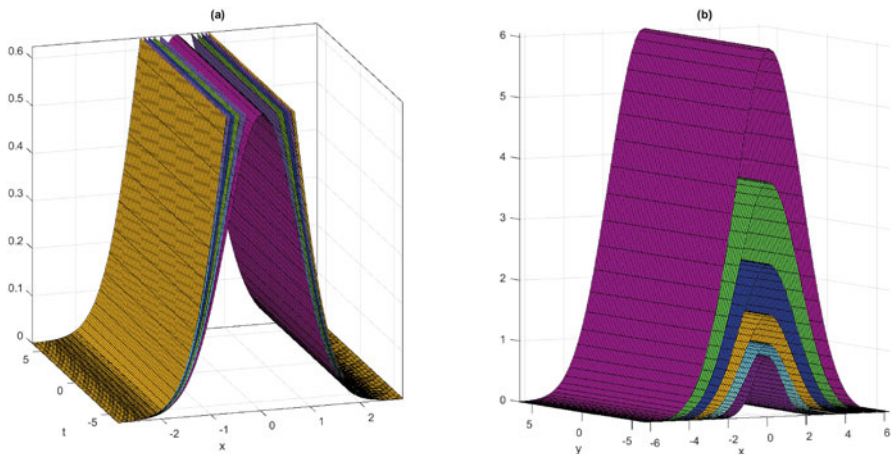


Fig. 3 Graph of seven-iteration solution of Equation (20) by FDM (a) $\nu = 0$ (b) $\nu = 10$

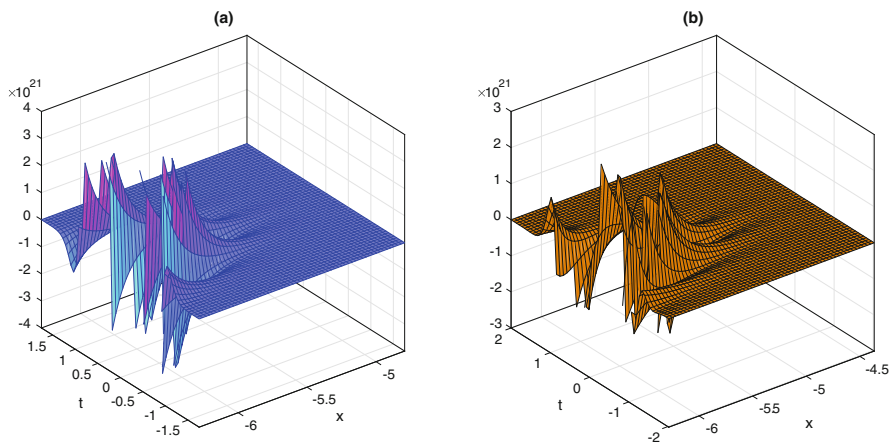


Fig. 4 Graph of eight-term approximation of Equation (20) by LDM (a) $\nu = 0$ (b) $\nu = 10$

$\nu = 0, \Delta t = 0.01, \Delta x = 0.5$ have been presented in the Table 4. It can be noted that $u(x, t)$ decreases with increase in both x and t . Figures 3 and 4 present the distortion of waves due to combined effect of nonlinearity, source term and dissipative parameter.

Consider the equation

$$u_t + uu_x + u = \nu u_{xx}, u(x, 0) = e^{-x} \tag{21}$$

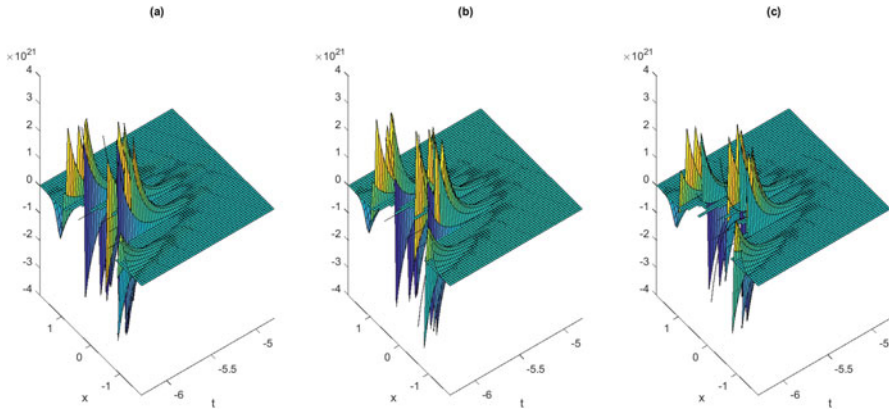


Fig. 5 Graph of error function of the Equation (20) by LDM (a) $\nu = 0$ (b) $\nu = 0.5$ (c) $\nu = 1.0$

Table 3 Relative errors of nine-term approximation solution of Equation (20) by LDM for $\nu = 0$

x	t	$\sum_0^8 u_n(x, t)$				Relative errors		
		0	1	2	3	1	2	3
0	0.001	0.9990	0.9970	0.9970	0.9970	0.0020	7.5173e-06	7.5113e-06
	0.002	0.9980	0.9940	0.9940	0.9940	0.0040	3.0139e-05	3.0139e-05
0.5	0.001	0.7780	0.7778	0.7778	0.7778	1.9288e-04	5.9595e-05	5.9595e-05
	0.002	0.7772	0.7769	0.7769	0.7769	3.8657e-04	1.9000e-05	1.9000e-05
1.0	0.001	0.3675	0.3685	0.3685	0.3685	0.0028	3.8392e-05	3.8392e-05
	0.002	0.3671	0.3692	0.3691	0.3691	0.0056	1.5934e-04	1.5934e-04

Table 4 Absolute errors of the solution of Equation (20) by FDM for $\nu = 0, \Delta t = 0.01, \Delta x = 0.5$

x	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0	0.0665	0.0625	0.0584	0.0545	0.0499
0.5	0.0459	0.0428	0.0399	0.0371	0.0344
1	0.0299	0.0277	0.0256	0.0236	0.0218

By proceeding as in Section 3.2 (see consider the equation) and by applying FDM and LDM on Equation (21), we obtain approximate solution and relative errors. Figures 6, 7 and 8, respectively, depict the solution by FDM, LDM [15] and error estimation by LDM for different dissipative values. The error estimation for specific values of x and t by LDM and FDM is presented in Tables 5 and 6. It is observed that ten-term approximation and u_{appr} are same for $n = 1$ only, indicating the rapid convergence of the method. Comparison of Figure 3 with Figure 6 and Figure 4 with Figure 7 indicates the influence of initial conditions on the final solution obtained by both FDM and LDM of given nonlinear equation. Figure 9 presents the graph of absolute errors of Equation (21) calculated by FDM with $\nu = 0$ clearly showing the convergence of the method.

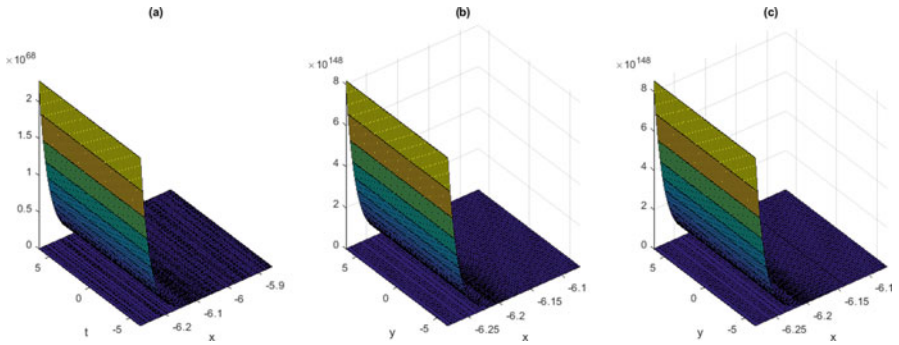


Fig. 6 Surface plot of approximation of Equation (21) by FDM (a) $\nu = 0$ (b) $\nu = 0.5$ (c) $\nu = 1.0$

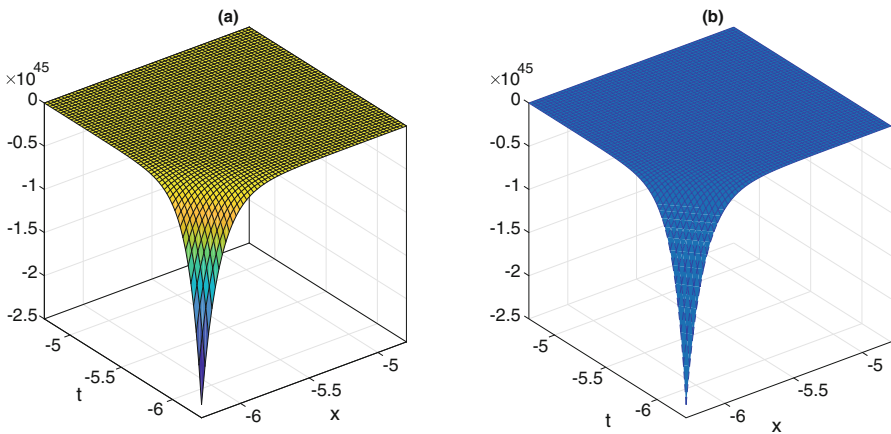


Fig. 7 Graph of approximation of Equation (21) by LDM (a) $\nu = 0$ (b) $\nu = 10$

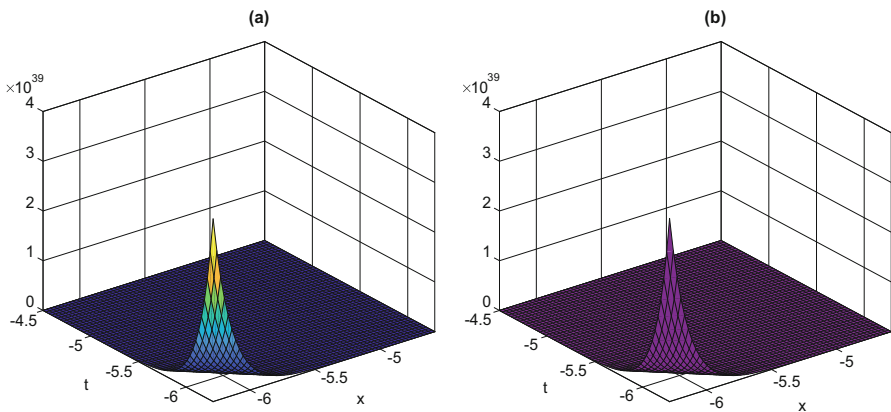


Fig. 8 Graph of absolute errors of Equation (21) by LDM (a) $\nu = 0$ (b) $\nu = 10$

Table 5 Relative errors of ten-term approximation solution of Equation (21) by LDM for $\nu = 0$ where $u_{appr} = \sum_0^9 u_n(x, t)$

x	t	u_{appr}	$\sum_0^n u_n(x, t)$					Relative errors				
			0	1	2	3	8	0	1	2	3	8
0	0.001	1.001	0.999	1.001	1.001	1.001	1.001	0.002	0	0	0	0
	0.002	1.0020	0.998	1.002	1.002	1.002	1.002	0.004	0	0	0	0
0.5	0.001	0.6069	0.6059	0.6069	0.6069	0.6069	0.6069	0.0016	0	0	0	0
	0.002	0.6073	0.6053	0.6073	0.6073	0.6073	0.6073	0.0032	0	0	0	0
1.0	0.001	0.3680	0.3675	0.3680	0.3680	0.3680	0.3680	0.0013	0	0	0	0
	0.002	0.3682	0.3671	0.3681	0.3682	0.3682	0.3682	0.0029	0	0	0	0

Table 6 Absolute errors of the solution of Equation (21) by FDM for $\nu = 0$

x	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$
0	0.0665	0.06245	0.05849	0.05446	0.04993
0.5	0.04594	0.04283	0.03987	0.03707	0.03439
1	0.02993	0.02768	0.02557	0.0236	0.02177

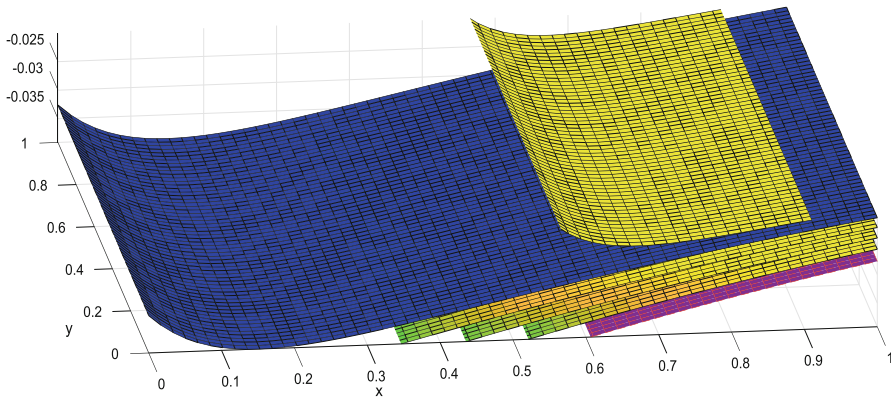


Fig. 9 Graph of absolute errors of Equation (21) by FDM for $\nu = 0$

4 Conclusion

In this paper, LDM and FDM have been successfully implemented to solve parabolic nonlinear PDEs, NWS equation, and Burgers’ equation. The impact of initial condition and dissipative coefficients on the wave propagation of the solution was studied through MATLAB graphs of numerical approximations. The convergence of the methods was ensured through error estimation for specific values of x and t and graphical representation. The wave profiles of solution of Burgers’ equation for various dissipative coefficients by both FDM and LDM are compared to show the rapid convergence of LDM. The approach can be extended to obtain physically relevant solutions to a wide range of nonlinear PDEs of real life phenomena involving nonlinear and dissipative effects by considering different source terms and initial conditions.

References

1. J.M. Burgers, *The Nonlinear Diffusion Equation*, Reidel, Dordrecht, 1974.
2. V.I. Karpman, *Non-Linear Waves in Dispersive Media*, Pergamon, Oxford, 1975.
3. S.A. Khuri, *A Laplace decomposition algorithm applied to a class of nonlinear differential equations*, Journal of Applied Mathematics. 1(2001), 141-155.
4. S.A. Khuri, *A new approach to Bratus problem*, Applied Mathematics and Computation. 1(2004), 131-136.
5. H. Jafari, *Application of the Laplace decomposition method for solving linear and nonlinear fractional diffusion-wave equations*, Applied Mathematics Letters. 24(2011), 1799-1805.
6. M. A. Hussain, *Modified Laplace Decomposition Method*, Applied Mathematical Science. 4(2010), 1769-1783.
7. M. E. Khan, *Application of Laplace Decomposition Method to Solve Nonlinear Coupled Partial Differential Equations*, World Applied Sciences Journal, 8(2010), 13-19.
8. Y. A. Khan, *Application of modified Laplace decomposition method for solving boundary layer equation*, Journal of King Saud University-Science, 23(2011), 115-119.
9. J. Fadaei, *Application of Laplace–Adomian Decomposition Method on Linear and Nonlinear System of PDEs*, Applied Mathematical Sciences, 5(2011), 1307-1315.
10. J. Patad, S. Bhalekar, *Approximate analytical solutions of Newell-Whitehead-Segel equation using a new iterative method*, World Journal of Modelling and Simulation. 11(2015), 94-103.
11. M. M. A. Mahgoub and A.K.H. Sedeeg, *On The Solution of Newell-Whitehead-Segel Equation*, American Journal of Mathematical and Computer Modelling. 1(2016), 21-24. <https://doi.org/10.11648/j.ajmcm.20160101.13>
12. P. Pue-on, *Laplace Adomian Decomposition Method for Solving Newell-Whitehead-Segel Equation*, Applied Mathematical Sciences. 7(2013), 6593 – 6600.
13. M. Hussain and M. Khan, *Modified Laplace Decomposition Method*, Applied Mathematical Sciences. 4(2010), 1769 - 1783
14. A. M. Wazwaz, *A new algorithm for calculating Adomian polynomials for nonlinear operators*, Applied Mathematics and Computation. 111(2000), 33-5.
15. A. M. Rao, and A. S. Warke, *Laplace Decomposition Method (LDM) for Solving Nonlinear Gas Dynamic Equation*, Annals of the Faculty of Engineering Hunedoara-International Journal of Engineering. 2(2015), 147-150.

Index

A

Abstract stochastic evolution, 163–178
Adaptive local iterative filtering (ALIF), 72, 80, 81
Additional food, 118, 124, 129, 131
Adomian polynomials, 490
Algorithmic convergence, 176–178
ALIF, *see* Adaptive local iterative filtering (ALIF)
Ambipolar diffusion (AD), 358–360, 363–368
ANN, *see* Artificial neural network (ANN)
Approximate solution, 54, 146, 147, 487–498
Artificial neural network (ANN), 83–101, 296, 297, 303
Asymmetric collocation, 105–114
AZ91, 471–484

B

Big data, 250, 255, 257, 309–318
Bilevel Knapsack problem, 35
Biomagnetic fluid, 403, 404, 406
Bloom, 117, 118, 121, 123, 124, 127–129, 131

C

Canonical dual, 4, 9, 12, 16–18, 20, 21, 23, 26–29, 35, 37, 39, 41, 42, 210, 214–220, 223, 224, 239, 241
Canonical duality, 3–46, 209–244
Canonical duality theory (CDT), 4, 7, 9, 10, 15, 19, 21, 23, 25, 31, 32, 34–38, 40, 41, 45, 209–244

Canonical penalty-duality method (CPD), 210, 211, 220–233, 235–238, 240, 243, 244
Carbon dioxide (CO₂), 117, 118, 126, 131
CDT, *see* Canonical duality theory (CDT)
Chemical reaction, 371–384
Complaint walls, 372
Complexity, 17, 117, 136, 144, 211, 223–224, 240, 259, 276, 282, 297, 310
Computational, 4, 5, 21, 36, 37, 52, 106, 135–147, 195–197, 209–211, 219, 223, 243, 254, 255, 262, 263, 283, 295, 310, 311, 318, 347, 349, 353, 450, 453, 461, 462, 474–478
Computational simulation, 283, 347–351
Controllability, 149–160
Convective condition, 457–468
Corrosion, 471–484
Couple stress, 371–384, 429–446
Couple stress fluid, 371–384, 429–446
Cylindrical cloak, 323

D

Darcy number, 416, 418–421, 425, 431
Data science, 295, 296
Dimensionality reduction, 295–298
Double diffusive convection, 431, 441, 446
Dynamical control system, 149

E

Elliptical cloaking phenomena, 324–334
Elliptic cloak, 323–334
Empirical mode decomposition (EMD), 71–74

Exascale computing, 283
 Exponential functions, 181–191

F

Fault model, 286–288
 Field programmable gates arrays (FPGAs),
 259–261, 264–268
 Finite difference method (FDM), 488, 491,
 494–498
 Finite element method (FEM), 9, 21, 31, 33,
 36, 209–211, 285, 323, 450, 451, 455,
 475
 Fixed points, 8, 24–25, 44, 149, 150, 156, 157,
 201–207
 FPGAs, *see* Field programmable gates arrays
 (FPGAs)
 Fractional inverse, 387–401

G

Gap function, 4, 15–17, 21, 22, 216
 Geometrical nonlinearity, 16, 17, 19, 46
 Global optimization, 3–46, 209, 210, 244

H

Hierarchical Tucker Format, 272–277
 High-performance computing (HPC),
 249–257, 259, 264, 268, 284, 299,
 309–318
 HPC architecture, 250–252, 257
 HT format, 272–280, 282
 Human impact, 117–132
 Hybrid finite element method, 449–455
 Hydromagnetic flow, 457–468

I

Inclined channel, 371–384
 Internal heat source, 429–446
 Irregular-shaped harbor, 450, 454

J

Jeffrey fluid, 415–425

K

Kansa's, 105, 106, 108, 110–114

L

Laplace decomposition method (LDM),
 488–498
 Lebesgue dominated convergence, 158
 Light intensity, 125
 Linear fractional transformation, 193
 Low rank, 271–282

M

Magnesium alloy, 471–484
 Magnetic field, 69, 357–359, 362–368, 372,
 373, 376, 379, 403–407, 409, 413, 416,
 430, 457, 458
 Magnetization, 403, 404, 406
 Magneto-hydrodynamic (MHD), 357–368,
 371–384, 403, 404, 416, 457
 Mean field dynamo, 357–360, 362, 365, 367,
 368
 Menger probabilistic metric spaces, 201–207
 MHD, *see* Magneto-hydrodynamic (MHD)
 MHD Dynamo, 357–368
 Mild slope equation (MSE), 90, 92–95, 97–99,
 267, 449, 450, 455
 Minimum connectivity inference (MCI),
 135–138, 140, 141, 146, 147
 Mixed integer, 25–26
 Modern Parallel Architecture, 259–268
 Modified Post-Widder, 181–191
 MSE, *see* Mild slope equation (MSE)
 Multidimensional data, 295–298, 300, 301,
 303
 Multi-scale modeling, 5–14
 Municipal solid waste (MSW), 84, 85,
 87–101
 Municipal solid waste management, 83

N

Neural process, 52
 Newell–Whitehead–Segel (NWS) equation,
 487, 489, 498
 Newtonian heating, 457–468
 Nonlinear PDE, 21, 487–489, 491, 498
 Nonstationary signal, 69–81
 NP-hardness, 3–46, 243
 Neural field, 51–66
 Numerical model, 453, 467, 473, 482
 Numerical simulation, 123–124, 259–268, 323,
 453–455

O

Objectivity, 3–6, 10, 11, 14, 15, 32, 240–242
 Operator splitting scheme, 163–178

P

Parabolic equation, 498
 Parallel algorithms, 271–282, 314
 Parallel Architecture, 259–268
 Peristalsis, 371–373, 415, 416, 421
 Peristaltic flow, 371, 372, 416
 Porous media, 338, 352, 429, 430, 440
 Post-Widder operator, 181–191
 Power efficiency, 253–254
 Properly posed problem, 5–14, 20, 40, 46

Q

Quantum tunneling, 283

R

Radial basis functions III-conditioned system,
 105–114
 Reduction rules, 137–146
 Regional economic development, 296,
 299–303
 Regularization, 4, 105–114, 261
 Riccati equation, 198, 199

S

SBF, *see* Simulated body fluid (SBF)
 Schwarzian derivative, 193, 195, 198, 199
 Scientific computing, 309–318
 Signal decomposition, 69–81
 Simulated body fluid (SBF), 471–484

Slip effects, 416

Solar magnetic fields, 357, 363
 Soret parameter, 444–446
 Spectral data, 390, 391, 394, 397, 399, 400
 Splitting operator, 163–178
 Stagnation point, 457–468
 Stochastic impulsive system, 150
 Stochastic neural field, 65
 Stretching surface, 403–413, 457–468

T

Tensor arithmetic, 271–282
 Tensors, 4, 6, 9, 14, 15, 218, 240, 241, 271–282
 Thermal conductivity, 404, 406, 409, 411–413,
 416, 459, 463
 Topology optimization, 32–36, 209–244
 Toxin, 117–132
 Triality, 4, 5, 7, 17, 19–21, 36, 41, 210, 243

V

Variable, 5, 8, 9, 20, 25, 30, 32, 33, 43, 44, 51,
 55, 77, 84, 88–100, 110, 126, 127, 137,
 146, 151, 194, 211–213, 224, 239, 297,
 347, 360, 361, 403–413, 416, 418, 431,
 433, 450, 459, 460, 479
 Variable viscosity, 404, 416
 Visualization, 255, 295–305, 309, 339

W

Wall features, 371–384
 Wave oscillation, 449–455
 Weak contraction, 201
 Web-based solution, 303