Fields Institute Communications

VOLUME 66

The Fields Institute for Research in Mathematical Sciences

The Fields Institute is a centre for research in the mathematical sciences, located in Toronto, Canada. The Institute's mission is to advance global mathematical activity in the areas of research, education and innovation. The Fields Institute is supported by the Ontario Ministry of Training, Colleges and Universities, the Natural Sciences and Engineering Research Council of Canada, and seven Principal Sponsoring Universities in Ontario (Carleton, McMaster, Ottawa, Toronto, Waterloo, Western and York), as well as by a growing list of Affiliate Universities in Canada, the U.S. and Europe, and several commercial and industrial partners.

For further volumes:
www.springer.com/series/10503

Roderick Melnik • Ilias Kotsireas
Editors

Advances in Applied Mathematics, Modeling, and Computational Science

The Fields Institute for Research
FIELDS in the Mathematical Sciences

Springer

Editors
Roderick Melnik
M²NeT Laboratory and
Department of Mathematics
Wilfrid Laurier University
Waterloo, ON
Canada

Ilias Kotsireas
Department of Physics and
Computer Science
Wilfrid Laurier University
Waterloo, ON
Canada

ISSN 1069-5265
Fields Institute Communications
ISBN 978-1-4614-5388-8
DOI 10.1007/978-1-4614-5389-5
Springer New York Heidelberg Dordrecht London

ISSN 2194-1564 (electronic)

ISBN 978-1-4614-5389-5 (eBook)

Library of Congress Control Number: 2012949685

Mathematics Subject Classification (2010): 00-02, 00B20, 65-02, 00A71, 92B05

Cover: Drawing of J.C. Fields by Keith Yeomans

Printed on acid-free paper

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

In memory of M.M.

"What you leave behind is not what is
engraved in stone monuments,
but what is woven into the lives of others."

(Pericles, 495 BC–429 BC)

Preface

Methods and tools of applied mathematics and computational science are becoming increasingly important in an ever increasing range of application areas. Mathematical models are becoming pervasive in our society, in its development and well-being. Along with more traditional areas of mathematics applications in natural sciences and engineering, mathematical models steadily and convincingly penetrate such areas as business and economics, urban planning and information management, health, social and political sciences, to name just a few. With such a wide spectrum of application areas, in describing recent advances in applied mathematics, modeling and computational science, we necessarily have to restrict ourselves by a selection of topics. As a result, this volume presents a selection of in-depth studies and state-of-the-art surveys of several challenging topics that are at the forefront of modern applied mathematics, mathematical modeling, and computational science. These three areas represent the foundation upon which the methodology of mathematical modeling and computational experiment is built as a major tool in all areas of applications of mathematics. The nine chapters of this book cover both fundamental and applied research, and provide the reader with state-of-the-art achievements in the development and application of new theories at the interfaces of applied mathematics, modeling and computational science.

The book can serve as a reference on several important current topics in modern applied mathematics and modeling, including random matrix theory with its innovative applications, dynamic blocking problems, elliptic curves over finite fields and their cryptographic applications, optimal control applications combining discrete and continuous features, among others. The reader can find in this book comprehensive accounts of recent advances in other important topics such as multiple scale methods coupling network and continuum models and their applications in various areas involving porous media, as well as statistical geometric and topological techniques and their applications in the life sciences. Two chapters of the book are devoted to recent developments in state-of-the-art numerical procedures for solving complex mathematical models based on partial differential equations. This includes energy stable weighted essentially non-oscillatory schemes with applications in fluid dynamics and aerospace sciences, as well as new efficient schemes

for hyperbolic equations based on the inverse Lax–Wendroff procedure for numerical boundary conditions. In these and other chapters of the book, the reader can find a wide spectrum of most advanced modeling techniques with their demonstration on a series of examples. Such techniques are explained with both, rigorous mathematics and details of numerical algorithms for their computer implementation. Several easy-to-run computer codes are also provided.

The material presented in this book aims at fostering interdisciplinary collaborations required to meet the modern challenges of applied mathematics, modeling and computational science. Due to a combination of rigorous mathematical and computational procedures with examples from a variety of applications ranging from engineering to life sciences, the book can be useful for a wide audience of professionals from different disciplines, including graduate students. It provides a rich source for graduate student projects, and can be used in courses or seminars on selected topics. Researchers in academia and industry and anyone who is interested in modern applications of mathematics and computational science, and advanced mathematical modeling techniques would benefit from this book.

We would like to thank our many colleagues around the world for their encouragement and fruitful discussions, and the NSERC, Ikerbasque, and the CRC Program for their support. We are thankful to the referees on both sides of the Atlantic for their invaluable professional help. We would like to express our gratitude to Dr. Carl Riehm and Debbie Iscoe from the Fields Institute for Research in Mathematical Sciences for their support, and the Springer team, in particular to Elizabeth Loew and Dahlia Fisch, for their assistance in completing this project.

Waterloo (Canada)–Bilbao (Spain) Roderick Melnik
July 2012 Ilias Kotsireas

Contents

Interconnected Challenges and New Perspectives in Applied Mathematical and Computational Sciences

Roderick V. N. Melnik and Ilias S. Kotsireas

Abstract More and more disciplines increasingly rely in their progress on the development of mathematical models, methods, and algorithms. Mathematical modeling and computational science are seen as a driving force in scientific discovery and innovative, mathematics-based technologies. They can facilitate new advances in systems-science-based approaches in applications and serve as a decisive factor in sustainable socio-economic development of the society. The chapter provides an overview of these new trends.

1 Mathematical Models and Algorithms

The methods and ideas of mathematics are taking on new and new spheres of influence. This process, started well before the famous quote by Galileo Galilei that "the Book of Nature is written in the language of mathematics", can be traced back to the thoughts of Pythagoras' school, while the origin of its path is hidden at the dawn of human civilization. Today, all existing advanced information technologies are essentially *mathematics-based technologies*, including mobile communication and internet. Mathematical models and mathematics-based quantitative analyses have penetrated to and are becoming increasingly important in the areas that were only recently labeled as non-traditional for conventional mathematics, ranging from life science and medicine, to business and economics, and to social, political and forensic sciences, to name just a few. What made this process developing at a much faster pace than ever before, in the last 70 years or so, is the ready availability of computer

R.V.N. Melnik (✉)
M²NeT Laboratory and Department of Mathematics, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada N2L 3C5
e-mail: rmelnik@wlu.ca

R.V.N. Melnik
Ikerbasque, Basque Foundation for Science and BCAM, Bilbao, 48011, Spain

I.S. Kotsireas
Department of Physics and Computer Science, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada N2L 3C5

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5_1,

power. This has led to a situation where many problems in sciences and engineering that could not be solved before found their way to be analyzed and explored computationally. Mathematics with its foundations in applications and computational science with its mathematics-based numerical *algorithms* become two wings of the same bird. What connects them together is the body of human knowledge which we call *mathematical modeling*. Under mathematical models we understand connections, patterns, or abstract constructs that characterize a phenomenon, a process, or a system such that they can be expressed in some mathematical form via equations, inequalities, formulae, sets of rules, etc. The original stimuli for the development of simplest mathematical models were agricultural needs of humans, subsequent development of geometric knowledge, arithmetics (as a precursor of number theory), and their applications. Probably, one of the most famous examples of such applications in the Ancient Times were the Egyptian Pyramids (with the earliest known, the Pyramid of Djoser, constructed 2630 BCE–2611 BCE). As pointed out by H. Poincare, "the true method of forecasting the future of mathematics lies in the study of its history and its present state" [1]. Many early examples of applications of mathematical knowledge demonstrate that such applications were closely interwoven with the development of mathematical algorithms, providing key stimuli for their applications. As abstract areas of mathematics are also based on models, mathematical algorithms have been fundamental in their development too. Unquestionably, algorithms are also central in computational mathematics, science, and engineering. The word "algorithm" stems from the name of Al-Khwarizmi (c. 780–c. 850), born in Khwarizm.[1] Translations of the works of this great scholar brought the decimal positional number system to the Western world, and the word "algorithm" was meant to indicate a technique with numerals. In the Ancient World, the development of algorithms was an important pillar of mathematical applications and was initially stimulated by geometrical considerations. Babylonian and Indian mathematicians already knew algorithms for approximating the area of a given circle. In the Ancient Greece, Antiphon the Sophist and Bryson of Heraclea in the 5th century BC were probably the first who developed an algorithm for calculating π by inscribing and then circumscribing a polygon around a circle, and calculating the polygons' areas. This methodology, known as the method of exhaustion, was a precursor to the integral calculus, and both Eudoxus of Cnidus and later Archimedes of Syracuse greatly contributed to its further development and applications. The development of number theory has also been a natural, rich source of algorithms. Early examples include Euclid's algorithm to determine the greatest common divisor of two integers, presented in his book the Elements. Later, the development of algorithms for finding solutions to mathematical models based on linear and non-linear equations, optimization and control problems has been dotted with the names of many outstanding scientists and mathematicians, including R. Descartes, I. Newton, L. Euler, C.F. Gauss, L. Kantorovich, G.B. Dantzig, J. von Neumann, L. Pontryagin, R. Bellman, and many others. While algorithms have always been central in the develop-

[1] Currently Khiva, Uzbekistan; geographically, Khwarizm or Chorasmia is a large oasis region on the Amu Darya river delta in western Central Asia.

ment of mathematical sciences, their importance has been substantially magnified with the advent of computers.

The speed of carrying out arithmetic and logical operations has increased drastically and continues to grow, leading to the ever increasing productivity in information processing and intellectual performance, not available to humans before. As a double-edge sword, this new power has to be used wisely and with caution. At the same time, it allows us to face new challenges in science, technology, and society. The process has led to an unprecedented boost in the development of new mathematical algorithms for the needs of computational science which has transformed the entire landscape of human activities, including scientific research, engineering design, production technologies, and education [2]. Through mathematical modeling, mathematics has become one of the major beneficiaries of this process, expanding its methods and tools to new areas. With the help of computational science, mathematical models now provide a stronger and more visible link between mathematics and the outside world, simultaneously strengthening links between different areas within mathematics itself.

2 Mathematical Modeling and Computational Experiments

The first step in mathematical modeling has always been to construct/derive a mathematical model. When studying phenomena, processes or systems in sciences and engineering or other areas of human endeavor, many mathematical models that are being constructed can be linked to the observable reality. In such cases, mathematical modeling can often complement natural experiments. The quality of mathematical models in such cases depends on the agreement between the results of mathematical modeling and natural experiments (experimental measurements). Therefore, the process of mathematical modeling is often a driving source for the development of new, better models, as well as for the development of hierarchies of mathematical models of different complexity (e.g., progressively more complex coupled models). This process helps determine the range of applicability of models and their possible simplifications. Based on such hierarchies, mathematical models can assist in explaining the behavior of a system under different conditions and the interaction of different system components. Notwithstanding that models for the same system can involve a range of mathematical structures and can be formalized with various mathematical tools (equation-based models, graphs, logical and game theoretic models, etc.).

Traditionally, the analysis of mathematical models has been based on a simplification of the model and some a priori assumptions made in such a way that something can be said analytically. Frequently, however, such simplifications lead to unrealistic assumptions and, as a result, to a large deviation from the reality of the phenomenon, process or system that is being described by the original model. The modern development of science and engineering convincingly demonstrates that the class of the models amenable to such simplifications, while keeping assumptions

realistic, is strikingly smaller compared to the general class of mathematical models that are at the forefront of modern applied mathematics, science, and engineering. As a rule, the problems in the latter class have to be treated numerically. Numerical methods for solving such problems require, in their turn, the development of efficient algorithms which are presented by sequences of mathematical/computational operations that would lead to a solution of the problem, often in the limit. Once such algorithms are implemented (programmed) computationally, we can run the model multiple times under varying conditions, helping us to answer outstanding questions quicker, more efficient, and providing us an option to improve the model when necessary. The triad "model-algorithm-implementation" is at the heart of mathematical modeling that leads to our ability to carry out *computational experiments* which become a pervasive and powerful theoretical tool in many areas of mathematical applications. Such areas embrace, for example, nonlinear phenomena in sciences and engineering, various spatio-temporal interactions, as well as complex systems studies, including the systems whose dynamics is only partially observable and systems with incomplete information.

While in the past, scientists, mathematicians, and engineers could use only simplified mathematical models to make them amenable to analytical treatments, the situation has now fundamentally changed. Indeed,

- we can now carry out computational experiments with more refined models, to solve them with more efficient methods, and to obtain more accurate results, unachievable with analytical techniques;
- we can carry out computational experiments in cases when natural experiments are impossible or difficult;
- with appropriate validation and verification procedures, we can provide reliable information more quickly and with less expense compared to natural experiments.

Computational experiments have become an intrinsic part of modern science and engineering and one of the essential tools in scientific discovery. Such experiments, sometimes referred to as "in silico", cover a wide spectrum of new computer-assisted disciplines such as computational physics, computational chemistry, computational materials science, computational biology, computational engineering, etc. In all these disciplines, we replace a traditional mathematical model by its computer equivalent, a computational model. Via computational experiments and theoretical analysis, we can assess the limits of applicability of mathematical models which can help set up natural experiments (sometimes referred to as "in vivo") and predict their results. This a two-way interaction. Indeed, most natural experiments are supported mathematically (via optimization of experiments, inverse problems, problems of identification, etc.). Not only can computational experiments be applied to traditionally applied areas of mathematical sciences mentioned above, but they can also be applied to their theoretical areas, and computational number theory, e.g., provides many beautiful examples. Even if an analytical solution to the original mathematical model is not known, based on a mathematical algorithm, a computer simulation can find an approximate solution. Computational experiments have promoted understanding in the society that new technologies developed with computational science and engineering tools are undoubtedly mathematics-based technologies.

An important feature of the current situation lies also in the fact that mathematical modeling becomes increasingly multidisciplinary. Computational science and engineering extend the applications of mathematics to new areas, reaffirming the versatility and ubiquitous nature of mathematical modeling. This provides new challenging problems in applied mathematics. At the same time, the development and implementation of new algorithms in these areas is closely connected with the development of information technology. Indeed, it is hard to overestimate the importance of the analysis of algorithms for emerging computer architecture. These two-way interactions, between information technology on the one hand and mathematical modeling and computational science on the other, continue to have a substantial impact on the analysis and *predictive capabilities* of mathematical models.

3 What Is Next

Computational experiments and the development of new methods and algorithms to support them help reveal some interconnected challenges in applied mathematical and computational sciences. Many systems that we encounter in nature, as well as man-made systems, are intrinsically inter-connected (coupled) with their parts interacting in non-trivial dynamic manner. Such systems require innovative mathematical and computational approaches, and their analysis demands fostering cross-fertilization between different disciplines and mathematics, as well as between different areas within mathematics. The areas where such systems frequently appear include, but not limited to:

- life sciences,
- earth, climate, environmental, and sustainability sciences,
- high energy and nuclear physics, fusion and energy problems,
- materials science and chemistry,
- cosmology, astrophysics, and celestial mechanics,
- aerospace sciences and new technologies,
- and others.

Facing many problems in these areas would require the development of new efficient algorithms, higher-order methods, systems-science-based approaches, and building multidisciplinary capabilities. New advances in mathematical modeling of intrinsically interconnected (coupled) systems will call upon an integration of a variety of methods and tools across multiple disciplines, including different areas of mathematics. Indeed, some such systems, especially those that are inherently stochastic and those with incomplete or partially observable dynamic behaviors, would require improved predictive capabilities of mathematical models describing them. The models in such cases should be able to predict the probability of an outcome in the best possible way. The success of predictive mathematical modeling is dependent also on further advances in information sciences, and the development of statistical and probabilistic methods, methods for uncertainty quantification, and machine learning

techniques. Furthermore, since coupled systems are often interacting on a range of spatio-temporal scales, the development of predictive multiscale models and multiscale algorithms (in particular, for filtering and data assimilation and various types of parametrization) is becoming an important new avenue of research.

Probabilistic and applied statistics approaches will also become increasingly important in the design, analysis, and optimization of computational experiments. This is due to three main sources of uncertainty caused by (a) parameters with uncertain values, (b) uncertainty in the model as a representation of the underlying phenomenon, process, or system, and (c) uncertainty in collecting/processing/measurements of data for model calibration. In quantifying and mitigating these uncertainties in mathematical models, new developments in novel statistical-stochastic methods and methods for efficient integration of data and simulation are expected.

With an unprecedented growth of the range of application areas where mathematical modeling and computational science play an increasingly important role, demands for professions with adequate mathematical skills should increase. This raises a number of challenges in education of such future professionals at the university level, a topic which lies beyond the scope of this introductory chapter.

4 What This Book Is About

The rest of the book consists of eight state-of-the-art chapters on important recent advances in applied mathematics, modeling and computational science. These chapters are based on selected invited contributions from leading specialists from North America, Australia, and Europe.

The vast range of research areas within these three interconnected fields has led us to a selection of topics covered in this book. Hence, the chapters that follow provide a selective, but at the same time broad spectrum of methods and tools that are at the forefront of modern developments in applied mathematics, mathematical modeling and computational science. Each chapter has both theoretical and applied parts, making the book a comprehensive combination of mathematical theories, model derivations, and development of efficient numerical procedures applicable to a wide range of applications. The book contains also a selection of 12 computer codes, written in MATLAB in a simple and transparent way to allow the reader immediate access to the idea, as well as a variety of state-of-the-art numerical algorithms with applications in fields ranging from population dynamics and medicine to cryptography and fire propagation.

- *Dynamic blocking problems and their applications.* The theory of dynamic blocking problems has a number of various applications such as the propagation of a wild fire in a forest, various traffic problems, the spatial spreading of a contaminating agent, etc. A comprehensive survey of this theory is given by A. Bressan (Penn State University, USA). Mathematical models of dynamic blocking problems can often be cast in the form of differential inclusions describing the growth

of sets in the plane. In order to restrain the expansion of such sets, it is usually assumed that barriers can be constructed. From both theoretical and practical points of view, it is important to be able to understand whether the growth of these sets can be blocked. Therefore, the author reviews the results known up to date on the existence and non-existence of blocking strategies. Another important issue addressed in this chapter is how to find the optimal location of the barriers, minimizing a cost criterion. The author provides all details on both necessary and sufficient conditions for optimality. He goes on to the description of a numerical algorithm for the computation of optimal barriers. The chapter is concluded with a survey of open problems in this new challenging area of research.

- *High order accurate numerical boundary conditions for solving hyperbolic equations.* The challenges of boundary treatment in mathematical models based on partial differential equations are well known in computational science and engineering. They are intrinsic to various numerical techniques applied for solving associated problems, including finite difference schemes and finite element type methods. A review and discussions on new developments in meeting these challenges are given in the current chapter by S. Tan and C.-W. Shu (Brown University, USA). They focus on the Inverse Lax-Wendroff (ILW) procedure and consider high order accurate finite difference methods for solving hyperbolic conservation laws involving complex static and moving geometries. Their methodology is applied on Cartesian grids, while the physical domain can be arbitrarily shaped. Some of the challenges that need to be address in this case include typically wide stencils for high order schemes and the fact that the grid lines may not necessarily coincide with the physical boundary. The authors give a very thorough description of their main methodology based on the ILW procedure for inflow boundary conditions and on a robust and high order accurate extrapolation for outflow boundary conditions. They provide details of the stability of the method, followed by a series of examples that include also problems involving interactions between shock waves and rigid boundaries.
- *Number theoretic and cryptographic issues of elliptic curves over finite fields.* Modern number theory has a rich history that can be traced back to the dawn of human civilization. The interest to elliptic curves is of more recent origin and one of the main reasons for the great interest to them for more than a century lies with the fact that they have a naturally associated with them group structure. A self-contained survey demonstrating strong links between many problems on elliptic curves over finite fields and a myriad of problems in analytic, algebraic, and computational number theory is given by I.E. Shparlinski (Macquarie University, Australia). The range of such problems has expanded further after the invention of elliptic curve cryptography, opening up many new research directions and applications. The author demonstrates that number theory remains central to both, a classical theory of elliptic curves and application driven research in these areas. After giving a detailed account of the structure of elliptic curves over finite fields, cryptographic applications of such curves are considered. The topics include, but not limited to, pairing friendly curves and pseudorandom number generators. The survey can serve as a reference for theoreticians to this new important direction,

as well as a source of new problems. It can be useful to the cryptographers and other practitioners working in applications of the modern number theory. The self-contained nature of the survey makes it also accessible to graduate students and those who are interested in learning about state-of-the-art developments in these areas.

- *Random matrix theory and its innovative applications.* Random matrices (matrix-valued random variables) are found an increasing number of applications in science and engineering. This process, starting at the beginning of the 20th century, has led to what we now know as Random Matrix Theory (RMT). Its many applications are often surprising and innovative. In number theory (see our previous chapter), for example, the distribution of zeros of the Riemann zeta function can be modelled by the distribution of eigenvalues of certain random matrices. Many novel examples of RMT applications are given in this chapter by A. Edelman (Massachusetts Institute of Technology, USA) and Y. Wang (Tufts University, USA). Such examples are ranging from health sciences to many engineering problems where the limiting densities are often needed to indicate the cutoff between "noise" and "signal". The chapter explains the theory behind these examples and associated applications. The reader can find here the Hermite and Laguerre ensembles, the description of four famous laws used in RMT as they govern the limiting eigenvalue distributions of random matrices. The details of matrix reductions and an overview of how these reductions can be used for efficient computation are also given. The chapter contains 12 codes, written in MATLAB, so that the readers can immediately embark on some of the RMT ideas in their specific areas of applications.
- *Energy stable weighted essentially non-oscillatory finite-difference schemes and their applications.* Complex scientific and engineering problems exhibiting discontinuities are ubiquitous in applications. For high fidelity simulations of such problems high-order weighted essentially non-oscillatory (WENO) schemes are often methods of choice. Near domain boundaries, however, these methods face serious challenges. Boundary closures of high order have recently been developed to overcome existing problems, as these closures, along with near-wall biasing mechanics, complement the periodic domain energy stable WENO methods (ESWENO). The chapter by M.H. Carpenter, T.C. Fisher (both NASA Langley Research Center, USA), and N.K. Yamaleev (North Carolina A&T University, USA) provides a summary of the recent developments in this field. It is demonstrated that a novel set of nonuniform flux interpolation points is necessary near the boundaries in order to simultaneously achieve accuracy, the summation-by-part convention, and WENO stencil biasing mechanics. One of the motivations for this work has been provided by simulations of sound that is generated by a shock-vortex interaction. Practical examples have been given to complement the theoretical part and to test the developed methodology. These include unsteady propagation of a two-dimensional Euler vortex and unsteady convection-diffusion-reaction of a supersonic hydrogen-air mixing layer, simulated by solving the two-dimensional Navier-Stokes equations. Detailed instructions on and numerical algorithms for the implementation of the new ESWENO schemes have also been provided.

- *Multiscale methods coupling network and continuum models in porous media.* Modeling transport in porous media applications is know to be a difficult task. This is especially true, for example, for two-phase flow simulations in porous media where nonlinearity and heterogeneity of small scale processes dictate the large scale flow behavior. This problem is addressed in the present chapter by J. Chu, B. Engquist, M. Prodanovic, and R. Tsai (all at the University of Texas at Austin, USA). They present a new numerical multiscale methodology for coupling a conservation law for mass at the continuum scale with a discrete network model that describes the pore scale flow in the porous medium. The developed methodology is based on the heterogeneous multiscale method (HMM). The HMM starts with an incomplete macroscopic model for macro variables on the macrogrid covering the full domain. The missing quantities and data in the macroscopic model are then obtained by solving an accurate microscale model locally over small domains. In the case at hand, the developed coupling method for the pressure equation uses local simulations on small sampled network domains at the pore scale to evaluate the continuum equation and thus solve for the pressure in the domain. The developed numerical algorithm for dynamic two-phase flows and its convergence are discussed in detail. The algorithm is given as a step-by-step procedure that is easy to implement for the interested reader. Computational experiments on both, single-phase flows with nonlinear flux-pressure dependence and the full two-phase flow, are presented.
- *Statistical geometry and topology applications in life sciences.* The use of concepts and methods of geometry, topology, and graph theory becomes increasingly important in life science applications. These applications include, but not limited to, the analysis of biological networks. A new method of characterizing tree networks based on structural triangulation of that network is developed in the present chapter by R.-K. Seong (Imperial, UK), P. Getreuer (Yale, USA), Y. Li, T. Girardi, C.M. Salafia (all Placental Analytics LLC, USA), and D.D. Vvedensky (Imperial, UK). A specific biological network is in the major focus in this chapter, namely the vasculature on the chorionic plate of a human placenta. For a graph-theoretical analysis, the vasculature is represented as edges and vertices. The developed method makes use of a triangulation of the tree network which enables one to formulate a partition function. Hence, the method can use then thermodynamic functions known from statistical mechanics as simple measures of the weighted topology of the corresponding tree network. The systematic variation of the introduced weights allows an examination of the development of the network. Several illustrative examples explaining the methodology and the algorithm behind are discussed. Finally, the developed methodology has been applied for the analysis of the arterial and venous vasculature of the chorionic plate of a variety of human placentas, and attempts have been made to examine the extent to which the entropy function is correlated to the infant birthweight with the sample set.
- *Optimal control applications with discrete and continuous features.* There is a wide range of applications where optimal control problems with discrete and continuous features arise naturally. This includes science and engineering, as well as

life science applications. A series of examples from such applications have been provided by S. Lenhart (University of Tennessee, USA), E. Bodine (Rhodes College, TN, USA), P. Zhong (Rutgers University, USA), and H.R. Joshi (Xavier University, OH, USA). Starting with several illustrative examples for optimal control with ordinary differential equations, the authors have explained in detail how the bang-bang and singular controls can be handled. Their examples include robotics applications where a mobile robot with one or more steerable drive wheels that steer together is considered. An example from population biology models species augmentation where two populations of the same species are described with a target endangered population and a reserve population. When the underlying systems are described by ODEs, problems which are linear in the control and have discrete values for the optimal control are emphasized. An extension to integro-difference models that are discrete in time and continuous in space is also given. This extension is illustrated with the optimal pest control problem where the underlying theory for characterization of an optimal control and necessary conditions are also discussed. Computational experiments for each of the examples described above complement the theory presented in this chapter. Details of numerical algorithms developed for such experiments are also given.

5 Concluding Remarks

An important pre-requisite of success of mathematical modeling and computational science as a driving force of scientific discovery and innovative, mathematics-based technologies is a closer collaboration of mathematicians with other disciplines. Only in this way, mathematical theories can evolve in a dynamic, sustainable manner as an essential part of human knowledge, and new mathematical models can stimulate new theoretical discoveries and new experimental findings.

At present, theoretical and experimental sciences, taken separately, are lacking systems-science-based approaches in studying phenomena, processes, and systems. Mathematical modeling and computational experiments can provide a missing link by complementing these two fundamental ways of human activities. They allow a two-way interaction between theory and natural experiment virtually in all areas where mathematical models can be constructed. They provide also a missing link in those cases where natural experiment is difficult or impossible to set up. This extends further the domain of mathematical modeling and mathematics-based technologies making them a decisive factor in sustainable socio-economic development and scientific-technological progress of our society.

References

1. Poincare, H.: L'avenir des mathématiques ("The Future of Mathematics"). Revue générale des sciences pures et appliquées, **19 (23)**, 930–939 (1908).
2. Xie, C., et al.: Computational Experiments for Science Education. Science, **332 (6037)**, 1516–1517 (2011).

Dynamic Blocking Problems for a Model of Fire Propagation

Alberto Bressan

Abstract This paper contains a survey of recent work on a class of dynamic blocking problems. The basic model consists of a differential inclusion describing the growth of a set in the plane. To restrain its expansion, it is assumed that barriers can be constructed, in real time. Here the issues of major interest are: (i) whether the growth of the set can be eventually blocked, and (ii) what is the optimal location of the barriers, minimizing a cost criterion. After introducing the basic definitions and concepts, the paper reviews various results on the existence or non-existence of blocking strategies. A theorem on the existence of an optimal strategy is then recalled, together with various necessary conditions for optimality. Sufficient conditions for optimality and a numerical algorithm for the computation of optimal barriers are also discussed, together with several open problems.

1 Introduction

Consider a set in the plane, which expands as time increases. Assume that its growth can be restrained by constructing barriers, in real time. In this setting, a natural problem is whether one can completely block the growth of the set, constructing barriers all around. In addition, given a cost criterion, it is also of interest to determine the optimal location of these barriers.

Dynamic blocking problems of this kind were first considered in [5], motivated by the optimal control of wild fires. Let $R(t) \subset \mathbb{R}^2$ denote the region burned by the fire at time t. To restrict its growth, assume that a barrier can be constructed, along a one-dimensional curve. We shall denote by $\gamma(t) \subset \mathbb{R}^2$ the portion of this barrier constructed within time t. In the case of a forest fire, one may think of a thin strip of land which is either soaked with water poured from an airplane or a helicopter, or cleared from all vegetation using a bulldozer, or sprayed with fire extinguisher by a team of firemen. In any case, this will prevent the fire from crossing that particular strip of land. In connection with this model, it is then natural to ask whether it is

A. Bressan (✉)
Department of Mathematics, Penn State University, University Park, PA 16802, USA
e-mail: bressan@math.psu.edu

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5_2,

possible to completely stop the fire, and what is the best strategy to achieve this goal, minimizing the total value of the burned region.

Aim of this paper is to survey the main concepts and results in the theory of dynamic blocking problems, developed in [5–7, 9–12, 15, 23], and discuss various open questions.

After a precise description of the mathematical model, in Sect. 2 we introduce an equivalent way to formulate both the blocking problem and the optimization problem. In Sect. 3 we recall the main results on the existence or non-existence of blocking strategies. The remainder of the paper is concerned with optimal strategies. The basic existence theorem [7] is presented in Sect. 4. Section 5 reviews the classification of arcs in an optimal barrier and the necessary conditions for optimality proved in [5, 11, 23], while Sect. 6 describes a recent result concerning sufficient conditions for optimality. A numerical algorithm for computing the optimal barrier, introduced in [9], is here presented in Sect. 7. Finally, in Sect. 8 we discuss several remaining open problems.

1.1 A Model for Fire Propagation

Several models for fire propagation have been proposed in the literature, see for example [17–19, 21]. In our model, the set $R(t) \subset \mathbb{R}^2$ burned by the fire up to time t is described as the reachable set for a differential inclusion. More precisely, consider the Cauchy problem

$$\dot{x} \in F(x), \qquad x(0) \in R_0, \tag{1}$$

where the upper dot denotes a derivative w.r.t. time. Here R_0 describes the region where the fire is initially burning at time $t = 0$, while $F(x)$ is a set of propagation velocities.

If barriers are not present, for each $t \geq 0$ the set $R(t)$ reached by the fire is defined as (see Fig. 1)

$$\begin{aligned} R(t) \doteq \big\{ & x(t);\, x(\cdot) \text{ is absolutely continuous, } x(0) \in R_0, \\ & \dot{x}(\tau) \in F\big(x(\tau)\big) \text{ for a.e. } \tau \in [0,t] \big\}. \end{aligned} \tag{2}$$

We shall always assume that the initial set $R_0 \subset \mathbb{R}^2$ is nonempty and bounded. Moreover, we assume that $F : \mathbb{R}^2 \mapsto \mathbb{R}^2$ is a Lipschitz continuous multifunction whose values $F(x)$ are compact, convex sets containing the origin. Clearly, this implies

$$R(t_1) \subseteq R(t_2) \quad \text{whenever } t_1 < t_2. \tag{3}$$

According to (2), the propagation speed of a fire front in the normal direction is computed by

$$h(x) = \max_{v \in F(x)} \big\langle \mathbf{n}(x), v \big\rangle. \tag{4}$$

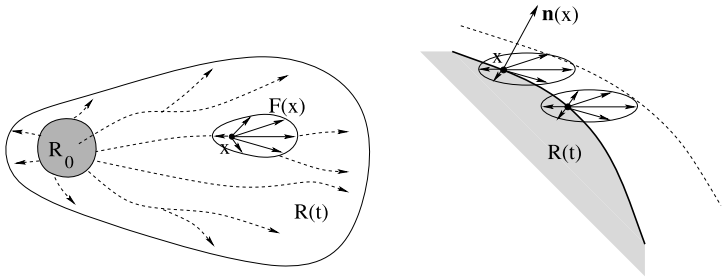


Fig. 1 *Left*: the region $R(t)$ burned by the fire at time $t > 0$ is described as the set reached by trajectories of the differential inclusion (1). *Right*: according to this model, the fire front propagates in the normal direction with speed given by (4)

Here x is any point along the boundary $\partial R(t)$ of the set burned up to time t, while $\mathbf{n}(x)$ denotes the unit outer normal vector to the boundary, at the point x. By $\langle \cdot, \cdot \rangle$ we denote the Euclidean inner product in $\mathbb{R}^2$.

An alternative way to describe this same model of fire propagation relies on the solution of a Hamilton-Jacobi (H-J) equation [17]. For each $x \in \mathbb{R}^2$, call

$$T(x) \doteq \inf\big\{t \geq 0; x \in R(t)\big\} \tag{5}$$

the minimum time taken by the fire to reach the point x, starting from the initial set R_0. The function $T(\cdot)$ can now be computed by solving the nonlinear PDE

$$H\big(x, \nabla T(x)\big) = 0, \qquad H(x, p) \doteq \max_{v \in F(x)} \langle p, v \rangle - 1, \tag{6}$$

with boundary data

$$T(x) = 0 \quad \text{for } x \in R_0. \tag{7}$$

The level set $\{x; T(x) = t\}$ describes the position of the fire front at time $t > 0$. We remark that, in general, the solution of (6)–(7) may not be smooth. In this case, the H-J equation (6) must be suitably interpreted in a viscosity sense [4].

While the representation based on differential inclusions is useful for theoretical analysis, the H-J equation leads to more efficient computational algorithms, based on the level set method [17, 20].

1.2 Barriers

We assume that the spreading of the fire can be controlled by constructing barriers, in real time. Intuitively, we think of a barrier as a curve (or a family of curves) in the plane, which the fire cannot cross. Since the wall is constructed in real time, simultaneously with fire propagation, a restriction on its length must be imposed. Calling $\gamma(t)$ the portion of the curve constructed up to time t, if σ is the speed at which the wall is constructed we thus have the constraint

$$\big[\text{length of } \gamma(t)\big] \leq \sigma t \quad \text{for every } t \geq 0. \tag{8}$$

A more general situation can be considered. Indeed, the construction of the barrier can be faster in certain places than others. For example, if water is sprayed by an helicopter on top of the fire, this operation can be carried out more quickly in areas close to a lake or a water reservoir. To model this fact, we consider a continuous, strictly positive function $\psi : \mathbb{R}^2 \mapsto \mathbb{R}_+$. Calling $\gamma(t) \subset \mathbb{R}^2$ the portion of the wall constructed within time $t \geq 0$, we make the following assumptions:

(H1) For every $0 \leq t_1 \leq t_2$ one has $\gamma(t_1) \subseteq \gamma(t_2)$.
(H2) Each $\gamma(t)$ is a rectifiable curve whose length satisfies

$$\int_{\gamma(t)} \psi \, dm_1 \leq t \quad \text{for every } t \geq 0. \tag{9}$$

The first assumption states that, after being constructed a barrier cannot be destroyed, or moved to another place. For a precise mathematical definition of *rectifiable set* we refer to [1]. Roughly speaking, a rectifiable set is the most general type of one-dimensional set for which a concept of *length* can be introduced. In the integral formula (9), m_1 denotes the one-dimensional Hausdorff measure, normalized so that $m_1(\Gamma)$ yields the usual length of a smooth curve Γ. Notice that $1/\psi(x)$ is the speed at which the wall can be constructed, at the location x. In particular, if $\psi(x) \equiv \sigma^{-1}$ is constant, then the constraint (9) reduces to (8). In the above model, we assume that the construction speed $1/\psi(x)$ depends only on the spatial location. It would be of interest to consider a model where this speed depends also on time. This more general case remains yet to be studied.

Remark 1 In general, the curve $\gamma(t)$ need not be connected. For example, it may be the union of two separate barriers, produced by two teams of firemen working independently at different locations.

A blocking strategy $t \mapsto \gamma(t)$ satisfying (H1)–(H2) will be called an *admissible strategy*. In addition, we say that the strategy γ is *complete* if it satisfies

(H3) For every $t \geq 0$ there holds

$$\int_{\gamma(t)} \psi \, dm_1 = t, \qquad \gamma(t) = \bigcap_{s>t} \gamma(s). \tag{10}$$

Moreover, if $\gamma(t)$ has positive upper density at a point x, i.e. if

$$\limsup_{r \to 0+} \frac{m_1(B(x,r) \cap \gamma(t))}{r} > 0,$$

then $x \in \gamma(t)$.

Here $B(x,r)$ denotes the open ball centered at x with radius r. As proved in [7], for every admissible strategy $t \mapsto \gamma(t)$ one can construct a second admissible strategy $t \mapsto \tilde{\gamma}(t) \supseteq \gamma(t)$, which is complete.

Remark 2 The assumption (H3) provides some weak regularity for the sets $\gamma(t)$. In general, one cannot require that these sets be closed, because the closure of a

set Γ can have much bigger total length. For example, consider a rectifiable set $\Gamma \subset \mathbb{R}^2$ containing all points with rational coordinates, and having total length $m_1(\Gamma) \leq 1$. In this case, the closure $\overline{\Gamma}$ is the entire plane, with one-dimensional measure $m_1(\overline{\Gamma}) = m_1(\mathbb{R}^2) = \infty$.

When a barrier is being constructed, the set reached by the fire is reduced. This leads to the definition of the new reachable set

$$R^\gamma(t) \doteq \big\{x(t);\, x(\cdot) \text{ absolutely continuous}, x(0) \in R_0,$$
$$\dot{x}(\tau) \in F(x(\tau)) \text{ for a.e. } \tau \in [0,t], x(\tau) \notin \gamma(\tau) \text{ for all } \tau \in [0,t]\big\}. \quad (11)$$

According to (11), at any time τ the fire cannot cross the portion $\gamma(\tau)$ of the wall which is already in place. Clearly, the burned set will thus be smaller: $R^\gamma(t) \subseteq R(t)$ for every $t \geq 0$.

The alternative description of the reachable sets based on the PDE (6) can also be implemented in this more general case, in the presence of barriers. A characterization of the minimum time function $T(\cdot)$ as the solution to a Hamilton-Jacobi equation with obstacles was recently proved in [15].

1.3 Blocking and Optimization Problems

In the above setting, two natural problems arise. The first one is concerned with dynamic control, the second with optimization.

(BP1) Blocking Problem: Given a multifunction F, a construction speed $1/\psi$ and a bounded initial set R_0, decide whether there exists an admissible strategy $t \mapsto \gamma(t)$ such that the corresponding reachable sets $R^\gamma(t)$ remain uniformly bounded for all $t \geq 0$.

In other words, we ask whether it possible to construct a barrier $\gamma(\cdot)$ such that

$$R^\gamma(t) \subseteq B_r \quad \text{for all } t > 0 \quad (12)$$

for some fixed ball B_r centered at the origin with radius r. If this is the case, we say that a blocking strategy exists. Here one should keep in mind that the barrier must be constructed in real time, simultaneously with the advancement of the fire front (Fig. 2). Clearly, a blocking strategy can exist only if the construction speed of the barrier is sufficiently fast, compared with the speed at which fire propagates.

To describe an optimization problem, one needs to introduce a cost functional. This should take into account:

- The value of the area burned by the fire.
- The cost of building the barrier.

Following [5], we consider two continuous, non-negative functions $\alpha, \beta : \mathbb{R}^2 \mapsto \mathbb{R}_+$ and define the functional

$$J(\gamma) = \int_{R^\gamma_\infty} \alpha \, dm_2 + \int_{\gamma_\infty} \beta \, dm_1. \quad (13)$$

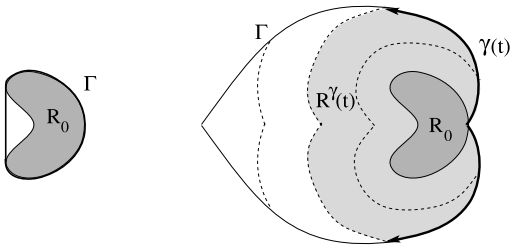


Fig. 2 *Left*: a "static" blocking problem. Here Γ is the curve of minimum length that entirely surrounds the set R_0. *Right*: a "dynamic" blocking problem, where the barrier Γ is constructed at the same time as the set R^γ burned by the fire expands. Here the *thick curve* $\gamma(t)$ denotes the portion of the barrier constructed up to time t, while the *shaded region* $R^\gamma(t)$ denotes the set burned by the fire, at time t

Here m_2 denotes the two-dimensional Lebesgue measure, while m_1 is the one-dimensional Hausdorff measure. Moreover, the domains of integration R^γ_∞, γ_∞ are defined respectively as

$$R^\gamma_\infty \doteq \bigcup_{t\geq 0} R^\gamma(t), \qquad \gamma_\infty \doteq \bigcup_{t\geq 0} \gamma(t). \tag{14}$$

In our model, $\alpha(x)$ is the value of a unit area of land at the point x, while $\beta(x)$ is the cost of building a unit length of wall at the point x. The set R^γ_∞ describes the entire region burned by the fire, while γ_∞ is the entire barrier. The first integral on the right hand side of (13) thus accounts for the value of the land burned by the fire, while the second integral yields the total cost of constructing the barrier. This leads to

(OP1) Optimization Problem: Find an admissible strategy $t \mapsto \gamma(t)$ for which the corresponding functional $J(\gamma)$ at (13) attains its minimum value.

In the remainder of this paper we discuss several recent results and open questions, in connection with the above blocking and optimization problems. In particular, the following issues are of interest:

- existence or non-existence of blocking strategies,
- existence of optimal strategies,
- necessary conditions for optimality,
- sufficient conditions for optimality,
- regularity of optimal barriers,
- numerical computation of optimal barriers.

For future reference, we list a set of assumptions which will be used in the remainder of the paper.

(A1) The initial set R_0 is nonempty, open and bounded. Its boundary satisfies $m_2(\partial R_0) = 0$.

(A2) The multifunction F is Lipschitz continuous w.r.t. the Hausdorff distance. For each $x \in \mathbb{R}^2$ the set $F(x)$ is nonempty, closed and convex and contains the origin in its interior.

(A3) For every $x \in \mathbb{R}^2$ one has $\alpha(x) \geq 0$, $\beta(x) \geq \beta_0 > 0$, and $\psi(x) \geq \psi_0 > 0$. Moreover, α is locally integrable, while β and ψ are both lower semicontinuous.

We recall that the Hausdorff distance between two compact sets X, Y is defined as

$$d_H(X, Y) \doteq \max\left\{\max_{x\in X} d(x, Y), \max_{y\in Y} d(y, X)\right\},$$

where

$$d(x, Y) \doteq \inf_{y\in Y} d(x, y)$$

and $d(x, y) \doteq |x - y|$ is the Euclidean distance on $\mathbb{R}^2$. The multifunction F is Lipschitz continuous if there exists a constant L such that

$$d_H\big(F(x), F(y)\big) \leq L \cdot d(x, y),$$

for every couple of points x, y. For the basic theory of multifunctions and differential inclusions we refer to [3].

2 An Equivalent Formulation

In its original formulation [5], an admissible strategy was defined as a set-valued map $t \mapsto \gamma(t) \subset \mathbb{R}^2$. Indeed, for each $t \geq 0$ one needs to describe the portion of the wall constructed within time t. The subsequent paper [9] showed that the both the blocking problem and the optimization problem can be reformulated in a simpler way, where an admissible strategy is determined by one single rectifiable set $\Gamma \subset \mathbb{R}^2$. This approach is particularly useful for the numerical computation of optimal strategies. We review here the main ideas.

Consider a rectifiable set $\Gamma \subset \mathbb{R}^2$ which is *complete*, in the sense that it contains all of its points of positive upper density:

$$\limsup_{r\to 0+} \frac{m_1(B(x, r) \cap \Gamma)}{r} > 0 \quad \Longrightarrow \quad x \in \Gamma.$$

The set reached at time t by trajectories of the differential inclusion (1) without crossing Γ is then defined as

$$R^{\Gamma}(t) \doteq \big\{x(t);\, x(\cdot) \text{ absolutely continuous}, x(0) \in R_0,$$
$$\dot{x}(\tau) \in F(x(\tau)) \text{ for a.e. } \tau \in [0, t], x(\tau) \notin \Gamma \text{ for all } \tau \in [0, t]\big\}. \qquad (15)$$

Throughout the following, $\overline{S}$ will denote the closure of a set S. We say that the rectifiable set Γ is *admissible* in connection with the differential inclusion (1) and the bound on the construction speed (9) if

$$\int_{\Gamma\cap\overline{R^\Gamma(t)}} \psi\, dm_1 \leq t \quad \text{for all } t \geq 0. \tag{16}$$

Of course, this means that the strategy

$$t \mapsto \gamma(t) \doteq \Gamma \cap \overline{R^\Gamma(t)} \tag{17}$$

is admissible according to (9).

In analogy with (14), we denote by

$$R^\Gamma_\infty \doteq \bigcup_{t\geq 0} R^\Gamma(t) \tag{18}$$

the entire region burned by the fire. Both the blocking problem (BP1) and the optimization problem (OP1) can now be reformulated in a simpler way, involving one single barrier Γ.

(BP2) Blocking Problem: Find an admissible rectifiable set Γ such that the corresponding region R^Γ_∞ is bounded.

(OP2) Optimization Problem: Find an admissible rectifiable set $\Gamma \subset \mathbb{R}^2$ such that the cost

$$J(\Gamma) = \int_{R^\Gamma_\infty} \alpha\, dm_2 + \int_\Gamma \beta\, dm_1 \tag{19}$$

attains the minimum possible value.

As proved in [9], under the assumptions (A1)–(A3) the two formulations are equivalent. In particular, if $t \mapsto \gamma(t)$ is a complete, optimal strategy for (OP1), then the rectifiable set

$$\Gamma \doteq \left(\bigcup_{t\geq 0} \gamma(t)\right) \setminus \left(\bigcup_{t\geq 0} R^\gamma(t)\right) \tag{20}$$

is admissible and provides an optimal solution to the minimization problem (OP2). Viceversa, if the set Γ provides an optimal solution to (OP2), then the strategy $\gamma(\cdot)$ in (17) is optimal for (OP1).

Remark 3 For each $t \geq 0$, the set $\gamma(t)$ in (17) is the part of the wall Γ touched by the fire at time t. This is the portion that actually needs to be put in place within time t, in order to restrain the fire. The remaining portion $\Gamma \setminus \gamma(t)$ can be constructed at a later time. On the other hand, given a strategy $\gamma(\cdot)$, the set Γ in (20) consists of the "useful" part of all walls constructed by γ. Portions of a wall, which are constructed in a region already reached by the fire, are clearly useless.

Remark 4 By the assumption (A2), the fire propagates with positive speed in every direction. Hence, for a given initial domain R_0, the set R^Γ_∞ in (18) burned by the fire can be characterized as the union of all connected components of $\mathbb{R}^2 \setminus \Gamma$ which intersect R_0.

Given a barrier Γ, we can define the minimum time function as

$$T^{\Gamma}(x) \doteq \inf\{t; x \in \overline{R^{\Gamma}(t)}\}. \tag{21}$$

A characterization of this function as a solution of a H-J equation with obstacles was recently provided by De Lellis and Robyr in [15]. Before stating their result, we recall that a map $u : \mathbb{R}^2 \mapsto \mathbb{R}$ is in the space SBV of Special functions with Bounded Variation if

- its distributional derivative Du is a measure,
- decomposing this measure $Du = \nabla u + D^{jump}u + D^{Cantor}u$ as the sum of an absolutely continuous part (w.r.t. Lebesgue measure), a jump part, and a Cantor part, the last component vanishes: $D^{Cantor}u \equiv 0$.

A family $\mathcal{S}^{\Gamma}$ of subsolutions to the H-J equation (6) with Γ as an obstacle can now be defined as follows. For any $t \geq 0$, we shall denote by $u \wedge t$ the truncated function

$$(u \wedge t)(x) \doteq \min\{u(x), t\}. \tag{22}$$

Definition 1 A function $u : \mathbb{R}^2 \mapsto [0, \infty]$ is in the set $\mathcal{S}^{\Gamma}$ if

(i) For every $t \geq 0$ one has $(u \wedge t) \in SBV$, and the set of jump points $J_{u\wedge t} \subseteq \text{Supp}(D^{jump}u)$ is contained inside the barrier. Namely, $m_1(J_{u\wedge t} \setminus \Gamma) = 0$.
(ii) $u = 0$ on R_0 and $H(x, \nabla u(x)) \leq 0$ for a.e. x.

As proved in [15], the function T^{Γ} defined by (21) admits the following characterization:

Theorem 1 *Let the assumptions* (A1)–(A2) *hold. Then the minimum time function* T^{Γ} *is the unique maximal element of* $\mathcal{S}^{\Gamma}$.

3 Existence of Blocking Strategies

In this section we discuss the existence or nonexistence of an admissible strategy $\gamma(\cdot)$ which restrains the fire within a uniformly bounded region.

We recall that a blocking problem is specified by assigning

- the multifunction $x \mapsto F(x)$ describing the propagation velocity of the fire,
- the set R_0, describing the initial location of the fire,
- the function $x \mapsto \psi(x)$ determining the speed $\sigma = 1/\psi$ at which the barrier can be constructed.

For different data, a simple but useful comparison result holds.

Lemma 1 *Consider two blocking problems, the first with data* (F, R_0, ψ), *the second with data* $(\widetilde{F}, \widetilde{R}_0, \tilde{\psi})$. *Assume that* $R_0 \subseteq \widetilde{R}_0$ *and*

$$F(x) \subseteq \widetilde{F}(x), \qquad \psi(x) \leq \tilde{\psi}(x) \quad \textit{for all } x \in \mathbb{R}^2. \tag{23}$$

If a blocking strategy for the second problem exists, then the first problem admits a blocking strategy as well.

Proof Indeed, let $t \mapsto \tilde\gamma(t)$ be an admissible strategy for the second problem, such that the corresponding reachable sets satisfy $\widetilde R^{\tilde\gamma}(t) \subseteq B_r$ for some fixed radius r and all $t \geq 0$. Since

$$\int_{\tilde\gamma(t)} \psi\, dm_1 \leq \int_{\tilde\gamma(t)} \tilde\psi\, dm_1 \leq t,$$

the strategy $\tilde\gamma(\cdot)$ is admissible for the first problem as well. Call $R^{\tilde\gamma}$, $\widetilde R^{\tilde\gamma}$ the corresponding sets reached by the fire in the first and second problem, respectively. By the assumptions,

$$R^{\tilde\gamma}(t) \subseteq \widetilde R^{\tilde\gamma}(t) \subseteq B_r \quad \text{for all } t \geq 0.$$

□

3.1 The Isotropic Case

We begin by discussing the isotropic case, where the fire propagates with unit speed in all directions, while the barrier is constructed at a constant speed $\sigma > 0$. In other words, we assume that $F(x) = \overline{B}_1$ is the closed disc centered at the origin with radius 1, and the constraint (8) holds.

We observe that, in this isotropic case, the family of solutions of (1) is invariant under rotations and translations. It is also invariant under a group of rescaling transformations. Namely, consider an initial set R_0 and an admissible strategy $t \mapsto \gamma(t)$, and let $R^\gamma(t)$ be the corresponding reachable sets, defined at (11). Given any $\lambda > 0$, define the rescaled barriers

$$\tilde\gamma(t) \doteq \lambda\gamma(t/\lambda)$$

and the initial set $\widetilde R_0 \doteq \lambda R_0$. It is now easy to check that the blocking strategy $\tilde\gamma(\cdot)$ is also admissible, and the corresponding sets reached by the fire are given by

$$R^{\tilde\gamma}(t) = \lambda R^\gamma(t/\lambda). \tag{24}$$

Combining (24) with the comparison result stated in Lemma 1, one can prove that the solvability of the blocking problem depends only on the speed σ, and not on the initial set R_0.

Lemma 2 *Let a construction speed $\sigma > 0$ be given. If there exists a (nonempty) bounded open set R_0^* for which the blocking problem can be solved, then a blocking strategy exists for every bounded set R_0.*

Proof Indeed, by a rescaling followed by a translation, every bounded set R_0 can be mapped into a subset of R_0^*. The result thus follows from Lemma 1. □

To understand the solvability of the blocking problem in the isotropic case, it thus suffices to study the case where

$$F(x) \equiv \overline{B}_1, \qquad R_0 = B_1, \qquad \psi(x) = \frac{1}{\sigma}, \tag{25}$$

where B_1 is the open disc centered at the origin with unit radius. By a comparison argument it is clear that there must be a constant $\sigma^* > 0$ such that

– for $\sigma > \sigma^*$ a blocking strategy exists,
– for $\sigma < \sigma^*$ blocking strategy does not exists.

At the present date, the exact value of this constant σ^* is not known. The analysis in [5, 10] has shown that $\sigma^* \in [1, 2]$. We recall the main results in this direction.

Theorem 2 *Let* (25) *hold. If* $\sigma > 2$, *a blocking strategy exists.*

Proof Assuming $\sigma > 2$, consider the positive constant $\lambda \doteq (\frac{\sigma^2}{4} - 1)^{-1/2}$, so that $\sigma = 2\sqrt{1+\lambda^2}/\lambda$. Using polar coordinates r, θ, let Γ be the closed curve consisting of two arcs of logarithmic spirals:

$$\Gamma \doteq \left\{(r\cos\theta, r\sin\theta);\, r = e^{\lambda|\theta|},\, \theta \in [-\pi, \pi]\right\}.$$

As in (17), let

$$\gamma(t) \doteq \Gamma \cap \overline{R^\Gamma(t)} = \left\{(r\cos\theta, r\sin\theta);\, r = e^{\lambda|\theta|},\, \theta \in [-\pi, \pi],\, r \leq 1 + t\right\}$$

be the portion of this barrier reached by the fire within time t. By the choice of λ, an elementary computation yields $m_1(\gamma(t)) = \sigma t$, showing that $\gamma(\cdot)$ is an admissible strategy. This strategy solves the blocking problem, because the reachable sets $R^\gamma(t)$ are all contained inside the region bounded by the closed curve Γ. □

On the other hand, proving the non-existence of a blocking strategy when σ is small is a far more difficult task. In this direction, two entirely different arguments were developed in [5] and [10].

Theorem 3 *Let* (25) *hold. If* $\sigma \leq 1$, *a blocking strategy does not exist.*

With reference to Fig. 3, this result can be motivated by the following intuitive argument. Assume that, for a construction speed $\sigma > 0$, a blocking strategy exists. Let Γ be the entire barrier constructed by this strategy, and let $\bar{x} \in \Gamma$ be the point where the "last brick" of the wall Γ is placed (see Fig. 3). Calling $T^\Gamma(\cdot)$ the minimum time function in (21), we must have

$$T^\Gamma(\bar{x}) = \sup_{x \in \Gamma} T^\Gamma(x) \geq \frac{m_1(\Gamma)}{\sigma}. \tag{26}$$

Indeed, the right hand side measures the total time needed to construct the barrier Γ. If the fire reaches the point $\bar{x}$ before this barrier is completed, it will spill outside, hence the reachable sets $R^\Gamma(t)$ will not remain within the bounded set enclosed by Γ.

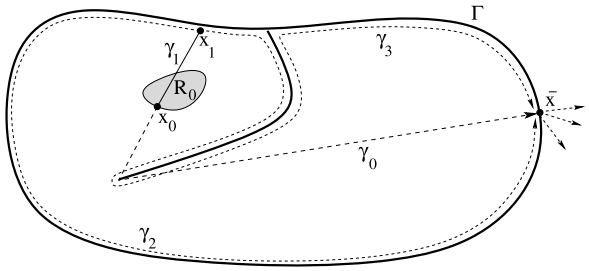


Fig. 3 If the barrier is constructed at speed $\sigma \leq 1$, then the fire reaches the point $\bar{x}$ and spills outside before the barrier is completed. Hence no blocking strategy can exist

We now estimate the left hand side of (26), i.e. the time needed by the fire to reach the point $\bar{x}$. Let γ_0 be a shortest path joining a point $x_0 \in R_0$ with $\bar{x}$ without crossing Γ. By prolonging this path backwards, we can find a point $x_1 \in \Gamma$ and a path $\gamma_1 \supset \gamma_0$, such that γ_1 is the shortest path joining x_1 with $\bar{x}$ without crossing Γ. An estimate on the length of γ_1 can be obtained as follows. Starting from x_1, move along Γ either clockwise or counterclockwise until the point $\bar{x}$ is reached. This yields two curves, say γ_2 and γ_3. The length of these curves satisfies

$$m_1(\gamma_2) + m_2(\gamma_3) \leq 2m_1(\Gamma). \tag{27}$$

Hence

$$T^{\Gamma}(\bar{x}) = m_1(\gamma_0) < m_1(\gamma_1) \leq \min\bigl\{m_1(\gamma_2), m_1(\gamma_3)\bigr\} \leq m_1(\Gamma). \tag{28}$$

If $\sigma \leq 1$, the inequality (28) is in contradiction with (26), hence a blocking strategy cannot exists. For a rigorous proof based on these ideas we refer to [10].

While the blocking problem on the entire plane is not yet fully understood, a sharp result is available in the case where fire propagation is restricted to a half plane $R_2^+ \doteq \{(x_1, x_2); x_2 \geq 0\}$. This models a situation where the presence of a river, or a lake, provides a natural barrier to fire propagation. In this case, the definition of reachable sets in (11) must be modified, requiring that all trajectories remain in the half plane R_+^2:

$$\begin{aligned} R^{\gamma}(t) \doteq \bigl\{&x(t); x(\cdot) \text{ absolutely continuous}, x(0) \in R_0, \\ &\dot{x}(\tau) \in F\bigl(x(\tau)\bigr) \text{ for a.e. } \tau \in [0,t], x(\tau) \in \mathbb{R}^2_+ \setminus \gamma(\tau) \text{ for all } \tau \in [0,t]\bigr\}. \end{aligned} \tag{29}$$

The following result was proved in [10].

Theorem 4 *Let* (25) *hold. Then, restricted to the half plane, for every bounded initial set* $R_0 \subset \mathbb{R}^2_+$ *a blocking strategy exists if and only if* $\sigma > 1$.

3.2 The Non-isotropic Case

In a realistic situation, as shown in Fig. 4, the fire propagates in various directions at different speeds. This happens, for example, if there is wind blowing in a preferred

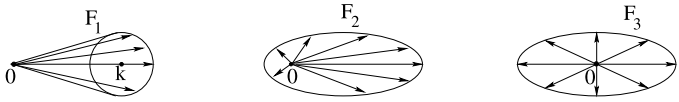


Fig. 4 The velocity sets F_1 and F_2 satisfy the assumptions (30), while the set F_3 does not

direction. This case is modeled by replacing the unit disc B_1 with more general velocity sets $F(x) \subset \mathbb{R}^2$. Sufficient conditions for the existence of a blocking strategy for the problem (1), (8) were derived in [6].

Theorem 5 *Assume that the velocity sets in* (1) *are independent of* x *and have the form*

$$F(x) \equiv F = \big\{(r\cos\theta, r\sin\theta);\, 0 \le r \le \rho(\theta), \theta \in [-\pi, \pi]\big\},$$

where the function $\rho : [-\pi, \pi] \mapsto \mathbb{R}_+$ *satisfies*

$$\rho(-\theta) = \rho(\theta), \qquad 0 \le \rho\big(\theta'\big) \le \rho(\theta) \quad \textit{for all } 0 \le \theta \le \theta' \le \pi. \tag{30}$$

If the wall construction speed satisfies

$$\sigma > [\textit{vertical width of } F] = 2 \max_{\theta\in[0,\pi]} \rho(\theta)\sin\theta$$

then, for every bounded initial set R_0, *a blocking strategy exists.*

Notice that the above result reduces to Theorem 3 in the case where $F = \overline{B}_1$ is the closed unit disc. Further results on existence or non-existence of blocking strategies can be obtained by comparison with the isotropic case, using Lemma 2.

4 Existence of Optimal Strategies

Consider the dynamic blocking problem with dynamics described by the differential inclusion (1) and by the admissibility condition (9). Let the cost functional be described by (13). The following result on the existence of optimal blocking strategies was proved in [7].

Theorem 6 *Let the assumptions* (A1)–(A3) *hold. If there exists an admissible strategy such that* $J(\gamma) < \infty$, *then the optimization problem* (OP1) *admits an optimal solution.*

Proof The proof given in [7] relies on the direct method of the Calculus of Variations. Consider a minimizing sequence $\gamma_n(\cdot)$ of admissible strategies, such that

$$J(\gamma_n) \to \inf_{\gamma\in\mathcal{A}} J(\gamma),$$

where the infimum is taken over the family of all admissible blocking strategies. An optimal strategy $\gamma^*(\cdot)$ is then obtained by taking a suitable limit of the $\gamma_n(\cdot)$.

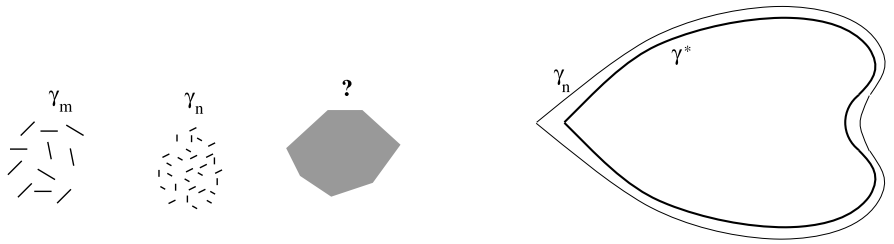


Fig. 5 *Left*: the limit of a sequence of bounded rectifiable sets can only be interpreted in measure sense, and may not yield a rectifiable set. *Right*: the limit of a sequence of *connected* rectifiable sets is a rectifiable set

One should be aware, however, that no regularity is known a priori about the rectifiable sets $\gamma_n(t)$. Hence (see Fig. 5) as $n \to \infty$ there is no guarantee that these sets will converge to a rectifiable set $\gamma^*(t)$. A more careful argument requires several steps. The key idea is to split each set $\gamma_n(t)$ into connected components of decreasing length, and take limits componentwise.

1. By possibly enlarging the sets $\gamma_n(t)$, is not restrictive to assume that the $\gamma_n(\cdot)$ are complete strategies, i.e., for every $t \geq 0$ one has

$$\left\{x \in \mathbb{R}^2; \limsup_{r \downarrow 0} \frac{m_1(B(x,r) \cap \gamma_n(t))}{r} > 0\right\} \subseteq \gamma_n(t) = \bigcap_{s>t} \gamma_n(s).$$

2. For each rational time τ, let the connected components of $\gamma_n(\tau)$ be ordered according to decreasing length, so that

$$\gamma_n(\tau) = \gamma_{n,1}(\tau) \cup \gamma_{n,2}(\tau) \cup \gamma_{n,3}(\tau) \cup \cdots$$

with

$$\ell_{n,1}(\tau) \geq \ell_{n,2}(\tau) \geq \ell_{n,3}(\tau) \geq \cdots, \qquad \ell_{n,i}(\tau) \doteq m_1\big(\gamma_{n,i}(\tau)\big).$$

Notice that, by completeness, each connected component $\gamma_{n,i}(\tau)$ must be closed. Taking a subsequence, as $n \to \infty$ we can assume that, for every rational time $\tau \geq 0$,

$$\ell_{n,i}(\tau) \to \ell_i(\tau), \qquad d_H\big(\gamma_{n,i}(\tau), \gamma_i(\tau)\big) \to 0.$$

Here $d_H(\cdot,\cdot)$ denotes the Hausdorff distance between compact sets. We then define $\gamma(\tau) \doteq \bigcup_{\ell_i(\tau)>0} \gamma_i(\tau)$. Finally, the optimal strategy $\gamma^*(\cdot)$ is defined as the completion of $\gamma(\cdot)$.

3. Using the lower semicontinuity of the functions ψ, β, for every $t \geq 0$ one obtains

$$\int_{\gamma^*(t)} \psi \, dm_1 \leq \liminf_{n\to\infty} \int_{\gamma_n(t)} \psi \, dm_1 \leq t, \qquad \int_{\gamma^*(t)} \beta \, dm_1 \leq \liminf_{n\to\infty} \int_{\gamma_n(t)} \beta \, dm_1.$$

By the first inequality, the limit strategy $\gamma^*(\cdot)$ is admissible.

4. The last (and most difficult) step is to prove the inequalities

$$\int_{R^{\gamma^*}(t)} \alpha \, dm_2 \leq \liminf_{n\to\infty} \int_{R^{\gamma_n}(t)} \alpha \, dm_2.$$

These are achieved by showing that, for every $t \geq 0$, the sets $R^{\gamma_n}(t)$ are "almost as big" as the reachable set $R^{\gamma^*}(t)$. In other words, assume that there exists a trajectory $\tau \mapsto x(\tau)$ for the fire, satisfying (1), and reaching a point $x(t) = \bar{x}$ without crossing the wall $\gamma^*(\tau)$ for any $\tau \in [0, t]$. Then for every $n \geq 1$ sufficiently large, there exists a trajectory $\tau \mapsto x_n(\tau)$ reaching a point $x_n(t)$ close to $\bar{x}$ without crossing the barriers $\gamma_n(\tau)$. The proof of this statement requires a careful analysis of solutions to the differential inclusion (1), involving both topological and measure-theoretic arguments. A key step calls for the partition of the plane $\mathbb{R}^2$ into a checkerboard, whose squares Q_k are colored either white or black depending on the length $m_1(\gamma_n(\tau) \cup Q_k)$. For all details we refer to [7].

An entirely different proof of Theorem 6, based on the analysis of the minimal time function T^Γ in (21), was recently developed in [15]. We review here the main ideas.

1. Consider a minimizing sequence of admissible barriers Γ_k, $k \geq 1$. Let $T_k \doteq T^{\Gamma_k}$ be the corresponding minimum time functions. For every $t \geq 0$, the functions $T_k \wedge t$ are SBV functions, which admit the characterization stated in Theorem 1.

2. Using the Ambrosio-De Giorgi compactness theorem for SBV functions [1], one obtains a convergent subsequence $T_k \to U$ such that

(i) Recalling the notation (22), for every $t \geq 0$ one has $(U \wedge t) \in SBV$.
(ii) The jump set of $U \wedge t$ satisfies

$$\int_{J_{(U\wedge t)}} \psi \, dm_1 \leq \liminf_{k\to\infty} \int_{J_{(T_k\wedge t)}} \psi \, dm_1 \leq t. \tag{31}$$

3. By (31) the rectifiable set Γ, obtained by taking the completion of J_U, is an admissible barrier. Standard lower semicontinuity estimates now yield

$$\int_{R^\Gamma_\infty} \alpha \, dm_2 \leq \liminf_{k\to\infty} \int_{R^{\Gamma_k}_\infty} \alpha \, dm_2, \qquad \int_\Gamma \beta \, dm_1 \leq \liminf_{k\to\infty} \int_{\Gamma_k} \beta \, dm_1.$$

Hence Γ is an optimal barrier. □

5 Necessary Conditions for Optimality

Given the minimization problem (OP2), assume that an optimal barrier Γ exists, consisting of the union of finitely many, sufficiently regular arcs. By deriving necessary conditions for optimality, one seeks to determine these optimal arcs, as solutions to a family of ODEs together with boundary conditions.

Following a standard procedure in the Calculus of Variations, necessary conditions are obtained by the analysis of perturbations. For the present problem, however, these conditions take different forms depending on the various types of arcs. As a preliminary, one must therefore introduce a classification of optimal arcs.

Let Γ be an admissible barrier for the differential inclusion (1), so that (16) holds. Observe that the presence of this barrier has two effects, namely: (i) it restricts the fire to the set R^Γ_∞, consisting of all connected components of $\mathbb{R}^2 \setminus \Gamma$ which intersect

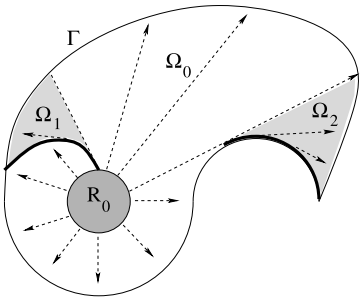


Fig. 6 Here we take $R_0 = B_1$, and $F(x) \equiv F = \overline{B}_1$, the unit disc centered at the origin. The *two thick arcs* denote the portion $\Gamma^d \subset \Gamma$ which contributes to slowing down the propagation of the fire. Notice that $T^\Gamma(x) = T(x)$ for $x \in \Omega_0$, but $T^\Gamma(x) < T(x)$ for $x \in \Omega_1 \cup \Omega_2$. The *thick arc* next to the *shaded region* Ω_1 lies in $\Gamma^d \setminus \Gamma^b$, while the *thick arc* next to Ω_2 lies in $\Gamma^d \cap \Gamma^b$

the initial domain R_0, and (ii) within the set R^Γ_∞, it can slow down the advancement of the fire.

This fact, illustrated in Fig. 6, can be better described as follows. Given the differential inclusion (1) and the barrier Γ, let $T^\Gamma(\cdot)$ be the minimum time function, defined at (21). Calling $T(\cdot)$ the minimum time function for the original problem (1) without any barrier, one clearly has

$$0 \leq T(x) \leq T^\Gamma(x) \quad \text{for all } x \in \mathbb{R}^2.$$

We say that a point $x \in \Gamma$ belongs to the delaying portion of the barrier if, by modifying the set Γ in an arbitrarily small neighborhood of x, one can change the minimal time function *somewhere else*:

Definition 2 The subset $\Gamma^d \subseteq \Gamma$ of *delaying walls* is the set of all points $x \in \Gamma$ such that, for some $\delta > 0$, the following holds. For every $\varepsilon > 0$ there exists an admissible rectifiable set Γ' with $\Gamma' \setminus B(x, \varepsilon) = \Gamma \setminus B(x, \varepsilon)$ and such that $T^{\Gamma'}(y) \neq T^\Gamma(y)$ at some point $y \notin B(x, \delta)$.

We think of Γ^d as a portion of the barrier Γ which contributes to slowing down fire propagation. In addition, the barrier Γ will contain an outer portion Γ^b, separating the burned from the unburned region.

Definition 3 The subset $\Gamma^b \subseteq \Gamma$ of *blocking walls* is defined as $\Gamma^b \doteq \Gamma \cap \partial(R^\Gamma_\infty)$.

Remark 5 If Γ is optimal and the construction cost β in (13) is strictly positive, then $\Gamma = \Gamma^d \cup \Gamma^b$. Indeed, any arc $\Gamma' \subset \Gamma$ contained in the interior of the reachable set R^Γ_∞ must be part of Γ^d, otherwise the alternative strategy $\widetilde{\Gamma} \doteq \Gamma \setminus \Gamma'$ would also be admissible, with a smaller cost. On the other hand, as shown in Fig. 6, one can have $\Gamma^d \cap \Gamma^b \neq \emptyset$.

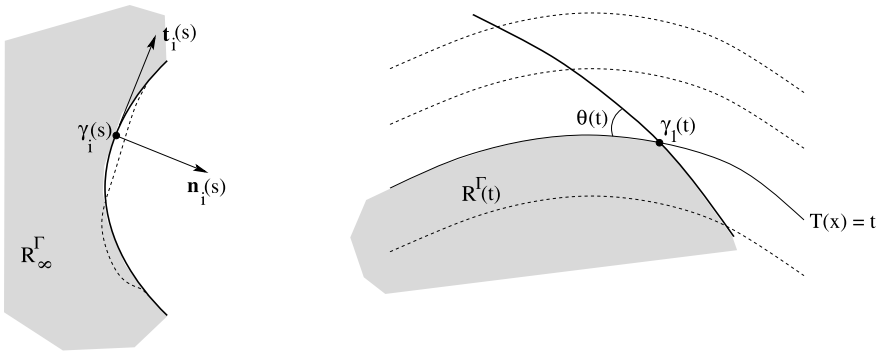


Fig. 7 *Left*: a free arc γ_i, parameterized by arc-length. Here any small perturbation having the same length (the *dotted line*) yields another admissible barrier. Hence the optimality conditions are the same as in isoperimetric problems. *Right*: a single boundary arc γ_1. In this case the admissibility condition already suffices to determine the arc

Given an admissible barrier Γ, a further classification of arcs can be achieved as follows. Define the set of times

$$\mathcal{S} \doteq \left\{ t \geq 0; \int_{\Gamma \cap \overline{R^{\Gamma}(t)}} \psi \, dm_1 = t \right\}. \tag{32}$$

These are the times where the admissibility constraint is *saturated*, i.e. it is satisfied as an equality. We can further classify points $x \in \Gamma$ by setting

$$\Gamma_{\mathcal{S}} \doteq \left\{ x \in \Gamma; T^{\Gamma}(x) \in \mathcal{S} \right\}, \qquad \Gamma_{\mathcal{F}} \doteq \left\{ x \in \Gamma; T^{\Gamma}(x) \notin \mathcal{S} \right\}.$$

Following [5], arcs lying in the subset $\Gamma_{\mathcal{F}}$ will be called *free arcs*, while arcs lying in $\Gamma_{\mathcal{S}}$ will be called *boundary arcs*. Notice that boundary arcs are constructed right at the edge of the advancing fire front. On the other hand, free arcs represent a preemptive strategy: they are put in place in advance, at locations which will be reached by the fire only at a later time.

In the following, for simplicity we discuss necessary conditions for free and boundary arcs, assuming that the functions $\beta(x) \equiv \beta$ and $\psi(x) \equiv \sigma^{-1}$ are both constant. Results valid in the general case can be found in [11]. Furthermore, in the case of delaying arcs, necessary conditions for optimality were recently derived in [23].

5.1 Free Arcs

Let Γ be an optimal barrier. Assume that, during a time interval $[t_0, t_1]$, this optimal strategy simultaneously constructs N free, blocking arcs: $\gamma_1, \ldots, \gamma_N \subset \Gamma_{\mathcal{F}} \cap \Gamma^b$. Referring to Fig. 7, let $s \mapsto \gamma_i(s)$, $s \in [a_i, b_i]$ be a parameterization of γ_i in terms of arc-length, so that $|\dot{\gamma}_i(s)| \equiv 1$. Consider the unit tangent vector $\mathbf{t}_i(s) \doteq \dot{\gamma}_i(s)$ and let

$\mathbf{n}_i(s)$ be the unit normal vector, oriented toward the exterior of the set R_∞^Γ burned by the fire. Let $\kappa_i(s)$ be the curvature of γ_i at the point $\gamma_i(s)$, so that

$$\ddot{\gamma}_i(s) = \dot{\mathbf{t}}_i(s) = \kappa_i(s)\mathbf{n}_i(s). \tag{33}$$

Since γ_i is a free arc, every curve $\tilde{\gamma}_i$ sufficiently close to γ_i, with the same length and the same endpoints, will also yield an admissible barrier. Notice that, if γ_i is not a segment, many such perturbations $\tilde{\gamma}$ can be constructed. The necessary conditions for optimality thus take the same form as in the classical isoperimetric problem of the Calculus of Variations.

Theorem 7 *Let $\gamma_1, \ldots, \gamma_N \subset \Gamma_{\mathcal{F}} \cap \Gamma^b$ be free arcs, simultaneously constructed by an optimal strategy. Then the curvature of each arc is proportional to the local value $\alpha(\cdot)$ of the land. Indeed, either $\kappa_i(s) \equiv 0$ (hence all arcs are straight segments), or there exists a Lagrange multiplier $\lambda \geq 0$ such that*

$$\left(\beta + \frac{\lambda}{\sigma}\right)\kappa_i(s) = \alpha\big(\gamma_i(s)\big) \quad \text{for all } i \in \{1, \ldots, N\},\ a_i < s < b_i. \tag{34}$$

In particular, if $\alpha(x) \equiv \alpha$ is a constant, then the curves γ_i are arcs of circumferences, all with the same radius $r = \frac{1}{\kappa} = \frac{\beta + (\lambda/\sigma)}{\alpha}$. It is important to notice that the constant λ is the same for every arc γ_i, $i \in \{1, \ldots, N\}$. Calling $r_i(s) = 1/\kappa_i(s)$ the radius of curvature, one has

$$\lambda = \big(\alpha\big(\gamma_i(s)\big) \cdot r_i(s) - \beta\big)\sigma. \tag{35}$$

As shown in [11], the Lagrange multiplier λ can be interpreted as the *instantaneous value of time*. The next paragraph provides an intuitive explanation of this concept.

Assume that, in an idealized situation, we could "buy time". In other words, assume that we had at our disposal a short time interval $[t, t+\varepsilon]$ to construct an additional portion of barrier, while in the meantime the fire front did not advance. Making the most of this advantage, we could thus reduce the total area eventually burned by the fire, and hence the total cost. Roughly speaking, we say that $V(t)$ is the *instantaneous value of time* (at time t) if

$$[\text{reduction of the total cost}] = \varepsilon \cdot V(t) + o(\varepsilon),$$

where the Landau symbol $o(\varepsilon)$ denotes a higher order infinitesimal as $\varepsilon \to 0$. In general, one can prove that $t \mapsto V(t)$ is a nonincreasing function. If at some time τ the fire propagation is extinguished, then $V(t) = 0$ for all $t > \tau$. In the situation described by Theorem 7, the value of time is actually constant: $V(t) \equiv \lambda$ during the interval of time when the free arcs $\gamma_1, \ldots, \gamma_N$ are being constructed.

5.2 A Single Boundary Arc

Next, assume that during a time interval $[t_1, t_2]$ the optimal strategy constructs one single boundary arc $\gamma_1 \subset \Gamma_{\mathcal{S}} \cap \Gamma^b$. We choose a parameterization $t \mapsto \gamma_1(t)$ of this arc so that each point $\gamma(t)$ lies on the level set $\{x;\, T^\Gamma(x) = t\}$.

In this case, an equation determining the arc γ_1 can already be derived from the admissibility condition (8), without using any optimality condition. Indeed, let $T(\cdot)$ be the minimum time function and let

$$h(x) \doteq \left|\nabla T(x)\right|^{-1} \tag{36}$$

be the propagation speed of the fire front in the normal direction, as in (4). We then have the identities

$$\left|\dot\gamma_1(t)\right| \equiv \sigma, \qquad T\big(\gamma_1(t)\big) = t \quad \text{for all } t \in [t_1, t_2]. \tag{37}$$

With reference to Fig. 7, let θ_1 be the angle between the curve γ_1 and the level curve of the minimum time function $T(\cdot)$, at a point x. By (37), one has

$$\sigma \cdot \sin\theta_1(x) = h(x). \tag{38}$$

If the initial point $\gamma_1(t_1)$ is known, from (38) one can recover the entire curve γ_1.

5.3 Several Boundary Arcs Constructed Simultaneously

We now consider a more general situation where ν boundary arcs $\gamma_1, \gamma_2, \ldots, \gamma_\nu \in \Gamma_{\mathcal{S}} \cap \Gamma^b$ are simultaneously constructed, on a time interval $t \in [t_1, t_2]$. Let each arc be parameterized by time, so that

$$T\big(\gamma_i(t)\big) = t \quad \text{for all } t \in [t_1, t_2],\ i \in \{1, \ldots, \nu\}. \tag{39}$$

The admissibility condition (8) yields

$$\sum_{i=1}^{\nu} \left|\dot\gamma_i(t)\right| = \sigma \quad \text{for all } t \in [t_1, t_2]. \tag{40}$$

In the present case where $\nu \geq 2$, the equations (39)–(40) are not enough to uniquely determine the arcs γ_i. Additional conditions, derived from the optimality of the barrier Γ, must be used. Following [5], we will show how the problem of determining these optimal arcs can be reduced to a standard problem of optimal control.

Call $w_i^*(t) \doteq \sigma^{-1}|\dot\gamma_i(t)|$ the portion of overall resources devoted to the construction of the arc γ_i, at time t. We regard the map $t \mapsto w^*(t) = (w_1^*(t), \ldots, w_\nu^*(t)) \in \Delta^\nu$ as a control function, taking values in the unit simplex

$$\Delta^\nu \doteq \left\{(w_1, \ldots, w_\nu);\, w_i \geq 0, \sum_{j=1}^{\nu} w_j = 1\right\}. \tag{41}$$

We now choose a set of coordinates $(t, s) \mapsto x_i(t, s)$, on a neighborhood of each arc γ_i. To fix the ideas (see Fig. 8), for each time t let the map $s \mapsto x_i(t, s)$ provide an arc-length parametrization of the level set $\{x;\, T(x) = t\}$, in a neighborhood of $\gamma_i(t)$. Moreover, let $\mathbf{e}_i = \frac{\partial x_i(t,s)}{\partial s}$ be the unit tangent vector to this level set.

Consider a second admissible strategy $t \mapsto w(t) = (w_1(t), \ldots, w_\nu(t))$. This will result in the construction of different arcs $t \mapsto y_i(t)$, determined by the equations

$$\left|\dot y_i(t)\right| = \sigma\, w_i(t), \qquad T\big(y_i(t)\big) = t,$$

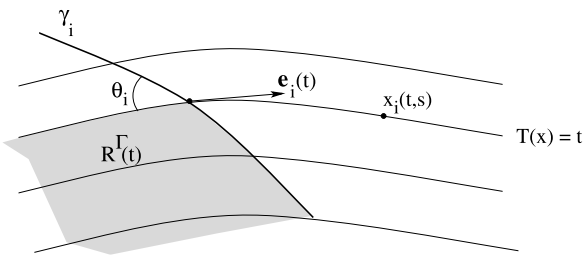


Fig. 8 Choice of the coordinates (t, s) in a neighborhood of the arc γ_i

with the same initial conditions at time $t = t_1$:

$$y_i(t_1) = \gamma_i(t_1) \quad \text{for all } i \in \{1, \ldots, \nu\}.$$

In our coordinate system, let $y_i(t) = x_i(t, s_i(t))$. For each i, the scalar function $s_i(\cdot)$ will satisfy an ODE of the form

$$\dot{s}_i = f_i\big(t, s_i(t), w_i(t)\big). \tag{42}$$

Here the right hand side f_i is implicitly determined by the scalar constraint

$$\left| \frac{\partial x_i(t, s_i)}{\partial t} + f_i(t, s_i, w_i) \frac{\partial x_i(t, s_i)}{\partial s_i} \right| = \sigma w_i. \tag{43}$$

This accounts for the fact that $|\dot{y}_i| = \sigma w_i$. We observe that (43) admits solutions provided that

$$h\big(x_i(t, s_i)\big) \leq \sigma w_i. \tag{44}$$

Indeed, the speed $|\dot{y}_i|$ at which the barrier is constructed cannot be smaller than the propagation speed of the fire front, in the normal direction. In the case of a strict inequality, (43) has exactly two solutions, say $f_i^- < f_i^+$. Clearly, the correct choice depends on the relative position of the burned region. For example, if the burned region locally has the representation $\{x_i(t, s); s < s_i(t)\}$ (as in Fig. 8), then one should choose $f_i = f_i^-$.

Consider the control system consisting of the ν equations (42), supplemented by the initial and terminal constraints

$$x_i\big(t_1, s_i(t_1)\big) = \gamma_i(t_1), \qquad x_i\big(t_2, s_i(t_2)\big) = \gamma_i(t_2) \quad i = 1, \ldots, \nu. \tag{45}$$

For this system, consider the optimization problem

$$\text{minimize:} \quad \Lambda(w) \doteq \sum_{i=1}^{\nu} \int_a^b L_i\big(t, s_i(t), w_i(t)\big)\, dt, \tag{46}$$

where the running costs have the form

$$L_i(t, s_i, w_i) \doteq \beta \sigma w_i(t) + \int_{\bar{s}_i}^{s_i} h\big(x_i(t, \xi)\big) \alpha\big(x_i(t, \xi)\big)\, d\xi, \tag{47}$$

with h as in (36). The minimum in (46) is sought among all control functions $w : [t_1, t_2] \mapsto \Delta^{\nu}$. Notice that the first term in (47) accounts for the cost of building

the wall, while the second term is related to the value of the burned area. Here the choice of the constants $\bar{s}_i$ is immaterial, because it does not affect the minimizers.

The system of ODEs (42) and boundary conditions (45), together with the integral functional at (46)–(47) yields an optimal control problem in standard form. The optimal control $t \mapsto w(t) = (w_1^*(t), \ldots, w_\nu^*(t))$ thus satisfies the Pontryagin maximum principle [8, 16, 22]. In turn, this yields a set of necessary conditions for the optimal arcs γ_i.

Theorem 8 *Let Γ be an optimal barrier. Let $\gamma_1, \ldots, \gamma_\nu \subset \Gamma_{\mathcal{S}} \cap \Gamma^b$ be boundary arcs simultaneously constructed during the time interval $[t_1, t_2]$, parameterized as in (39). Call $\mathbf{e}_i(t)$ the unit vector tangent to the boundary of the reachable set $R(t)$ at the point $\gamma_i(t)$, oriented toward the exterior of R_∞^Γ (see Fig. 8).*

Then there exists a constant $\lambda_0 \geq 0$ and nontrivial solutions to the adjoint equations

$$\dot{p}_i(t) = \frac{\langle \dot{\gamma}_i(t), \dot{\mathbf{e}}_i(t) \rangle}{\langle \dot{\gamma}_i(t), \mathbf{e}_i(t) \rangle} p_i(t) - \lambda_0 h\big(\gamma_i(t)\big)\alpha\big(\gamma_i(t)\big) \tag{48}$$

for $i = 1, \ldots, \nu$, such that the functions

$$V_i(t) \doteq \left\langle \frac{\dot{\gamma}_i(t)}{|\dot{\gamma}_i(t)|}, \mathbf{e}_i(t) \right\rangle^{-1} \cdot p_i(t) \tag{49}$$

all coincide, at each time $t \in [t_1, t_2]$.

A proof of this theorem was first obtained in [5]. For a more general result, valid for general functions $\beta(\cdot)$ and $\psi(\cdot)$, we refer to [11].

With reference to Fig. 8, the inner product in (48) is related to the angle θ_i between the barrier and the fire front. Namely

$$\left\langle \frac{\dot{\gamma}_i(t)}{|\dot{\gamma}_i(t)|}, \mathbf{e}_i(t) \right\rangle = \cos\theta_i(t).$$

According to Theorem 8, the quantity

$$V(t) \doteq \left(\frac{p_i(t)}{\cos\theta_i(t)} - \beta \right) \sigma \tag{50}$$

is independent of $i = 1, \ldots, \nu$. For a suitable choice of the boundary conditions for the p_i in (48), the function $V(\cdot)$ can be interpreted as the *instantaneous value of time* [11].

5.4 Necessary Conditions at Junctions

The optimality conditions derived in the previous subsections provide a set of ODEs satisfied by the optimal arcs. In order to uniquely determine these arcs, one needs suitable boundary conditions. These can be obtained by studying what happens at points where arcs originate, or at junctions between arcs of different types. Referring to Fig. 9, we recall here two results proved in [5].

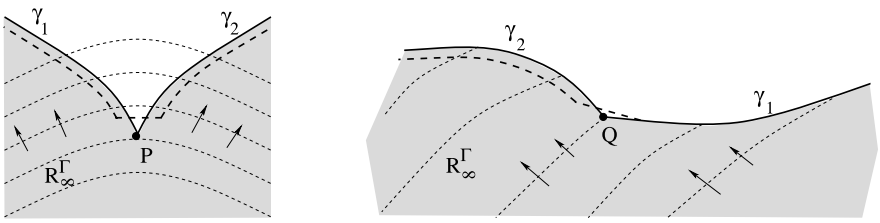


Fig. 9 *Left*: The boundary arcs γ_1, γ_2 originate at the same point P. They can be replaced by a new admissible barrier (the *thick dotted line*), reducing the total cost. *Right*: the free arc γ_1 and the boundary arc γ_2 join non-tangentially at the point Q. They can again be replaced by a new admissible barrier (the *thick dotted line*), reducing the total cost. Here the *thin lines* are the level set of the minimum time function, while the *arrows* give the direction of fire propagation

- *A barrier containing two boundary arcs γ_1, γ_2 originating from the same point cannot be optimal.*
- *If an optimal barrier contains a free arc γ_1 and a boundary arc γ_2 with a point in common, then they must be tangent at the point of junction.*

Further necessary conditions, valid at junction points, can be found in [11]. In particular, at the time t when a free arc joins a boundary arc, the two expressions for the instantaneous value of time (35) and (50) coincide.

6 Sufficient Conditions for Optimality

We shall consider the isotropic case, where the fire is initially burning on the open unit disc $B_1 \subset \mathbb{R}^2$ and propagates with unit speed in all directions. Let a construction speed $\sigma > 2$ and constant costs $\alpha > 0$, $\beta \geq 0$ be given. As suggested by the necessary conditions derived in [5, 11], the optimal barrier Γ^* which minimizes the cost J in (19) should consist of an arc of circumference and two arcs of logarithmic spirals.

We now describe more precisely this barrier (Fig. 10). For every $\tau > 0$ small enough, there exists a unique arc of circumference $\tilde{\gamma}_\tau$ with the following properties:

(i) $\tilde{\gamma}_\tau$ is symmetric w.r.t. the x_1-axis and has length $m_1(\tilde{\gamma}_\tau) = \sigma\tau$,
(ii) the endpoints P^-, P^+ lie on the circumference $\{|x| = 1 + \tau\}$,
(iii) the angle α between the two circumferences at P^- and at P^+ satisfies $\sin\alpha = 2/\sigma$.

In addition, consider the two arcs of logarithmic spirals $\gamma_\tau^+, \gamma_\tau^-$, defined as

$$\gamma_\tau^\pm = \big\{(r\cos\theta, \pm r\sin\theta);\, r = r_0 e^{\lambda\theta}, r \geq 1 + \tau, \theta \leq \pi\big\}. \tag{51}$$

Here $\lambda = \sqrt{\frac{4}{\sigma^2-4}}$, while the constant r_0 is chosen so that the two arcs start from the points P^+, P^- respectively. The above choice of the constants α, λ implies that the arcs $\gamma_\tau^\pm$ meet the circular arc $\tilde{\gamma}_\tau$ tangentially at $P^\pm$.

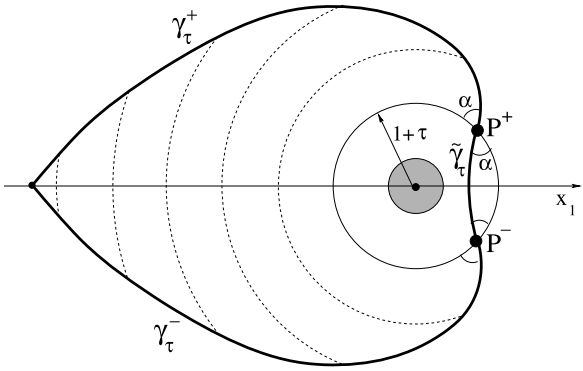


Fig. 10 Construction of the barrier $\Gamma_\tau = \tilde\gamma_\tau \cup \gamma_\tau^+ \cup \gamma_\tau^-$

For every fixed τ, the union $\Gamma_\tau \doteq \tilde\gamma_\tau \cup \gamma_\tau^+ \cup \gamma_\tau^-$ of these three arcs is a simple closed curve. By minimizing the cost $J(\Gamma_\tau)$ over the scalar parameter τ, we single out the curve

$$\Gamma^* \doteq \Gamma_{\tau^*}, \quad \tau^* \doteq \arg\min_{\tau>0} J(\Gamma_\tau). \tag{52}$$

At present, it is not known whether the barrier Γ^* is a global minimizer. A partial result in this direction was recently proved in [12].

Theorem 9 *The barrier Γ^* is optimal within the family of all admissible Jordan curves with finite length.*

In other words, if Γ is any simple closed curve with length $m_1(\Gamma) < \infty$, which is admissible according to (8), then

$$J\big(\Gamma^*\big) \le J(\Gamma). \tag{53}$$

In fact, one can show that the inequality in (53) is strict, except when Γ is the image of Γ^* by a rotation around the origin. Observe that, if the construction cost β is strictly positive, then any optimal curve must have finite length.

Referring to Fig. 11, we sketch the three main steps in the proof of Theorem 9.

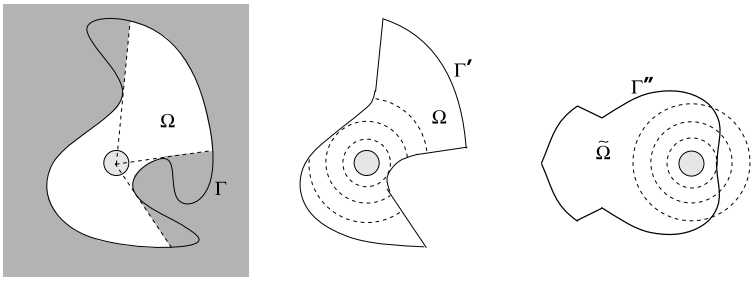


Fig. 11 The mains steps in the proof of Theorem 9

Proof 1. Consider any admissible, simple closed curve Γ. Thinking of Γ as a wall which blocks the light, let $\Omega \subset \mathbb{R}^2$ be the set of points illuminated by a light source located at the origin $0 \in \mathbb{R}^2$. Then, using techniques developed in [2], one can show that the boundary $\Gamma' \doteq \partial\Omega$ is also an admissible, simple closed curve. Moreover, $J(\Gamma') \leq J(\Gamma)$.

2. By the previous construction, the domain Ω is star-shaped. Indeed, it can be represented in polar coordinates as

$$\Omega = \big\{(r\cos\theta, r\sin\theta);\, r \leq r(\theta), \theta \in [-\pi, \pi]\big\} \tag{54}$$

for some (possibly discontinuous) function $\theta \mapsto r(\theta)$ having bounded variation. As in [14], let $\theta \mapsto \tilde{r}(\theta)$ be the symmetric, non-decreasing rearrangement of the map $\theta \mapsto r(\theta)$. Consider the symmetric domain

$$\widetilde{\Omega} = \big\{(r\cos\theta, r\sin\theta);\, r \leq \tilde{r}(\theta), \theta \in [-\pi, \pi]\big\}. \tag{55}$$

The boundary $\Gamma'' \doteq \partial\widetilde{\Omega}$ of this new domain is an admissible, simple closed curve. Moreover, $J(\Gamma'') \leq J(\Gamma')$.

3. Since the barrier Γ'' is optimal within the class of simple closed curves, it must satisfy the necessary conditions stated in Sect. 5. As a result, we conclude that the symmetric curve Γ'' is a concatenation of

– free arcs, which must be arcs of circumferences
– boundary arcs, which must be arcs of logarithmic spirals of the form

$$\big\{(r\cos\theta, \pm r\sin\theta);\, r = ce^{\lambda\theta}, r \in [r_1, r_2]\big\}, \quad \text{with } \lambda = \sqrt{\frac{4}{\sigma^2 - 4}}.$$

Moreover, each arc must join tangentially with the previous one.

A simple geometric argument now shows that the curves Γ_τ considered at (52), and their images under rigid rotations around the origin, are the only symmetric curves with $\theta \mapsto r(\theta)$ nondecreasing, which satisfy all these necessary conditions. This concludes the proof. For all details we refer to [12]. □

7 Numerical Computation of Optimal Barriers

A first algorithm for the computation of optimal barriers was developed in [9]. This construction relies on two basic assumptions about the optimal barrier, namely:

(i) The optimal barrier Γ is a simple closed curve, which can be represented as

$$\Gamma = \big\{(r\cos\theta, r\sin\theta);\, r = \rho(\theta)\big\}$$

for some positive continuous function $\rho(\cdot)$. In particular, the corresponding set R_∞^Γ burned by the fire is star-shaped.

(ii) The optimal barrier does not contain delaying arcs. Hence, restricted to the set R_∞^Γ enclosed by the curve Γ, the minimal time function satisfies $T^\Gamma(x) = T(x)$. In particular, the function $T(\cdot)$ can be determined in advance, before the computation of Γ.

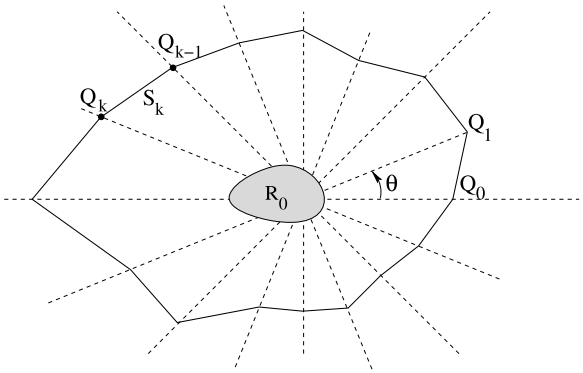


Fig. 12 A polygonal approximation to the optimal barrier

Under the assumptions (i)–(ii), barriers can be approximated by polygonal curves, whose vertices lie on a family of rays through the origin (see Fig. 12). To define one of these approximations, we fix an integer $n \geq 3$ and consider closed polygonal curves $\mathcal{P}$ having vertices at points $Q_k = (\rho_k \cos k\theta, \rho_k \sin k\theta)$, $0 \leq k \leq n$. Here $\rho_0, \ldots, \rho_n$ are positive numbers with $\rho_0 = \rho_n$, while $\theta \doteq 2\pi/n$. We call S_k the edge of the polygonal joining Q_{k-1} with Q_k. Its length is computed by

$$\|S_k\| = \sqrt{\rho_k^2 + \rho_{k-1}^2 - 2\rho_k \rho_{k-1} \cos\theta}.$$

Setting $\boldsymbol{\rho} \doteq (\rho_1, \ldots, \rho_n)$, the area enclosed by the polygonal is computed as

$$A(\boldsymbol{\rho}) = \frac{1}{2} \sin\theta \cdot \sum_{k=1}^{n} \rho_k \rho_{k-1},$$

while the total length is

$$L(\boldsymbol{\rho}) = \sum_{k=1}^{n} \sqrt{\rho_k^2 + \rho_{k-1}^2 - 2\rho_k \rho_{k-1} \cos\theta}.$$

Moreover, the minimum time needed by the fire to reach some point on the segment S_k is defined as

$$T(S_k) \doteq \inf\big\{T(x);\, x \in S_k\big\}.$$

A discrete approximation to the constrained optimization problem (19), (8) with $\alpha(x) \equiv \alpha$ and $\beta(x) \equiv \beta$ constant, is given by

$$\min_{\boldsymbol{\rho}} \big\{\alpha \cdot A(\boldsymbol{\rho}) + \beta \cdot L(\boldsymbol{\rho})\big\}, \tag{56}$$

subject to the family of constraints

$$\sum_{k=1}^{m} \|S_{i_k}\| - \sigma \cdot T(S_{i_m}) \leq 0 \quad \text{for all } m = 1, 2, \ldots, n. \tag{57}$$

Here $(i_1, i_2, \ldots, i_n)$ is some permutation of the indices $(1, 2, \ldots, n)$ such that

$$T(S_{i_1}) \leq T(S_{i_2}) \leq \cdots \leq T(S_{i_n}).$$

We denote by $\mathcal{F}_{\boldsymbol{\rho}}$ the collection of all such possible permutations, for a given $\boldsymbol{\rho}$. Observe that, if the constraints (57) are satisfied for one permutation $\alpha \in \mathcal{F}_{\boldsymbol{\rho}}$, then they are necessarily satisfied for every permutation $\beta \in \mathcal{F}_{\boldsymbol{\rho}}$.

In a first step, the algorithm constructs a local minimum for the finite-dimensional problem (56)–(57). In the next step, the number of vertices is doubled, replacing θ with $\theta/2$. A further minimization process is carried out, starting with the local minimizer constructed at the previous step, etc....

The results of some numerical experiments using this algorithm are reported in [9].

8 Open Problems

8.1 Isotropic Blocking Problem

On the entire plane $\mathbb{R}^2$, assume that $F(x) \equiv \overline{B}_1$, so that the fire propagates with unit speed in all directions. Let the barrier be constructed at speed $\sigma > 0$. For any (nonempty, open, bounded) initial set R_0, it is then known that a blocking strategy exists if $\sigma > 2$, and does not exist if $\sigma \leq 1$. The case $1 < \sigma \leq 2$ remains open. A reasonable conjecture is that, if $\sigma \leq 2$, then the fire cannot be blocked. In this direction, the following observations may be useful.

- As shown in Sect. 3, it is not restrictive to assume that the initial set $R_0 \subset \mathbb{R}^2$ is the unit disc centered at the origin.
- If the barrier Γ is a simple closed curve, then the estimate (27) can be replaced by

$$m_1(\gamma_2) + m_1(\gamma_3) \leq m_1(\Gamma).$$

  The same is true if Γ is the union of finitely many simple closed curves. In this case, the argument given at (26)–(28) already implies that the construction speed for a blocking strategy must be $\sigma > 2$. We conclude that, if a blocking strategy exists for $\sigma \leq 2$, the barrier Γ must also contain purely delaying arcs.
- Since the fire front propagates with unit speed in the normal direction, to be effective any boundary arc γ_i must be constructed at speed $\sigma_i(t) > 1$. Indeed, let $t \mapsto \gamma_i(t)$ be a parameterization of this arc according to the construction time, so that $T^{\Gamma}(\gamma_i(t)) = t$. Then (see Fig. 8), the angle θ_i between the barrier and the level curve $\{x;\, T^{\Gamma}(x) = t\}$ is determined by

$$\left|\dot{\gamma}_i(t)\right| = \sigma_i(t) = \frac{1}{\sin\theta_i(t)}.$$

  If $\sigma \leq 2$, then only one boundary arc can be constructed, at any given time.

For example, assume that the initial set R_0 is the unit disc. Then the construction of one single arc γ_1 along the edge of the advancing fire front produces a spiraling curve (see Fig. 13). As shown in [6], this curve eventually closes on itself, blocking

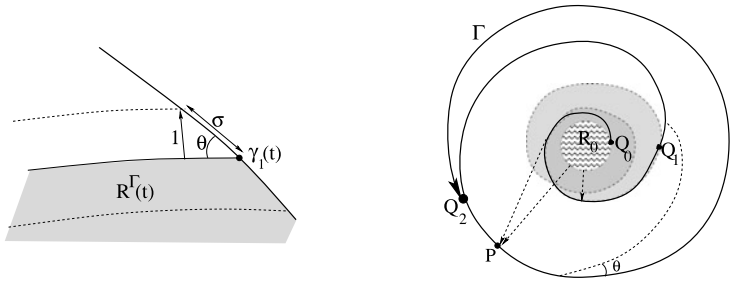


Fig. 13 *Left*: if the fire front advances with unit speed, and the barrier γ_1 is constructed at speed σ, then the angle θ is determined by $\sin\theta = 1/\sigma$. *Right*: a fire starting on the unit disc is encircled by a spiraling barrier. *Shaded areas* denote regions reached by the fire at various times. The first portion of the wall, between Q_0 and Q_1, is exactly a logarithmic spiral. This is a delaying arc. For example, the minimum time needed to reach the point P without crossing Γ is $T^\Gamma(P) > T(P) = |P| - 1$. As proved in [6], if $\sigma > \sigma^\dagger \approx 2.614430844$, then the spiraling barrier eventually closes on itself

the fire, only if the construction speed is $|\dot\gamma_1(t)| = \sigma > \sigma^\dagger = 2.614430844\ldots$. More precisely, the constant $\sigma^\dagger$ is here defined as

$$\sigma^\dagger \doteq \max_{\lambda>0} \frac{1}{\cos\theta(\lambda)},$$

where $\lambda \mapsto \theta(\lambda) \in [2\pi, 5\pi/2]$ is the function implicitly defined by

$$e^{\lambda\theta}\cos\theta - 1 = \lambda e^{\lambda\theta}\sin\theta.$$

In particular, this analysis shows that a strategy constructing a single spiraling wall cannot block the fire if $\sigma \le 2$.

For additional remarks on this problem, and a cash prize for its solution, one may look at the author's web page: http://www.math.psu.edu/bressan/

8.2 Existence of Optimal Strategies

The existence theorem proved in [7, 15] already covers very general situations. However, it relies on one key assumption: every velocity set $F(x)$ should contain a neighborhood of the origin. In other words, the fire should propagate with strictly positive speed in every direction. This assumption guarantees the Lipschitz continuity of the minimum time function T^Γ, away from the barrier.

It is not known whether Theorem 6, on the existence of optimal strategies, remains valid if we only assume that $0 \in F(x)$ for every $x \in \mathbb{R}^2$. This extension would cover situations where the wind pushes the fire only in one direction. For example, consider the "ice-cream cone" case:

$$F(x) = \big\{(\lambda x_1, \lambda x_2);\, (x_1 - 2)^2 + x_2^2 \le 1, \lambda \in [0, 1]\big\} \subset \mathbb{R}^2.$$

8.3 Sufficient Conditions for Optimality

At the present date, not one single example is known of a blocking strategy which is provably optimal.

In the isotropic case, according to Theorem 9, the barrier Γ^* consisting of an arc of circumference and two arcs of logarithmic spirals is the one which encloses the minimum area, among all simple closed curves satisfying the admissibility condition (16). One conjectures that Γ^* is the global minimizer among all admissible curves, regardless of their topological structure.

In the classical setting of Calculus of Variations and optimal control, sufficient conditions for optimality are obtained by studying the value function [4, 8, 13, 16]. If the state space is finite-dimensional, this value function can be often characterized as the unique (viscosity) solution to a Hamilton-Jacobi PDE.

For a dynamic blocking problem, the "state" of the system at time $t \geq 0$ can be described by the couple $(R^\gamma(t), \gamma(t))$. Here $R^\gamma(t)$ is the set burned by the fire at time t, as in (11), while $\gamma(t)$ is the portion of the barrier constructed up to time t. One can now introduce a value function: $V = V(R_0, \gamma_0)$, defined as the infimum among the costs of all admissible strategies, assuming that at the initial time $t = 0$ the fire is burning on the region R_0 and a barrier is already in place along the rectifiable set γ_0.

We remark that the space of all couples (R_0, γ_0), where $R_0 \subset \mathbb{R}^2$ is Lebesgue measurable and $\gamma_0 \subset \mathbb{R}^2$ is a rectifiable set, does not have the structure of a vector space. A characterization of the value function V in terms of a PDE is out of the question. Yet, it may be of interest to study some properties of the function V, possibly related to a dynamic programming principle.

An alternative approach relies on the observation that the couple $(R^\gamma(t), \gamma(t))$ can be recovered from the truncated minimum time function $x \mapsto \min\{T^\gamma(x), t\}$, which is a Special function of Bounded Variation. One may thus consider a corresponding value function $V(\cdot)$ defined on elements of the functional space SBV.

8.4 Regularity

According to the existence results proved in [7, 15], under the general hypotheses of Theorem 6 the optimal barrier Γ^* is a complete rectifiable set, which can be decomposed as

$$\Gamma^* = \left(\bigcup_{i\geq 1} \Gamma_i\right) \cup \Gamma_0.$$

Here the countably many sets Γ_i are disjoint, compact, and connected, while Γ_0 is a set with 1-dimensional Hausdorff measure $m_1(\Gamma_0) = 0$.

Unfortunately, the above result does not allow us to derive any of the necessary conditions for optimality in [5, 11, 23], which require stronger regularity assumptions. Indeed, at the present date these optimality conditions are known to hold only for an optimal barrier which is the union of finitely many regular arcs.

It would be of interest to close this gap, proving that any optimal barrier Γ^* must satisfy additional regularity conditions. In particular, assuming that the initial set R_0 has a smooth boundary and the cost functions $\alpha(\cdot)$, $\beta(\cdot)$ are smooth, the following questions naturally arise:

- What is the regularity of an optimal barrier? Is Γ^* the union of finitely many $\mathcal{C}^1$ arcs?
- If R_0 is connected, does this imply that the optimal barrier Γ^* is connected?
- Does the optimal barrier contain purely delaying arcs $\gamma_i \subset \Gamma^d \setminus \Gamma^b$?

All the above problems are open even in the case where α, β are constant, and R_0 is a convex set. In particular, no example is known of an optimal barrier containing a purely delaying arc.

Results in this direction would be relevant also toward an understanding of the isotropic blocking problem. Indeed, for a given speed $\sigma \leq 2$, assume that a blocking strategy exists. Then there exists also an optimal blocking strategy, minimizing the total area of the burned region. Hence the search for blocking strategies can be restricted to barriers which satisfy necessary optimality conditions. Of course, this approach is useful only if some a priori regularity of optimal barriers is known, in order to apply the necessary conditions in [5, 11, 23].

Acknowledgements This work was partially supported by NSF through grant DMS 1108702 "Problems of Nonlinear Control".

References

1. L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, London, 2000.
2. L. Ambrosio, A. Colesanti and E. Villa, Outer Minkowski content for some classes of closed sets, *Math. Ann.*, **342** (2008), 727–748.
3. J. P. Aubin and A. Cellina, *Differential Inclusions*, Springer, Berlin, 1984.
4. M. Bardi and I. Capuzzo-Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.
5. A. Bressan, Differential inclusions and the control of forest fires, *J. Differ. Equ.*, **243** (2007), 179–207. (Special volume in honor of A. Cellina and J. Yorke.)
6. A. Bressan, M. Burago, A. Friend, and J. Jou, Blocking strategies for a fire control problem, *Anal. Appl.*, **6** (2008), 229–246.
7. A. Bressan and C. De Lellis, Existence of optimal strategies for a fire confinement problem, *Commun. Pure Appl. Math.*, **62** (2009), 789–830.
8. A. Bressan and B. Piccoli, *Introduction to the Mathematical Theory of Control*, American Institute of Mathematical Sciences, Springfield, 2007.
9. A. Bressan and T. Wang, Equivalent formulation and numerical analysis of a fire confinement problem, *ESAIM Control Optim. Calc. Var.*, **16** (2010), 974–1001.
10. A. Bressan and T. Wang, The minimum speed for a blocking problem on the half plane, *J. Math. Anal. Appl.*, **356** (2009), 133–144.
11. A. Bressan and T. Wang, Global necessary conditions for a dynamic blocking problem. *ESAIM Control Optim. Calc. Var.*, **18** (2012), 124–156.
12. A. Bressan and T. Wang, On the optimal strategy for an isotropic blocking problem. *Calc. Var. PDEs*, **45** (2012), 125–145.

13. L. Cesari, *Optimization Theory and Applications*, Springer, Berlin, 1983.
14. A. Cianchi and N. Fusco, Functions of bounded variation and rearrangements, *Arch. Rational Mech. Anal.*, **165** (2002), 1–40.
15. C. De Lellis and R. Robyr, Hamilton-Jacobi equations with obstacles, *Arch. Rational Mech. Anal.*, **200** (2011), 1051–1073.
16. W. H. Fleming and R. W. Rishel, *Deterministic and Stochastic Optimal Control*, Springer, New York, 1975.
17. V. Mallet, D. E. Keyes, and F. E. Fendell, Modeling wildland fire propagation with level set methods, *Comput. Math. Appl.*, **57** (2009), 1089–1101.
18. G. D. Richards, An elliptical growth model of forest fire fronts and its numerical solution, *Int. J. Numer. Meth. Eng.*, **30** (1990), 1163–1179.
19. R. C. Rothermel, A mathematical model for predicting fire spread in wildland fuels, USDA Forest Service, Intermountain Forest and Range Experiment Station, Research Paper INT-115, Ogden, Utah, USA, 1972.
20. J. A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge University Press, Cambridge, 1999.
21. A. L. Sullivan, Wildland surface fire spread modelling, 1990–2007, *Int. J. Wildland Fire*, **18** (2009), 349–403.
22. R. Vinter, *Optimal Control*, Birkhäuser, Boston, 2000.
23. T. Wang, Optimality conditions for a blocking strategy involving delaying arcs, *J. Optim. Theory Appl.*, **152** (2012), 307–333.

Inverse Lax–Wendroff Procedure for Numerical Boundary Conditions of Hyperbolic Equations: Survey and New Developments

Sirui Tan and Chi-Wang Shu

Abstract In this paper, we give a survey and discuss new developments and computational results for a high order accurate numerical boundary condition based on finite difference methods for solving hyperbolic equations on Cartesian grids, while the physical domain can be arbitrarily shaped. The challenges result from the wide stencil of the high order interior scheme and the fact that the physical boundary does not usually coincide with grid lines. There are two main ingredients of the method. The first one is an inverse Lax-Wendroff procedure for inflow boundary conditions and the other one is a robust and high order accurate extrapolation for outflow boundary conditions. The method is high order accurate, stable under standard CFL conditions determined by the interior schemes, and easy to implement. We show applications in simulating interactions between compressible inviscid flows and rigid (static or moving) boundaries.

Keywords Numerical boundary conditions · Hyperbolic conservation laws · Cartesian mesh · Inverse Lax-Wendroff procedure · Extrapolation

1 Introduction

We consider high order accurate finite difference methods for solving hyperbolic conservation laws involving complex static or moving geometries. For such problems, body-fitted meshes which conform to the geometry are often used due to the ease of imposing boundary conditions. In finite difference methods, body-fitted grids are usually generated by a curvilinear transformation that maps the physical domain to a computational domain. However, grid generation could be difficult for geometrically complicated domains and could also be very time-consuming for moving geometries, since the grids have to be updated every few time steps. On the

S. Tan · C.-W. Shu (✉)
Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
e-mail: shu@dam.brown.edu

S. Tan
e-mail: sirui@dam.brown.edu

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5_3,

other hand, fixed Cartesian grids can be used to discretize the physical domain regardless of its geometry. The advantage is that the grid generation is trivial and the numerical schemes are usually more efficient than those on body-fitted grids. The biggest challenge with Cartesian grids is however to accurately impose the boundary conditions while retaining the stability under the standard CFL conditions determined by the interior schemes. The challenge mainly results from two facts. First, the physical boundary does not usually coincide with grid lines and can intersect the Cartesian grids in an arbitrary fashion. Secondly, a high order interior scheme needs a suitable treatment of *several* ghost points near the boundary because of its wide numerical stencil.

There are many successful numerical methods based on Cartesian grids. For example, the immersed boundary method (IBM) introduced by Peskin [29] is widely used to solve incompressible flows in complicated time-varying geometries. See also [28] for an overview of the method and its applications. The IBM is extended for compressible viscous flows in [6, 7, 11]. An immersed interface method (IIM) is developed for elliptic equations in [25, 26] and for streamfunction-vorticity equations in [24]. We would like to emphasize that high order numerical boundary conditions for hyperbolic equations are somewhat more difficult than elliptic or convection-diffusion type equations mentioned above due to the possible presence of strong discontinuities near the boundaries. The boundary treatment should be robust for such cases when strong shock waves reflect off rigid boundaries.

To solve hyperbolic equations in complex static or moving geometries with Cartesian grids, most methods in the literature are based on finite volume schemes. The difficulty mainly comes from the "small-cell" problem. Namely, one obtains irregular cut cells near the boundary, which may be orders of magnitude smaller than the regular grid cells, leading to a severe time step restriction. The so-called h-box method [4, 15] is developed to overcome this problem. The basic idea is to approximate numerical fluxes at the interface of a small cell based on initial values specified over regions of a regular length h. For one-dimensional advection equations, the h-box method is shown to be conservative and stable on arbitrary irregular grids in [4]. Numerical results confirm the same conclusion for Euler equations in [15]. Another method to avoid the small-cell problem for Euler equations is to maintain cut boundary cells as whole (ghost) cells and obtain ghost cell values by reflecting locally the flow field around the boundary which is approximated by straight lines [10]. This method is stable and applicable to any finite volume method, but conservation at the boundary is violated although this error may be relatively small. For Euler equations involving moving geometries, an idea similar to [10] is developed in [9] but the ghost cell values are obtained by a mirror flow extrapolation. A cell merging technique combined with dimension splitting is developed by Falcovitz et al. [8]. Shyue [31] proposes a moving-boundary tracking algorithm based on finite volume wave-propagation method. The methods is stable even if there are very small cut cells near the tracked interface. Arienti et al. [2] develop an Eulerian-Lagrangian coupling scheme. Hu et al. [16] develop a conservative interface method based on the level set technique. In terms of accuracy, all the finite volume schemes mentioned above are at most second order. In particular, the errors at the boundaries sometimes fall short of second order, see numerical examples in [2, 10, 16].

For methods based on the finite difference formulation, a second order accurate Cartesian embedded boundary method is developed to solve the wave equations with Dirichlet or Neumann boundary conditions in [19–21] and to solve hyperbolic conservation laws in [32]. The basic idea is to assign values to ghost points *outside* the domain by using extrapolation. To avoid oscillations near shock waves, slope limiters are combined with extrapolation in [32]. The method in [32] is essentially based on a three point interior scheme so that only points just outside the boundary are ghost points. The feasibility and effectiveness to generalize this method to higher order remain to be demonstrated. Recently, Appelö and Petersson [1] modify the embedded boundary method in [19–21] by assigning values to boundary points *inside* the domain via interpolation. This method is fourth order accurate and is based on a compact finite difference scheme.

A Lax-Wendroff type boundary condition procedure is introduced by Huang et al. [17] for solving static Hamilton-Jacobi equations with a third order finite difference method. It is extended to fifth order in [36] and to discontinuous Galerkin method in [37] for the same type of problems. The idea is to repeatedly use the partial differential equations (PDEs) to write the normal derivatives to the inflow boundary in terms of the tangential derivatives of the given boundary condition. With these normal derivatives, we can obtain accurate values of ghost points by a Taylor expansion from a point located on the boundary. Tan and Shu [33, 34] have systematically extended this procedure to solve time dependent hyperbolic conservation laws involving complex static or moving geometries with high order finite difference schemes. For time dependent problems, the boundary treatment procedure is in essence repeatedly using the PDEs to convert normal spatial derivatives to time and tangential derivatives of the given boundary condition. It is in some sense an inverse to the standard Lax-Wendroff procedure [23], in which the PDEs are repeatedly used to convert time derivatives to spatial derivatives when discretizing the PDEs in time with high order accuracy. We therefore refer to our method as the inverse Lax-Wendroff procedure (ILW). This procedure is first proposed by Goldberg and Tadmor [12, 13] for analyzing numerical boundary conditions of linear hyperbolic equations in one dimension with boundaries aligned with grid lines. Lombard et al. [27] applied a similar idea to arbitrary-shaped free boundaries in finite difference schemes for linear elastic waves. Thus [33, 34] can also be regarded as an extension of [12, 13, 27] to nonlinear hyperbolic problems. In particular, strong discontinuities near the boundaries, which are absent in linear elastic waves, are handled by a high order weighted essentially non-oscillatory (WENO) type extrapolation to prevent overshoot or undershoot. However, the algebra of the ILW procedure can be very heavy for fully nonlinear systems of equations in two dimensions (2D). For this reason, in [33, 34] we only implement a third order boundary treatment for 2D Euler equations, although our method was designed to achieve arbitrarily high order of accuracy. To address this difficulty, a simplified implementation is developed in [35] for hyperbolic systems with or without source terms in static geometries.

In this paper, we first review the ILW procedure for high order numerical boundary conditions of hyperbolic conservation laws in static geometries in Sects. 2–4. We only discuss the simplified implementation, which is proposed in [35] for problems

in static geometries. Then we extend the simplified implementation for 2D Euler equations in moving geometries in Sect. 5. Some numerical examples computed by a third order boundary treatment in [34] are now tested by a fifth order boundary treatment. Finally, concluding remarks are given in Sect. 6.

2 Interior Finite Difference Schemes

We consider hyperbolic conservation laws possibly with source terms for $\mathbf{U} = \mathbf{U}(x, y, t) \in \mathbb{R}^2$

$$\begin{cases} \mathbf{U}_t + \mathbf{F}(\mathbf{U})_x + \mathbf{G}(\mathbf{U})_y = \mathbf{S}(\mathbf{U}) & (x, y) \in \Omega(t),\ t > 0, \\ \mathbf{U}(x, y, 0) = \mathbf{U}_0(x, y) & (x, y) \in \bar{\Omega}(t), \end{cases} \tag{1}$$

on a bounded domain $\Omega(t)$ with appropriate boundary conditions prescribed on $\Gamma(t) = \partial\Omega(t)$ at time t. We assume we always have an analytical expression for the geometry of $\Gamma(t)$. $\Omega(t)$ is covered by a uniform Cartesian mesh with mesh size $\Delta x = \Delta y = h$, but the boundary $\Gamma(t)$ does not need to coincide with any grid lines. The semi-discrete approximation of (1) is given by

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{U}_{i,j}(t) = &-\frac{1}{h}(\hat{\mathbf{F}}_{i+1/2,j} - \hat{\mathbf{F}}_{i-1/2,j}) \\ &-\frac{1}{h}(\hat{\mathbf{G}}_{i,j+1/2} - \hat{\mathbf{G}}_{i,j-1/2}) + \mathbf{S}\big(\mathbf{U}_{i,j}(t)\big), \end{aligned} \tag{2}$$

where $\hat{\mathbf{F}}_{i+1/2,j}$ and $\hat{\mathbf{G}}_{i,j+1/2}$ are the numerical fluxes. We use a third order total variation diminishing Runge-Kutta (RK) method [30] to integrate the system of ordinary differential equations (2) in time

$$\begin{aligned} \mathbf{U}_{i,j}^{(1)} &= \mathbf{U}_{i,j}^{n} + \Delta t\mathcal{L}\big(\mathbf{U}_{i,j}^{n}\big), \\ \mathbf{U}_{i,j}^{(2)} &= \frac{3}{4}\mathbf{U}_{i,j}^{n} + \frac{1}{4}\mathbf{U}_{i,j}^{(1)} + \frac{1}{4}\Delta t\mathcal{L}\big(\mathbf{U}_{i,j}^{(1)}\big), \\ \mathbf{U}_{i,j}^{n+1} &= \frac{1}{3}\mathbf{U}_{i,j}^{n} + \frac{2}{3}\mathbf{U}_{i,j}^{(2)} + \frac{2}{3}\Delta t\mathcal{L}\big(\mathbf{U}_{i,j}^{(2)}\big), \end{aligned} \tag{3}$$

where $\mathcal{L}(\cdot)$ is the operator defined by the right-hand side of (2).

Special care must be taken when we impose a *time dependent* boundary condition $\mathbf{g}(t)$ in the two intermediate stages of the RK method (3). Following the idea in [5], we can easily show that for the hyperbolic equations (1) with source terms the following match of time levels at the boundary maintains the third order accuracy

of (3)

$$\begin{aligned}
\mathbf{U}^n &\sim \mathbf{g}(t_n), \\
\mathbf{U}^{(1)} &\sim \mathbf{g}(t_n) + \Delta t \mathbf{g}'(t_n), \\
\mathbf{U}^{(2)} &\sim \mathbf{g}(t_n) + \frac{1}{2}\Delta t \mathbf{g}'(t_n) + \frac{1}{4}\Delta t^2 \mathbf{g}''(t_n).
\end{aligned} \tag{4}$$

For simplicity, we denote the boundary conditions for all three stages at time level $t = t_n$ by $\mathbf{g}(t_n)$, although $\mathbf{g}(t_n)$ is actually different for each stage according to (4).

Even though the boundary treatment method discussed in this paper is applicable to general finite difference schemes, we use the fifth order finite difference WENO scheme with the Lax-Friedrichs flux splitting [18] to form the numerical fluxes $\hat{\mathbf{F}}_{i+1/2,j}$ and $\hat{\mathbf{G}}_{i,j+1/2}$ in (2). The scheme requires a seven point stencil in both x and y directions. Near $\Gamma(t)$ where the numerical stencil is partially outside of $\Omega(t)$, up to three ghost points are needed in each direction. In some cases of moving boundaries, up to four ghost points may be needed in each direction, see Sect. 5 for details. We concentrate on how to define these values of the ghost points at time level $t = t_n$ in the rest of the paper.

3 1D Scalar Conservation Laws

To illustrate the essential idea of the ILW procedure, we use 1D scalar conservation laws as an example

$$\begin{cases}
u_t + f(u)_x = 0 & x \in (-1, 1),\ t > 0, \\
u(-1, t) = g(t) & t > 0, \\
u(x, 0) = u_0(x) & x \in [-1, 1].
\end{cases}$$

We assume $f'(u(-1, t)) \geqslant \alpha > 0$ and $f'(u(1, t)) \geqslant \alpha > 0$ for $t > 0$. This assumption guarantees the left boundary $x = -1$ is an inflow boundary where a boundary condition is needed and the right boundary $x = 1$ is an outflow boundary where no boundary condition is needed.

Let us discretize the interval $(-1, 1)$ by a uniform mesh

$$-1 + h/2 = x_0 < x_1 < \cdots < x_N = 1 - h/2.$$

Notice that both x_0 and x_N are not located on the boundary, which is chosen this way on purpose since it is usually not possible to align boundary with grid points in a 2D domain with complex geometry. We assume $u_0, \ldots, u_N$ have been updated from time level t_{n-1} to time level t_n. Here we suppress the t_n dependence without causing any confusion. We describe our fifth order boundary treatment, which is the same order as our interior scheme.

3.1 Robust and High Order Extrapolation for Outflow Boundary Conditions

At the outflow boundary $x = 1$, values of ghost points $u_{N+1}, \ldots, u_{N+3}$ are approximated by a fourth order Taylor expansion

$$u_j = \sum_{k=0}^{4} \frac{(x_j - 1)^k}{k!} u_R^{*(k)}, \quad j = N+1, \ldots, N+3,$$

where $u_R^{*(k)}$ is a $(5-k)$th order approximation of $\frac{\partial^k u}{\partial x^k}|_{x=1,t=t_n}$. At the outflow boundary, $u_R^{*(k)}$ should be imposed by the values of the interior points, $u_0, \ldots, u_N$, because of the outgoing characteristics, even if a boundary condition is improperly prescribed. If u is smooth near the boundary, $u_R^{*(k)}$ can be easily obtained by

$$u_R^{*(k)} = \left. \frac{\mathrm{d}^k p_4(x)}{\mathrm{d}x^k} \right|_{x=1},$$

where $p_4(x)$ is a Lagrange polynomial of degree 4 interpolating u_j, $j = N-4, \ldots, N$. However, if a discontinuity appears in the stencil $\{x_{N-4}, \ldots, x_N\}$, Lagrange extrapolation may lead to severe overshoot or undershoot near the discontinuity. In this situation, we prefer a possibly lower order accurate but more robust extrapolation. The fifth order WENO type extrapolation is developed for this purpose. It is described in detail in Sect. 2.2 of [35] and omitted here.

3.2 The ILW Procedure for Inflow Boundary Conditions

At the inflow boundary $x = -1$, we construct a fifth order approximation of ghost point values $u_{-3}, \ldots, u_{-1}$ by a Taylor expansion

$$u_j = \sum_{k=0}^{4} \frac{(x_j + 1)^k}{k!} u_L^{*(k)}, \quad j = -3, \ldots, -1, \tag{5}$$

where $u_L^{*(k)}$ is a $(5-k)$th order approximation of $\frac{\partial^k u}{\partial x^k}|_{x=-1,t=t_n}$. We impose $u_L^{*(0)} = g(t_n)$. To obtain the approximation of the spatial derivative $\frac{\partial u}{\partial x}|_{x=-1,t=t_n}$, we utilize the PDE

$$u_t + f'(u) u_x = 0$$

and evaluate it at $x = -1$, $t = t_n$. We impose

$$\begin{aligned} u_L^{*(1)} &= -\frac{u_t(-1, t_n)}{f'(u(-1, t_n))} \\ &= -\frac{g'(t_n)}{f'(g(t_n))}, \end{aligned} \tag{6}$$

where $f'(g(t_n))$ is bounded away from zero by the assumption that $x = -1$ is an inflow boundary.

Notice that we obtain the spatial derivative from the time derivative by using the PDE in (6). This idea comes from the original Lax-Wendroff scheme [23]. Since we convert spatial derivatives to time derivatives, our method is called the *inverse* Lax-Wendroff procedure. The time derivatives can be obtained by either using the analytical derivatives of $g(t)$ if available or numerical differentiation. In the case of discontinuities going through the boundary, $g(t)$ is discontinuous. The stencil used for numerical differentiation should not contain any discontinuity. For example, an essentially non-oscillatory (ENO) procedure [14] or a WENO procedure [18] can be used for this numerical differentiation.

Higher order spatial derivatives $u_L^{*(k)}$, $k \geq 2$, can be obtained by repeated use of the PDE, see Sect. 2.1 of [33] for the formula of $k = 2$. The algebra of converting higher order derivatives can be very heavy if the PDE is complicated, which is usually the case if we consider 2D fully nonlinear systems (1). In [35], we investigate each term in the Taylor expansion (5) and find that only $u_L^{*(1)}$ is crucial to be implemented by the ILW procedure to ensure stability. In other words, $u_L^{*(k)}$, $k = 2, 3, 4$, can be simply obtained by the extrapolation discussed in Sect. 3.1. We remark that instability is observed in our numerical experiments if the first order spatial derivative $u_L^{*(1)}$ is also obtained by extrapolation. Therefore, the ILW procedure is crucial to ensure stability.

3.3 Computational Results

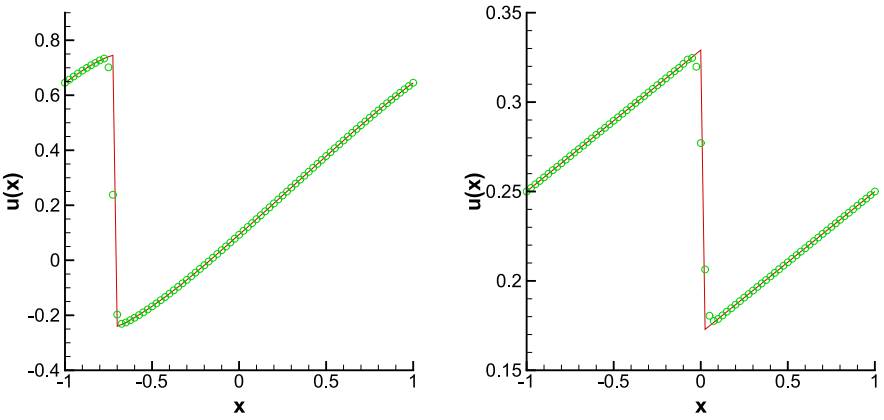
Example 1 We test the Burgers equation

$$\begin{cases} u_t + (\frac{1}{2}u^2)_x = 0 & x \in (-1, 1),\ t > 0, \\ u(x, 0) = 0.25 + 0.5\sin(\pi x) & x \in [-1, 1], \\ u(-1, t) = g(t) & t > 0. \end{cases} \tag{7}$$

The boundary condition $g(t)$ is taken from the exact solution of the initial value problem on $(-1, 1)$ with periodic boundary conditions. For all t, the left boundary $x = -1$ is an inflow boundary and the right boundary $x = 1$ is an outflow boundary. At $t = 0.3$, we have a smooth solution. The errors are listed in Table 1. We achieve the designed fifth order accuracy. At $t = 1.1$, a shock is fully developed in the interior of the computational domain. A shock enters the inflow boundary at $t = 8$ and moves to $x = 0$ at $t = 12$. We can see from Fig. 1 that the shock is well captured in both scenarios by our method. This example and later examples when shocks come from the boundary indicate that, even though the inverse Lax-Wendroff procedure is based on the assumption of smoothness of the boundary data, its application to discontinuous but piecewise smooth boundary data coupled with a non-oscillatory internal scheme does not lead to spurious oscillations or instability. Figure 2 illustrates the important role of WENO type extrapolation. The shock is very close to

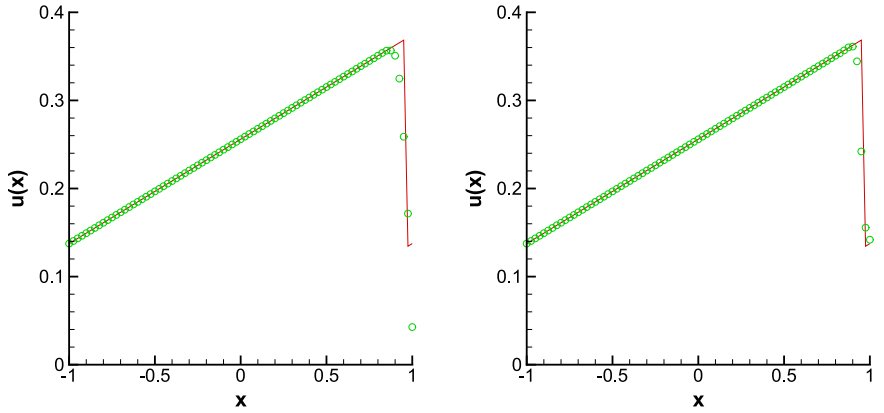
Table 1 Errors of the Burgers equation (7). $\Delta t = O(h^{5/3})$ and $t = 0.3$

h	L^1 error	Order	L^∞ error	Order
1/40	4.73E–06		5.88E–05	
1/80	1.46E–07	5.02	1.58E–06	5.21
1/160	3.48E–09	5.39	2.68E–08	5.89
1/320	7.56E–11	5.53	3.62E–10	6.21
1/640	1.89E–12	5.32	1.97E–11	4.20
1/1280	7.01E–14	4.76	8.00E–13	4.62



(a) t = 1.1 (b) t = 12

Fig. 1 Burgers equation (7). $h = 1/40$ and CFL $= 0.6$. *Solid line*: exact solution; *Symbols*: numerical solution



(a) Lagrange extrapolation (b) WENO type extrapolation

Fig. 2 Burgers equation (7). $h = 1/40$, CFL $= 0.6$, and $t = 7.8$. *Solid line*: exact solution; *Symbols*: numerical solution

the outflow boundary in this case. Lagrange extrapolation gives severe undershoot, while WENO type extrapolation maintains excellent performance.

4 2D Euler Equations in Static Geometries

We consider 2D compressible Euler equations in static geometries

$$\mathbf{U}_t + \mathbf{F}(\mathbf{U})_x + \mathbf{G}(\mathbf{U})_y = 0, \quad (x, y) \in \Omega,\ t > 0, \tag{8}$$

where

$$\mathbf{U} = \begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{pmatrix} = \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix}, \qquad \mathbf{F}(\mathbf{U}) = \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(E+p) \end{pmatrix}, \qquad \mathbf{G}(\mathbf{U}) = \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(E+p) \end{pmatrix},$$

with appropriate boundary conditions and initial conditions. ρ, u, v, p and E describe the density, x-velocity, y-velocity, pressure, and total energy, respectively. The equation of state has the form

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho\left(u^2 + v^2\right), \tag{9}$$

where the ratio of specific heats $\gamma = 1.4$ for air at ordinary temperatures.

4.1 General Framework

We assume the values of all the grid points inside Ω have been updated from time level t_{n-1} to time level t_n. For a ghost point $P = (x_i, y_j)$, we find a point $P_0 = (x_0, y_0) = \mathbf{x}_0$ on the boundary Γ such that the normal $\mathbf{n}(\mathbf{x}_0)$ at P_0 goes through P. The sign of the normal $\mathbf{n}(\mathbf{x}_0)$ is chosen in such a way that it is positive if it points to the exterior of Ω. The point P_0 and the normal $\mathbf{n}(\mathbf{x}_0)$ can be obtained analytically, since we assume we have an explicit expression for the geometry of Γ. We set up a local coordinate system at P_0 by

$$\begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \mathbf{T} \begin{pmatrix} x \\ y \end{pmatrix}, \tag{10}$$

where θ is the angle between the normal $\mathbf{n}(\mathbf{x}_0)$ and the x-axis and $\mathbf{T}$ is a rotational matrix. The $\hat{x}$-axis then points in the same direction as $\mathbf{n}(\mathbf{x}_0)$ and the $\hat{y}$-axis points in the tangential direction, see Fig. 3. In this local coordinate system, the Euler equations (8) are written as

$$\hat{\mathbf{U}}_t + \mathbf{F}(\hat{\mathbf{U}})_{\hat{x}} + \mathbf{G}(\hat{\mathbf{U}})_{\hat{y}} = 0, \tag{11}$$

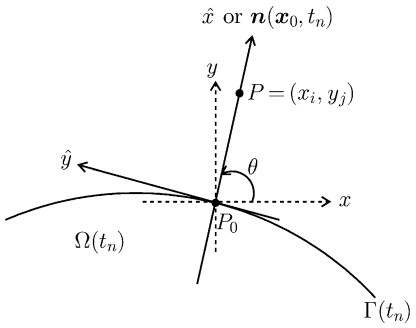


Fig. 3 The local coordinate system (10). For static geometries, t_n dependence can be suppressed

where

$$\hat{\mathbf{U}} = \begin{pmatrix} \hat{U}_1 \\ \hat{U}_2 \\ \hat{U}_3 \\ \hat{U}_4 \end{pmatrix} = \begin{pmatrix} \rho \\ \rho\hat{u} \\ \rho\hat{v} \\ E \end{pmatrix}, \qquad \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix} = \mathbf{T} \begin{pmatrix} u \\ v \end{pmatrix}.$$

For a fifth order boundary treatment, the value of the ghost point P is imposed by the Taylor expansion

$$(\hat{U}_m)_{i,j} = \sum_{k=0}^{4} \frac{\Delta^k}{k!} \hat{U}_m^{*(k)}, \quad m = 1, \ldots, 4, \tag{12}$$

where Δ is the $\hat{x}$-coordinate of P and $\hat{U}_m^{*(k)}$ is a $(5-k)$th order approximation of the normal derivative $\frac{\partial^k \hat{U}_m}{\partial \hat{x}^k}|_{(x,y)=\mathbf{x}_0, t=t_n}$. To decide the inflow and outflow boundary conditions at time t, we need a local characteristic decomposition of (11). We assume $\hat{\mathbf{U}}_0$ is the value of a grid point nearest to P_0 among all the grid points inside Ω. We denote the Jacobian matrix of the normal flux by

$$\mathbf{A}_\perp(\hat{\mathbf{U}}_0) = \left.\frac{\partial \mathbf{F}(\hat{\mathbf{U}})}{\partial \hat{\mathbf{U}}}\right|_{\hat{\mathbf{U}}=\hat{\mathbf{U}}_0}.$$

$\mathbf{A}_\perp(\hat{\mathbf{U}}_0)$ has four eigenvalues $\lambda_1 = \hat{u}_0 - c_0$, $\lambda_2 = \lambda_3 = \hat{u}_0$, $\lambda_4 = \hat{u}_0 + c_0$ and a complete set of left eigenvectors $\mathbf{l}_m(\hat{\mathbf{U}}_0)$, $m = 1, \ldots, 4$, which forms a matrix

$$\mathbf{L}(\hat{\mathbf{U}}_0) = \begin{pmatrix} \mathbf{l}_1(\hat{\mathbf{U}}_0) \\ \mathbf{l}_2(\hat{\mathbf{U}}_0) \\ \mathbf{l}_3(\hat{\mathbf{U}}_0) \\ \mathbf{l}_4(\hat{\mathbf{U}}_0) \end{pmatrix} = \begin{pmatrix} l_{1,1} & l_{1,2} & l_{1,3} & l_{1,4} \\ l_{2,1} & l_{2,2} & l_{2,3} & l_{2,4} \\ l_{3,1} & l_{3,2} & l_{3,3} & l_{3,4} \\ l_{4,1} & l_{4,2} & l_{4,3} & l_{4,4} \end{pmatrix}.$$

The number of prescribed boundary conditions depends on the signs of λ_m, $m = 1, \ldots, 4$. For definiteness, we assume $\lambda_1 < 0$ and $\lambda_4 > \lambda_2 = \lambda_3 > 0$. Thus one boundary condition is needed at P_0. For example, the normal momentum is prescribed $\hat{U}_2(\mathbf{x}_0, t) = g_2(t)$.

The local characteristic variables V_m at grid points near P_0 are defined by

$$(V_m)_{\mu,\nu} = \mathbf{l}_m(\hat{\mathbf{U}}_0)\hat{\mathbf{U}}_{\mu,\nu}, \quad m = 1, \ldots, 4, \qquad (x_\mu, y_\nu) \in \mathcal{E}_{i,j}, \tag{13}$$

where $\mathcal{E}_{i,j} \subset \Omega$ is a set of grid points used for extrapolation. We extrapolate $(V_m)_{\mu,\nu}$ to P_0 and denote the extrapolated kth order $\hat{x}$-derivative of V_m by $V_m^{*(k)}$, $k = 0, \ldots, 4$. The choice of $\mathcal{E}_{i,j}$ and the fifth order WENO type extrapolation in 2D are described in detail in Sect. 2.4 of [35] and omitted here.

We solve $\hat{U}_m^{*(0)}$ by a linear system of equations

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ l_{2,1} & l_{2,2} & l_{2,3} & l_{2,4} \\ l_{3,1} & l_{3,2} & l_{3,3} & l_{3,4} \\ l_{4,1} & l_{4,2} & l_{4,3} & l_{4,4} \end{pmatrix} \begin{pmatrix} \hat{U}_1^{*(0)} \\ \hat{U}_2^{*(0)} \\ \hat{U}_3^{*(0)} \\ \hat{U}_4^{*(0)} \end{pmatrix} = \begin{pmatrix} g_2(t_n) \\ V_2^{*(0)} \\ V_3^{*(0)} \\ V_4^{*(0)} \end{pmatrix}. \tag{14}$$

Here the first equation is the prescribed boundary condition. The other equations represent extrapolation of the three outgoing characteristic variables V_m, $m = 2, \ldots, 4$. Next, we use the ILW procedure for $\hat{U}_2$. The second equation of (11) gives us

$$\begin{aligned} \frac{\partial \hat{U}_2}{\partial t} = &-\left(\frac{\gamma-3}{2}\frac{\hat{U}_2^2}{\hat{U}_1^2} + \frac{\gamma-1}{2}\frac{\hat{U}_3^2}{\hat{U}_1^2}\right)\frac{\partial \hat{U}_1}{\partial \hat{x}} - (3-\gamma)\frac{\hat{U}_2}{\hat{U}_1}\frac{\partial \hat{U}_2}{\partial \hat{x}} \\ &+ (\gamma-1)\frac{\hat{U}_3}{\hat{U}_1}\frac{\partial \hat{U}_3}{\partial \hat{x}} - (\gamma-1)\frac{\partial \hat{U}_4}{\partial \hat{x}} - \frac{\partial}{\partial \hat{y}}\left(\frac{\hat{U}_2\hat{U}_3}{\hat{U}_1}\right). \end{aligned}$$

At the boundary, the left-hand side of the above equation is the known function $g_2'(t)$. The tangential derivative on the right-hand side can be computed by numerical differentiation, since we have obtained $\hat{U}_m^{*(0)}$ of all the ghost points. Thus $\hat{U}_m^{*(1)}$ can be solved by the linear system

$$\mathbf{A}^{*(0)} \begin{pmatrix} \hat{U}_1^{*(1)} \\ \hat{U}_2^{*(1)} \\ \hat{U}_3^{*(1)} \\ \hat{U}_4^{*(1)} \end{pmatrix} = \begin{pmatrix} -g_2'(t_n) - \frac{\partial}{\partial \hat{y}}\left(\frac{\hat{U}_2^{*(0)}\hat{U}_3^{*(0)}}{\hat{U}_1^{*(0)}}\right) \\ V_2^{*(1)} \\ V_3^{*(1)} \\ V_4^{*(1)} \end{pmatrix}, \tag{15}$$

where

$$\mathbf{A}^{*(0)} = \begin{pmatrix} \frac{\gamma-3}{2}\left(\frac{\hat{U}_2^{*(0)}}{\hat{U}_1^{*(0)}}\right)^2 + \frac{\gamma-1}{2}\left(\frac{\hat{U}_3^{*(0)}}{\hat{U}_1^{*(0)}}\right)^2 & (3-\gamma)\frac{\hat{U}_2^{*(0)}}{\hat{U}_1^{*(0)}} & (1-\gamma)\frac{\hat{U}_3^{*(0)}}{\hat{U}_1^{*(0)}} & \gamma-1 \\ l_{2,1} & l_{2,2} & l_{2,3} & l_{2,4} \\ l_{3,1} & l_{3,2} & l_{3,3} & l_{3,4} \\ l_{4,1} & l_{4,2} & l_{4,3} & l_{4,4} \end{pmatrix}.$$

As discussed in Sect. 3.2, $\hat{U}_m^{*(k)}$, $k = 2, 3, 4$, can be simply obtained by extrapolation

$$\mathbf{L}(\hat{\mathbf{U}}_0)\begin{pmatrix}\hat{U}_1^{*(k)}\\ \hat{U}_2^{*(k)}\\ \hat{U}_3^{*(k)}\\ \hat{U}_4^{*(k)}\end{pmatrix} = \begin{pmatrix}V_1^{*(k)}\\ V_2^{*(k)}\\ V_3^{*(k)}\\ V_4^{*(k)}\end{pmatrix}. \tag{16}$$

4.2 No-Penetration Boundary Condition

An important class of boundary conditions for Euler equations is the no-penetration boundary condition at rigid boundaries, i.e., $\hat{u} = 0$ or $\hat{U}_2 = 0$. In this case, the eigenvalues $\lambda_1 \approx -c_0 < 0$, $\lambda_4 \approx c_0 > 0$ and $\lambda_2 = \lambda_3 \approx 0$. Since only one boundary condition is prescribed, we consider V_m, $m = 2, \ldots, 4$, to be outgoing and V_1 to be ingoing, which falls into the same case as discussed in Sect. 4.1. The first equation of (14) gives us $\hat{U}_2^{*(0)} = 0$. Then the first equation of (15) reduces to

$$\frac{\gamma-1}{2}\left(\frac{\hat{U}_3^{*(0)}}{\hat{U}_1^{*(0)}}\right)^2 \hat{U}_1^{*(1)} + (1-\gamma)\frac{\hat{U}_3^{*(0)}}{\hat{U}_1^{*(0)}}\hat{U}_3^{*(1)} + (\gamma-1)\hat{U}_4^{*(1)} = \frac{(\hat{U}_3^{*(0)})^2}{R\hat{U}_1^{*(0)}}, \tag{17}$$

where R is the radius of curvature of Γ at P_0. Notice that there is no tangential derivative in (17). Therefore, we do not need to do any numerical differentiation, which simplifies the implementation.

4.3 Algorithm Flow Chart

We now summarize our fifth order boundary treatment for the 2D Euler equations (8). We assume the values of all the grid points inside Ω have been updated from time level t_{n-1} to time level t_n. Our goal is to impose the value of $(\hat{U}_m)_{i,j}$, $m = 1, \ldots, 4$, for each ghost point (x_i, y_j) at time level $t = t_n$.

1. For each ghost point (x_i, y_j), we do the following three steps:

   - Decide the local coordinate system (10). Compute the eigenvalues $\lambda_m(\hat{\mathbf{U}}_0)$ and left eigenvectors $\mathbf{l}_m(\hat{\mathbf{U}}_0)$ of the Jacobian matrix $\mathbf{A}_\perp(\hat{\mathbf{U}}_0)$ for $m = 1, \ldots, 4$. Decide the prescribed inflow boundary conditions $g_m(t)$ according to the signs of $\lambda_m(\hat{\mathbf{U}}_0)$.
   - Form the local characteristic variables $(V_m)_{\mu,\nu}$, $(x_\mu, y_\nu) \in \mathcal{E}_{i,j}$, as in (13). Extrapolate $(V_m)_{\mu,\nu}$ to the boundary to obtain $V_m^{*(k)}$, $k = 0, \ldots, 4$, with fifth order WENO type extrapolation. See Sect. 2.4 of [35] for details of the 2D extrapolation.

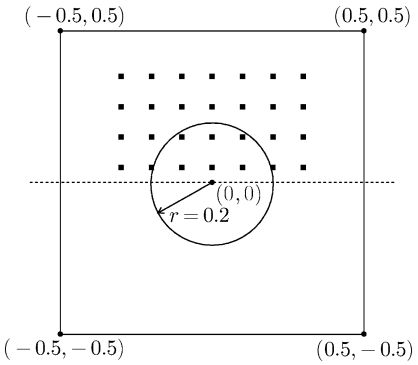


Fig. 4 Physical domain of shock reflection from a cylinder. The *square points* indicate some of the grid points near the cylinder. Illustrative sketch, not to scale

- Solve for $\hat{U}_m^{*(0)}$, $m = 1, \ldots, 4$, by the prescribed boundary conditions and extrapolated values $V_m^{*(0)}$, such as (14).

2. For each ghost point (x_i, y_j), use the ILW procedure to write the first derivative of $g_m(t)$ as a linear combination of first normal derivatives plus tangential derivatives. Together with the extrapolation equations, form a linear system with $\hat{U}_m^{*(1)}$ as unknowns, such as (15). Solve for $\hat{U}_m^{*(1)}$, $m = 1, \ldots, 4$. For $k = 2, 3, 4$, solve for $\hat{U}_m^{*(k)}$ by extrapolation equations (16) where the ILW procedure is not needed.
3. Impose the values of the ghost points by the Taylor expansion (12).

If no-penetration boundary condition is considered at rigid boundaries, then the first equation of (15) is replaced by (17) in Step 2 with other steps unchanged.

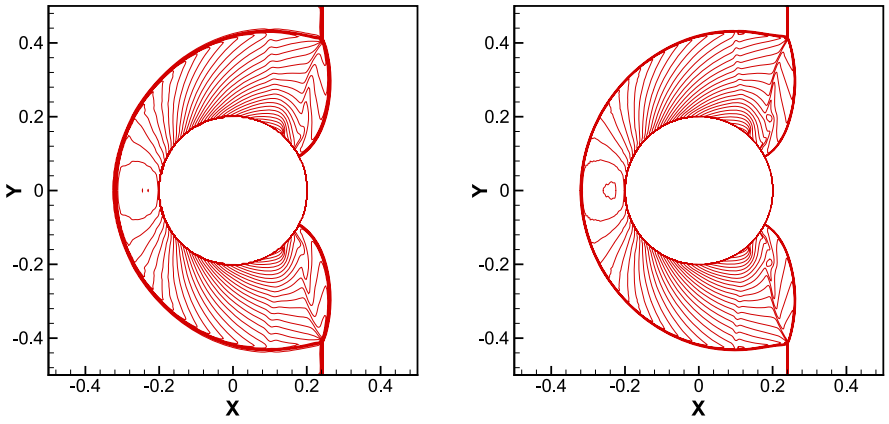
We finally remark that this algorithm can be easily extended for boundary treatment of Euler equations with source terms representing chemical reactions, see [35].

4.4 Computational Results

We provide one numerical example here. More results can be found in Sect. 3.2 of [35].

Example 2 We consider a shock reflection from a circular cylinder. The physical domain is $[-0.5, 0.5] \times [-0.5, 0.5]$. The cylinder is located at $(0, 0)$ and has a radius of 0.2. Initially a Mach 3 shock is located at $x = -0.3$. The state in front of the shock is $\rho = 1.4$, $u = v = 0$, $p = 1$. This problem is considered by Forrer and Jeltsch [10]. Helzel et al. [15] test the same problem but with a Mach 2 shock.

The computational domain is the upper half of the physical domain due to the symmetry of the problem, see Fig. 4. Figure 4 also shows some of the grid points near the cylinder which indicate the boundary cuts the grid in an arbitrary fashion. We apply our fifth order boundary treatment at the surface of the cylinder and take CFL number to be 0.6. Figure 5 shows the density contours at $t = 0.18$ with $h =$



(a) h = 1/320 (b) h = 1/640

Fig. 5 Density contours in Example 2. 39 contours from 0 to 13

$1/320$ and $h = 1/640$. The density contour plot with $h = 1/320$ in Fig. 5(a) agrees well with the result computed on a mesh with $h = 1/300$ in Fig. 11 of [10].

5 2D Euler Equations in Moving Geometries

In this section, we consider 2D Euler equations in moving geometries, i.e., the domain $\Omega(t)$ varies with time t. To describe the boundary conditions, we let $\mathbf{X}_b(\mathbf{a}, t)$ represent the position vector (in Eulerian coordinates) of a point $\mathbf{a}$ on $\Gamma(t)$. Here $\mathbf{a}$ is the Lagrangian coordinate of the point determined by the condition $\mathbf{X}_b(\mathbf{a}, 0) = \mathbf{a}$. We mainly consider rigid bodies moving at a prescribed motion. Namely, $\mathbf{X}_b(\mathbf{a}, t)$ and thus $\Omega(t)$ are explicitly given. The no-penetration boundary condition for inviscid flows is then

$$\mathbf{u}(\mathbf{x}, t) \cdot \mathbf{n}(\mathbf{x}, t) = \mathbf{V}_b(t) \cdot \mathbf{n}(\mathbf{x}, t), \quad \text{for all } \mathbf{x} = \mathbf{X}_b(\mathbf{a}, t) \in \Gamma(t), \tag{18}$$

where $\mathbf{V}_b(t) = \frac{\partial \mathbf{X}_b}{\partial t}$ is the prescribed velocity. Notice that $\mathbf{V}_b(t)$ is independent of $\mathbf{a}$ for rigid bodies. The normal $\mathbf{n}(\mathbf{x}, t)$ is defined in the same way as in Sect. 4.1, see Fig. 3. If the motion of a rigid body is induced by the fluid, the acceleration can be expressed as

$$\frac{\partial^2 \mathbf{X}_b}{\partial t^2} = -\frac{1}{M_b} \int_{\Gamma(t)} p\mathbf{n} ds, \tag{19}$$

where M_b is the rigid body mass. Although $\mathbf{X}_b(\mathbf{a}, t)$ is not explicitly given in this case, we can obtain it at each time level by integrating (19) in time.

There are two underlying issues with the moving boundary problem. First, the no-penetration condition (18) is prescribed in the Lagrangian specification of normal velocity. Thus in the ILW procedure we should use material derivatives of primitive variables, i.e., ρ, u, v, p, instead of Eulerian time derivatives of conservative variable $\mathbf{U}$. Secondly, in expansion flows there may be grid points which are outside $\Omega(t_{n-1})$ at the previous time level t_{n-1} but enter $\Omega(t_n)$ at the current time level t_n. Such grid points are called newly emerging points in [34]. The newly emerging points do not have any value at time level t_n. To update their values to time level t_{n+1}, we not only need to construct their values at time level t_n, but also need one extra ghost point in each direction if we assume $\Gamma(t)$ travels a distance of at most h in each direction from t_n to t_{n+1}, i.e.,

$$\Delta t < \frac{h}{\max_{t\in[t_n,t_{n+1}]} \|\mathbf{V}_b(t)\|_\infty}.$$

See Fig. 3.1 in [34] for a demonstration of newly emerging points and ghost points. If the same method is used to construct values of ghost points and newly emerging points, there is actually no need to distinguish them in implementation. We only need to construct values of four ghost points (instead of three ghost points for the static boundary problem) in each direction at time level t_n.

We start to describe a fifth order boundary treatment for the no-penetration boundary condition (18) in which $\mathbf{X}_b(\mathbf{a}, t)$ and thus $\mathbf{V}_b(t)$ are prescribed. A third order boundary treatment is developed in [34] where the ILW procedure is used to convert up to second order derivatives. Here we use the ILW procedure only for the first derivatives, which significantly alleviates the complicated algebra and makes the implementation of the fifth order boundary treatment practical.

We assume the values of all the grid points inside $\Omega(t_{n-1})$ have been updated by the interior scheme from time level t_{n-1} to time level t_n. At time level t_n, we set up the local coordinate system (10) for each ghost point $P = (x_i, y_j)$. In this local coordinate system, the Euler equations (8) are written in terms of primitive variable as

$$\frac{\partial \hat{\mathbf{W}}}{\partial t} + \mathbf{A}(\hat{\mathbf{W}})\frac{\partial \hat{\mathbf{W}}}{\partial \hat{x}} + \mathbf{B}(\hat{\mathbf{W}})\frac{\partial \hat{\mathbf{W}}}{\partial \hat{y}} = 0, \tag{20}$$

where

$$\hat{\mathbf{W}} = \begin{pmatrix} \hat{W}_1 \\ \hat{W}_2 \\ \hat{W}_3 \\ \hat{W}_4 \end{pmatrix} = \begin{pmatrix} \rho \\ \hat{u} \\ \hat{v} \\ p \end{pmatrix}, \qquad \mathbf{A}(\hat{\mathbf{W}}) = \begin{pmatrix} \hat{u} & \rho & 0 & 0 \\ 0 & \hat{u} & 0 & \frac{1}{\rho} \\ 0 & 0 & \hat{u} & 0 \\ 0 & \rho c^2 & 0 & \hat{u} \end{pmatrix},$$

$$\mathbf{B}(\hat{\mathbf{W}}) = \begin{pmatrix} \hat{v} & 0 & \rho & 0 \\ 0 & \hat{v} & 0 & 0 \\ 0 & 0 & \hat{v} & \frac{1}{\rho} \\ 0 & 0 & \rho c^2 & \hat{v} \end{pmatrix}.$$

Our ILW procedure is carried out by the use of (20). The procedure is similar to the static boundary problem and thus we mainly focus on the difference. The notations introduced in Sect. 4.1 are carried over with the exceptions that the conservative variable $\hat{\mathbf{U}}$ is replaced by the primitive variable $\hat{\mathbf{W}}$ and that the Jacobian matrix of the normal flux $\mathbf{A}_\perp(\hat{\mathbf{U}}_0)$ is replaced by $\mathbf{A}(\hat{\mathbf{W}})$.

For a fifth order boundary treatment, the value of the ghost point (x_i, y_j) is imposed by the Taylor expansion

$$(\hat{W}_m)_{i,j} = \sum_{k=0}^{4} \frac{\Delta^k}{k!} \hat{W}_m^{*(k)}, \quad m = 1, \ldots, 4.$$

The constant term $\hat{W}_m^{*(0)}$, $m = 1, \ldots, 4$, is solved by a linear system of equations

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ l_{2,1} & l_{2,2} & l_{2,3} & l_{2,4} \\ l_{3,1} & l_{3,2} & l_{3,3} & l_{3,4} \\ l_{4,1} & l_{4,2} & l_{4,3} & l_{4,4} \end{pmatrix} \begin{pmatrix} \hat{W}_1^{*(0)} \\ \hat{W}_2^{*(0)} \\ \hat{W}_3^{*(0)} \\ \hat{W}_4^{*(0)} \end{pmatrix} = \begin{pmatrix} \mathbf{V}_b(t_n) \cdot \mathbf{n}(\mathbf{x}_0, t_n) \\ V_2^{*(0)} \\ V_3^{*(0)} \\ V_4^{*(0)} \end{pmatrix}. \tag{21}$$

Next, we take the first material derivative $\frac{D}{Dt} = \frac{\partial}{\partial t} + \hat{u}\frac{\partial}{\partial \hat{x}} + \hat{v}\frac{\partial}{\partial \hat{y}}$ of (18) and obtain

$$\frac{D\hat{u}}{Dt} + \hat{\mathbf{u}} \cdot \frac{D\mathbf{n}}{Dt} = \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{V}_b \cdot \mathbf{n}),$$

where $\hat{\mathbf{u}} = (\hat{u}, \hat{v})^T$. Converting the material derivative $\frac{D\hat{u}}{Dt}$ to spatial derivatives by the second equation of (20), we have

$$\frac{\partial p}{\partial \hat{x}} = \rho \left[\hat{\mathbf{u}} \cdot \frac{D\mathbf{n}}{Dt} - \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{V}_b \cdot \mathbf{n}) \right].$$

The right-hand side of the above equation is already known if evaluated at P_0. As a result, $\hat{W}_m^{*(1)}$, $m = 1, \ldots, 4$, can be solved by

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ l_{2,1} & l_{2,2} & l_{2,3} & l_{2,4} \\ l_{3,1} & l_{3,2} & l_{3,3} & l_{3,4} \\ l_{4,1} & l_{4,2} & l_{4,3} & l_{4,4} \end{pmatrix} \begin{pmatrix} \hat{W}_1^{*(1)} \\ \hat{W}_2^{*(1)} \\ \hat{W}_3^{*(1)} \\ \hat{W}_4^{*(1)} \end{pmatrix} = \mathbf{b}, \tag{22}$$

where

$$\mathbf{b} = \begin{pmatrix} \hat{W}_1^{*(0)}[(\hat{W}_2^{*(0)}, \hat{W}_3^{*(0)})^T \cdot \frac{D\mathbf{n}}{Dt} - \frac{\mathrm{d}}{\mathrm{d}t}(\mathbf{V}_b \cdot \mathbf{n})]|_{(x,y)=\mathbf{x}_0, t=t_n} \\ V_2^{*(1)} \\ V_3^{*(1)} \\ V_4^{*(1)} \end{pmatrix}.$$

Higher order spatial derivatives $\hat{W}_m^{*(k)}$, $k = 2, 3, 4$, can be obtained by extrapolation

$$\mathbf{L}(\hat{\mathbf{W}}_0)\begin{pmatrix}\hat{W}_1^{*(k)}\\ \hat{W}_2^{*(k)}\\ \hat{W}_3^{*(k)}\\ \hat{W}_4^{*(k)}\end{pmatrix} = \begin{pmatrix}V_1^{*(k)}\\ V_2^{*(k)}\\ V_3^{*(k)}\\ V_4^{*(k)}\end{pmatrix}. \tag{23}$$

Notice that the procedure we have described so far is for time level t_n only. As mentioned in Sect. 2, we need to match the time levels when constructing values of ghost points in the two intermediate stages of the RK method (3) by (4). Namely, for the first and second intermediate stages, we replace $\mathbf{V}_b(t_n)\cdot\mathbf{n}(\mathbf{x}_0, t_n)$ in (21) by

$$\mathbf{V}_b(t_n)\cdot\mathbf{n}(\mathbf{x}_0, t_n) + \Delta t\left.\frac{\partial \hat{u}}{\partial t}\right|_{(x,y)=\mathbf{x}_0, t=t_n}$$

and

$$\mathbf{V}_b(t_n)\cdot\mathbf{n}(\mathbf{x}_0, t_n) + \frac{\Delta t}{2}\left.\frac{\partial \hat{u}}{\partial t}\right|_{(x,y)=\mathbf{x}_0, t=t_n} + \frac{\Delta t^2}{4}\left.\frac{\partial^2 \hat{u}}{\partial t^2}\right|_{(x,y)=\mathbf{x}_0, t=t_n}$$

respectively. The Eulerian time derivatives can be obtained by a standard Lax-Wendroff procedure, since all the necessary spatial derivatives have been obtained at time level t_n.

The linear system (22) should also be adjusted. We convert Eulerian time derivatives instead of material derivatives. The first equation of (22) is then replaced by

$$-\hat{W}_2^{*(0)}\hat{W}_2^{*(1)} - \frac{\hat{W}_4^{*(1)}}{\hat{W}_1^{*(0)}} = \left.\frac{\partial \hat{u}}{\partial t}\right|_{(x,y)=\mathbf{x}_0, t=t_n} + \Delta t\left.\frac{\partial^2 \hat{u}}{\partial t^2}\right|_{(x,y)=\mathbf{x}_0, t=t_n} + \hat{W}_3^{*(0)}\frac{\partial \hat{u}}{\partial \hat{y}}$$

in the first stage; replaced by

$$-\hat{W}_2^{*(0)}\hat{W}_2^{*(1)} - \frac{\hat{W}_4^{*(1)}}{\hat{W}_1^{*(0)}} = \left.\frac{\partial \hat{u}}{\partial t}\right|_{(x,y)=\mathbf{x}_0, t=t_n} + \frac{\Delta t}{2}\left.\frac{\partial^2 \hat{u}}{\partial t^2}\right|_{(x,y)=\mathbf{x}_0, t=t_n} + \hat{W}_3^{*(0)}\frac{\partial \hat{u}}{\partial \hat{y}}$$

in the second stage. Equation (23) is shared by all the RK stages.

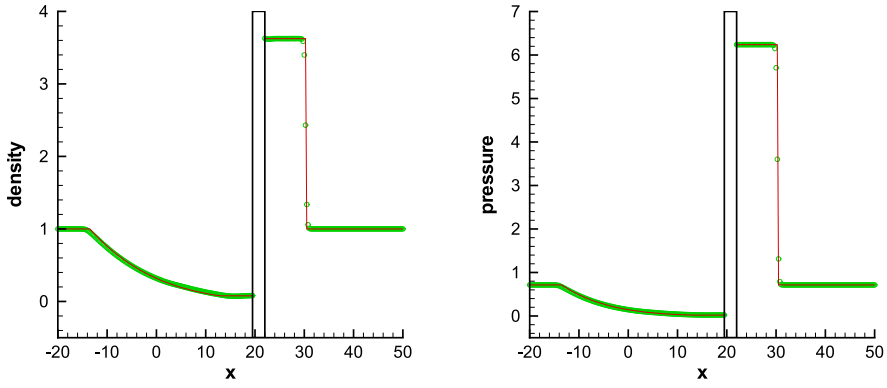
We remark that the boundary treatment for moving geometries provides an alternative to the no-penetration boundary condition in static geometries discussed in Sect. 4.2. Namely, besides conservative variables, primitive variables can also be used.

Example 3 We redo Example 1 in [34] with our fifth order boundary treatment. We consider a 1D gas confined between two rigid walls. The right wall is fixed at $x_r = 1.0$ while the left wall is positioned at $x_l(t)$ and moving. The initial conditions are

$$\begin{aligned}\rho(x, 0) &= 1 + 0.2\cos\big[2\pi(x - 0.5)\big],\\ u(x, 0) &= x - 1,\\ p(x, 0) &= \rho(x, 0)^\gamma,\end{aligned}$$

Table 2 Entropy errors and convergence rates in Example 3. $\Delta t = O(h^{5/3})$ and $t = 0.5$

h	$x_l(t) = 0.5(1-t)$				$x_l(t) = 0.5(1-\sin t)$			
	L^1 error	Order	L^∞ error	Order	L^1 error	Order	L^∞ error	Order
1/40	7.47E–07		1.24E–06		7.31E–07		1.35E–06	
1/80	1.17E–08	5.99	2.88E–08	5.43	1.16E–08	5.98	2.82E–08	5.59
1/160	3.49E–10	5.07	6.36E–10	5.50	3.45E–10	5.07	6.19E–10	5.51
1/320	9.55E–12	5.19	1.72E–11	5.21	9.88E–12	5.12	3.67E–11	4.08



(a) Density (b) Pressure

Fig. 6 Density and pressure profiles in Example 4. The piston is represented by the rectangle. *Solid lines*: exact solutions; *Symbols*: numerical solutions with $h = 0.25$

such that the initial entropy $s(x, 0) = 1$. Here $\gamma = 1.4$ is the ratio of specific heats in the equation of state (9). As long as the solution stays smooth, we have isentropic flow, i.e., $s(x, t) = 1$. Thus the numerical value of the entropy can be used for the analysis of convergence. Our fifth order boundary treatment is used at the left moving boundary while the standard reflection technique is used at the right fixed boundary. We measure the L^1 errors and L^∞ errors in entropy at $t = 0.5$. We can see from Table 2 that our method achieves the designed fifth order accuracy in the L^1 norm.

Example 4 This is a 1D problem involving shocks and rarefaction waves. A piston with width $10h$ is initially centered at $x = -5h$ inside a shock tube. The piston instantaneously moves with a constant velocity $u_p = 2$ into an initially quiescent fluid with $\rho = 1$ and $p = 5/7$. This problem is equivalent to two independent Riemann problems and thus the exact solution can be obtained. A shock forms ahead of the piston and a rarefaction wave forms in the rear. We take $h = 0.25$ and set the CFL number to be 0.5. The density and pressure profiles at $t = 11$ are shown in Fig. 6, together with the exact solution. Our fifth order boundary treatment gives

Table 3 Entropy errors and convergence rates in Example 5. $\Delta t = O(h^{5/3})$

h	$\mathbf{X}_c = (-0.5\sin t, 0), t = 0.7$				$\mathbf{X}_c = (-0.5\sin t, 0.3t), t = 0.5$			
	L^1 error	Order	L^∞ error	Order	L^1 error	Order	L^∞ error	Order
1/5	4.11E–03		2.47E–03		3.93E–03		1.79E–03	
1/10	3.86E–04	3.41	3.60E–04	2.78	4.05E–04	3.28	1.93E–04	3.21
1/20	1.21E–05	5.00	1.12E–05	5.00	1.20E–05	5.07	9.20E–06	4.39
1/40	2.43E–07	5.64	6.08E–07	4.21	2.21E–07	5.77	3.73E–07	4.62

Table 4 Center of the cylinder in Example 6

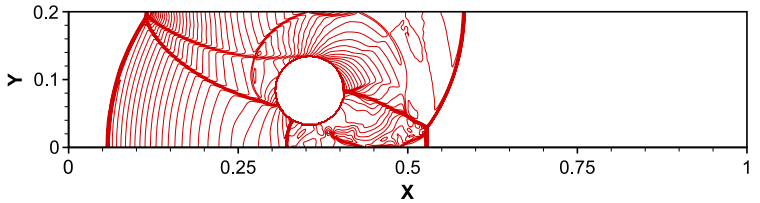
h	$t = 0.1641$		$t = 0.30085$	
	x-coordinate	y-coordinate	x-coordinate	y-coordinate
1/320	3.5878E–01	8.4001E–02	6.4601E–01	1.4713E–01
1/640	3.5542E–01	8.3778E–02	6.3708E–01	1.4717E–01
1/1280	3.5434E–01	8.3827E–02	6.3320E–01	1.4640E–01
1/2560	3.5422E–01	8.4140E–02	6.3357E–01	1.4624E–01

a good non-oscillatory resolution for both shock wave and rarefaction wave on this relatively coarse mesh. The results are similar to those obtained by a third order boundary treatment in Example 2 of [34].

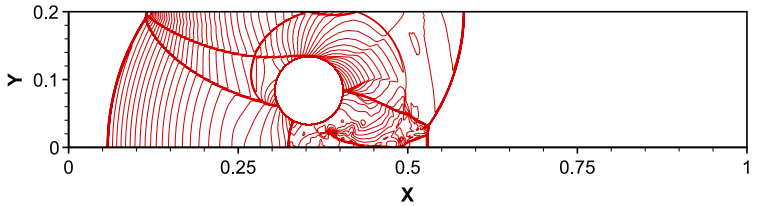
Example 5 We redo Example 4 in [34] with our fifth order boundary treatment. We construct a 2D isentropic flow such that we are able to measure the entropy errors. The computational domain is $[-4, 4] \times [-4, 4]$ with all the boundaries as rigid walls. A rigid cylinder with radius $R = 1$ is initially centered at $(0, 0)$ and starts moving. The center of the cylinder is positioned at $\mathbf{X}_c(t)$. We use our fifth order boundary treatment at the surface of the moving cylinder and the reflection technique at the fixed walls. The initial conditions are the same as in [34]. The entropy errors in the region $[-2, 2] \times [-2, 2]$ are listed in Table 3 for a cylinder moving horizontally and a cylinder moving in the 2D space. We can see that the designed fifth order convergence is achieved in the L^1 norm.

Example 6 We test the so-called cylinder lift-off problem which is first proposed by Falcovitz et al. [8] and considered in [2, 9, 16, 31, 34] later. In this problem, a rigid cylinder initially resting on the floor of a 2D channel is driven and lifted by a strong shock. The problem setup is the same as in [2, 9, 31, 34].

We use our fifth order boundary treatment at the surface of the moving cylinder and the reflection technique at the top and bottom walls. Since the cylinder initially rests exactly on the floor, a stencil for high order extrapolation may be too wide to be contained in the computational domain. We have to use low order extrapolation in this situation and turn to high order extrapolation otherwise. We list the center of the cylinder at two fixed times for different meshes in Table 4. Compared with

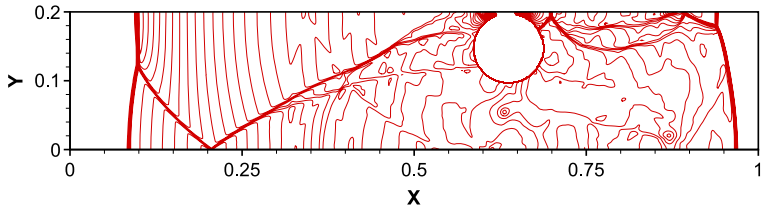


(a) h = 1/ 640

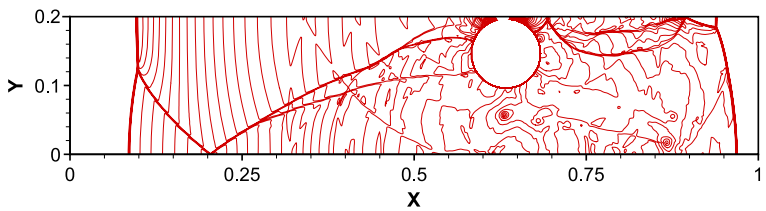


(b) h = 1/ 1280

Fig. 7 Pressure contours in Example 6, 53 contours from 2 to 28. CFL = 0.6 and $t = 0.1641$



(a) h = 1/ 640



(b) h = 1/ 1280

Fig. 8 Pressure contours in Example 6, 53 contours from 2 to 28. CFL = 0.6 and $t = 0.30085$

the results computed by a third order boundary treatment in Example 5 of [34], the center of the cylinder tends to converge to the same limit. We plot pressure contours at $t = 0.1641$ and $t = 0.30085$ in Figs. 7 and 8 respectively. The flow structures agree with those in [2, 16, 31, 34].

6 Concluding Remarks

We give a review and discuss new developments of the ILW procedure for numerical boundary conditions of hyperbolic equations. It is based on Cartesian grids, which is very challenging for accurate and stable boundary treatment because of the wide stencil and the fact that the physical boundary is not necessarily aligned with the grids. Our method is high order accurate, stable under the standard CFL conditions determined by the interior schemes, and easy to implement. It is successfully applied to no-penetration boundary condition of compressible inviscid flows involving complex static or moving geometries. In particular, it has good performance for problems involving complicated interactions between shock waves and rigid boundaries.

The challenge of boundary treatment is not limited to finite difference schemes. Even for finite element type methods, difficulties sometimes arise if unstructured, straight-sided element meshes are used to fit curved geometries, see [3]. Krivodonova and Berger [22] propose an accurate implementation of no-penetration boundary condition for discontinuous Galerkin (DG) methods on such meshes. In future work, we will try to extend our ILW procedure to DG methods on rectangular meshes. Moreover, in some applications, geometrically complicated structures are deformable and the fluid is considered to be viscous. The effectiveness and robustness of our method for these types of problems are also subject to future research.

Acknowledgements Research is supported by AFOSR grant FA9550-09-1-0126 and NSF grant DMS-1112700.

References

1. Appelö, D., Petersson, N.A.: A fourth-order accurate embedded boundary method for the wave equation. Preprint. www.its.caltech.edu/~appelo/preprints/PadeEBPaper.pdf
2. Arienti, M., Hung, P., Morano, E., Shepherd, J.E.: A level set approach to Eulerian-Lagrangian coupling. J. Comput. Phys. **185**, 213–251 (2003)
3. Bassi, F., Rebay, S.: High-order accurate discontinuous finite element solution of the 2D Euler equations. J. Comput. Phys. **138**, 251–285 (1997)
4. Berger, M.J., Helzel, C., LeVeque, R.J.: h-box methods for the approximation of hyperbolic conservation laws on irregular grids. SIAM J. Numer. Anal. **41**, 893–918 (2003)
5. Carpenter, M.H., Gottlieb, D., Abarbanel, S., Don, W.-S.: The theoretical accuracy of Runge-Kutta time discretizations for the initial boundary value problem: a study of the boundary error. SIAM J. Sci. Comput. **16**, 1241–1252 (1995)
6. De Palma, P., de Tullio, M.D., Pascazio, G., Napolitano, M.: An immersed-boundary method for compressible viscous flows. Comput. Fluids **35**, 693–702 (2006)
7. de Tullio, M.D., De Palma, P., Iaccarino, G., Pascazio, G., Napolitano, M.: An immersed boundary method for compressible flows using local grid refinement. J. Comput. Phys. **225**, 2098–2117 (2007)
8. Falcovitz, J., Alfandary, G., Hanoch, G.: A two-dimensional conservation laws scheme for compressible flows with moving boundaries. J. Comput. Phys. **138**, 83–102 (1997)
9. Forrer, H., Berger, M.: Flow simulations on Cartesian grids involving complex moving geometries. In: Jeltsch, R. (ed.) Proc. 7th Intl. Conf. on Hyperbolic Problems, pp. 315–324. Birkhäuser, Basel (1998)

10. Forrer, H., Jeltsh, R.: A high-order boundary treatment for Cartesian-grid methods. J. Comput. Phys. **140**, 259–277 (1998)
11. Ghias, R., Mittal, R., Dong, H.: A sharp interface immersed boundary method for compressible viscous flows. J. Comput. Phys. **225**, 528–553 (2007)
12. Goldberg, M., Tadmor, E.: Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. I. Math. Comput. **32**, 1097–1107 (1978)
13. Goldberg, M., Tadmor, E.: Scheme-independent stability criteria for difference approximations of hyperbolic initial-boundary value problems. II. Math. Comput. **36**, 603–626 (1981)
14. Harten, A., Engquist, B., Osher, S., Chakravarthy, S.R.: Uniformly high order accurate essentially non-oscillatory schemes. III. J. Comput. Phys. **71**, 231–303 (1987)
15. Helzel, C., Berger, M.J., LeVeque, R.J.: A high-resolution rotated grid method for conservation laws with embedded geometries. SIAM J. Sci. Comput. **26**, 785–809 (2005)
16. Hu, X.Y., Khoo, B.C., Adams, N.A., Huang, F.L.: A conservative interface method for compressible flows. J. Comput. Phys. **219**, 553–578 (2006)
17. Huang, L., Shu, C.-W., Zhang, M.: Numerical boundary conditions for the fast sweeping high order WENO methods for solving the Eikonal equation. J. Comput. Math. **26**, 336–346 (2008)
18. Jiang, G.-S., Shu, C.-W.: Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**, 202–228 (1996)
19. Kreiss, H.-O., Petersson, N.A.: A second order accurate embedded boundary method for the wave equation with Dirichlet data. SIAM J. Sci. Comput. **27**, 1141–1167 (2006)
20. Kreiss, H.-O., Petersson, N.A., Yström, J.: Difference approximations for the second order wave equation. SIAM J. Numer. Anal. **40**, 1940–1967 (2002)
21. Kreiss, H.-O., Petersson, N.A., Yström, J.: Difference approximations of the Neumann problem for the second order wave equation. SIAM J. Numer. Anal. **42**, 1292–1323 (2004)
22. Krivodonova, L., Berger, M.: High-order accurate implementation of solid wall boundary conditions in curved geometries. J. Comput. Phys. **211**, 492–512 (2006)
23. Lax, P.D., Wendroff, B.: Systems of conservation laws. Commun. Pure Appl. Math. **13**, 217–237 (1960)
24. LeVeque, R.J., Calhoun, D.: Cartesian grid methods for fluid flow in complex geometries. In: Fauci, L.J., Gueron, S. (eds.) Computational Modeling in Biological Fluid Dynamics, IMA Vol. Math. Appl., vol. 124, pp. 117–143. Springer, New York (2001)
25. LeVeque, R.J., Li, Z.: The immersed interface method for elliptic equations with discontinuous coefficients and singular sources. SIAM J. Numer. Anal. **31**, 1019–1044 (1994)
26. LeVeque, R.J., Li, Z.: Immersed interface methods for Stokes flow with elastic boundaries or surface tensions. SIAM J. Sci. Comput. **18**, 709–735 (1997)
27. Lombard, B., Piraux, J., Gélis, C., Virieux, J.: Free and smooth boundaries in 2-D finite-difference schemes for transient elastic waves. Geophys. J. Int. **172**, 252–261 (2008)
28. Mittal, R., Iaccarino, G.: Immersed boundary methods. Annu. Rev. Fluid Mech. **37**, 239–261 (2005)
29. Peskin, C.S.: Flow patterns around the heart valves. J. Comput. Phys. **10**, 252–271 (1972)
30. Shu, C.-W., Osher, S.: Efficient implementation of essentially non-oscillatory shock-capturing schemes. J. Comput. Phys. **77**, 439–471 (1988)
31. Shyue, K.-M.: A moving-boundary tracking algorithm for inviscid compressible flow. In: Benzoni-Gavage, S., Serre, D. (eds.) Hyperbolic Problems: Theory, Numerics, Applications, pp. 989–996. Springer, Berlin (2008)
32. Sjögreen, B., Petersson, N.A.: A Cartesian embedded boundary method for hyperbolic conservation laws. Commun. Comput. Phys. **2**, 1199–1219 (2007)
33. Tan, S., Shu, C.-W.: Inverse Lax-Wendroff procedure for numerical boundary conditions of conservation laws. J. Comput. Phys. **229**, 8144–8166 (2010)
34. Tan, S., Shu, C.-W.: A high order moving boundary treatment for compressible inviscid flows. J. Comput. Phys. **230**, 6023–6036 (2011)
35. Tan, S., Wang, C., Shu, C.-W., Ning, J.: Efficient implementation of high order inverse Lax-Wendroff boundary treatment for conservation laws. J. Comput. Phys. **231**, 2510–2527 (2012)

36. Xiong, T., Zhang, M., Zhang, Y.-T., Shu, C.-W.: Fifth order fast sweeping WENO scheme for static Hamilton-Jacobi equations with accurate boundary treatment. J. Sci. Comput. **45**, 514–536 (2010)
37. Zhang, Y.-T., Chen, S., Li, F., Zhao, H., Shu, C.-W.: Uniformly accurate discontinuous Galerkin fast sweeping methods for Eikonal equations. SIAM J. Sci. Comput. **33**, 1873–1896 (2011)

Elliptic Curves over Finite Fields: Number Theoretic and Cryptographic Aspects

Igor E. Shparlinski

Abstract We present a collection of several natural questions about elliptic curves, mostly over finite fields, that have led to some interesting number theoretic questions and whose solutions require rather involved techniques from various area of number theory. Some of these questions are of intrinsic value for the theory of elliptic curves; they stem from their application to cryptography.

1 Introduction

1.1 Motivation

The purpose of this work is to exhibit strong links and exciting connections which exist between many problems on elliptic curves over finite fields and various problems of analytic and algebraic number theory. We do not go into the details, but rather try to demonstrate the rich variety of these links.

The main reason of the great interest to elliptic curves is that they have naturally associated with them a group structure. Studying these groups, theoretically and experimentally, has been one of most active directions of research for more than a century. This has led to remarkable achievements and intriguing conjectures, see [142].

Since the invention of elliptic curve cryptography by Koblitz [93] and Miller [120] many new directions of research have appeared in this area. However, whether it is a classical theory of elliptic curves or an application driven investigation, which we discuss in the survey, number theory has always played a prominent role.

The main purpose of this short and mostly self-contained survey is to give some taste of this beautiful area, show what kind of problems have been considered and also demonstrate some capabilities of modern number theory.

With only very few exceptions, most of the necessary facts and notions the reader needs to know about elliptic curves are given here, so this survey should be accessible to essentially any mathematician, computer scientist or cryptographer.

I.E. Shparlinski (✉)
Department of Computing, Macquarie University, Sydney, NSW 2109, Australia
e-mail: igor.shparlinski@mq.edu.au

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5_4,

Some of the corresponding number theoretic problems are of intrinsic interest and well-known (and also known to be notoriously hard). Some are entirely motivated by the applications to elliptic curves. Some of the problems have partial solutions, some have conditional solutions under various assumptions such as the *Generalised Riemann Hypothesis* and the *Dickson Prime s-tuplets Conjecture*; some are still wide open and even heuristically the situation is not well understood.

There are also very deep and intriguing connections between various properties of elliptic curves over the rationals (and other fields of zero characteristic) and number theory. For example, there are strong links between

- the number and distribution of integral and S-integral points on elliptic curves on one side and Diophantine problems and geometry of numbers on the other, see [60, 81];
- average ranks and a related problem of the distribution of zeros of L-function for some natural families of elliptic curves and bounds of exponential and multiplicative sums, see [11, 25, 80, 151].
- bounds on the torsion of elliptic curves in algebraic number fields and the distribution of solutions to the congruence $xy \equiv a \pmod{p}$, see [119].

However, here we are mostly interested in the case of elliptic curves over finite fields.

We hope that this collection of problems will stimulate both number theorists to work on new exciting problems, and experts in elliptic curves, in elliptic curve cryptography in particular, posing such problems to the number theoretic community.

1.2 Arranging a Group Structure on Elliptic Curves

Here we give a rather informal explanation how elliptic curves get equipped with a group structure.

Although the argument below works in any field, it becomes more visually obvious for the case of real numbers $\mathbb{F} = \mathbb{R}$. Generally speaking, an elliptic curve $\mathbf{E}$ over a field $\mathbb{F}$ is given by a cubic equation, which we assume to be of the form

$$Y^2 = X^3 + aX + b \tag{1}$$

(this can always be assumed if the characteristic of $\mathbb{F}$ is different from 2 and 3, see Sect. 1.4).

We take two points P_1 and P_2 on $\mathbf{E}$ and try to design a natural composition law that leads to a group. To do so, we take a straight line that passes through $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$. Since the curve is defined by a cubic equation, this line has one more intersection point, say Q. This may naively suggest one to call $Q = (x_0, y_0)$ the result of "addition" of P_1 and P_2, and we write $Q = P_1 \oplus P_2$. However, this definition also leads to such strange relations as $P_1 \oplus Q = P_2$ and $P_2 \oplus Q = P_1$ and thus has to be modified. To do so we note that the point $P = (x_0, -y_0)$

also belongs to the curve $\mathbf{E}$ and this is exactly how we define the composition law, by setting $P = P_1 \oplus P_2$. Now one can easily check that all properties of a group operation are satisfied. This is, however, not the end of the story, as we have to consider several special cases. For example, what if $P_1 = P_2$? In this case a natural analogue of a line passing through P_1 and P_2 is a tangent line to the point $P_1 = P_2$. Furthermore, what if $P_1 = (x, y)$ and $P_2 = (x, -y)$? The vertical line that passes through P_1 and P_2 has no intersection point with $\mathbf{E}$ (to see this we recall that the graph of $\mathbf{E}$ is symmetric with respect to the y-axis, so if it intersects $\mathbf{E}$ it intersects it in at least two points Q_1, Q_2, giving in total four intersection points P_1, P_2, Q_1, Q_2, which is impossible for a line and cubic curve. To remedy the situation we add one more point, which we call the *point at infinity* which is not the graph of $\mathbf{E}$. This point can easily be demystified in a *projective model* of $\mathbf{E}$ given by the homogeneous equation

$$Y^2Z = X^3 + aXZ^2 + bZ^3$$

where this is the point $(0, 1, 0)$.

It also remains to observe that in computing $P = P_1 \oplus P_2$ one does not have to appeal to the geometric interpretation. There are rather simple explicit algebraic formulas that give the coordinates of P as rational functions of coordinates of P_1 and P_2.

1.3 General Notation

The letter p (possibly subscripted) always denotes a prime; k, m and n always denote integers (and so do K, M and N).

As usual, we say that $n \geq 2$ is y-smooth if all prime divisors p of n satisfy $p \leq y$.

We use ε to denote a small positive parameter on which implied constants may depend.

Calligraphic letters, for example, $\mathcal{A} = (a_n)$, usually denote sets or sequences of integers. For a prime power q, we use $\mathbb{F}_q$ to denote the finite field of q elements.

For an integer m, we use $\mathbb{Z}_m$ to denote the residue ring modulo m.

We use some standard notations for most common arithmetic functions. For $m \geq 2$ be an integer, we denote by:

- $P(m)$ the largest prime divisor of m,
- $\varphi(m)$ the Euler (totient) function of m,
- $\omega(m)$ the number of distinct prime divisors of m,
- $\tau(m)$ the number of positive integer divisors of m.

We also define $P(1) = \omega(1) = 0$ and $\tau(1) = \varphi(1) = 1$.

Letting $x \geq 0$ be a real number, we denote by:

- $\pi(x)$ the number of primes $p \leq x$,
- $\pi(x; q, a)$ the number of primes $p \leq x$ such that $p \equiv a \pmod{q}$.

We abbreviate the Generalised Riemann Hypothesis as GRH.

We use the Vinogradov notation '$f(x) \ll g(x)$' which is equivalent to the Landau notation $f(x) = O(g(x))$. If convenient, we also write $g(x) \gg f(x)$ instead of $f(x) \ll g(x)$. The notation $f(x) \asymp g(x)$ is equivalent to $f(x) \ll g(x) \ll f(x)$.

Finally, $\log x$ denotes the natural logarithm; we always assume that the argument is large enough for the whole expression to make sense (as well as in the case of iterated logarithms).

1.4 Basic Facts on Elliptic Curves

Let $\mathbf{E}$ be an elliptic curve defined over $\mathbb{F}_q$, given by an *affine Weierstraß equation* of the form

$$Y^2 + (a_1X + a_3)Y = X^3 + a_2X^2 + a_4X + a_6,$$

with coefficients $a_1, a_2, a_3, a_4, a_6 \in \mathbb{F}_q$, such that the partial derivatives $a_1Y - 3X^2 - a_2X - a_4$ and $2Y + a_1X + a_3$ do not vanish simultaneously at points of the curve $(x, y) \in \mathbf{E}(\overline{\mathbb{F}}_q)$ over the algebraic closure $\overline{\mathbb{F}}_q$ of $\mathbb{F}_q$, see [22, 142]. We put $h(X) = a_1X + a_3$ and $f(X) = X^3 + a_2X^2 + a_4X + a_6$, thus the Weierstraß equation becomes

$$Y^2 + h(X)Y - f(X) = 0.$$

If p is the characteristic of $\mathbb{F}_q$ then for $p > 2$ one can always take $h = 0$ and for $p > 3$ also $a_2 = 0$; for $p = 2$ at least one of a_1, a_3 must be nonzero. In the case that $p > 3$ every elliptic curve over $\mathbb{F}_q$ can be given by a Weierstraß equation (1) for some $a, b \in \mathbb{F}_q$ with $4a^3 + 27b^2 \neq 0$.

We use $\mathbf{E}_{a,b}$ to denote the elliptic curve defined by (1).

We recall that the set $\mathbf{E}(\mathbb{F}_q)$ of $\mathbb{F}_q$-rational points forms an abelian group of order which satisfies the Hasse–Weil inequality,

$$\left|\#\mathbf{E}(\mathbb{F}_q) - q - 1\right| \le 2q^{1/2}, \tag{2}$$

and with the *point at infinity* $\mathcal{O}$ as the neutral element of this group (which does not have affine coordinates), see [5, 22, 100, 142] for this and other general properties of elliptic curves. In particular, it is common to call $[q + 1 - 2q^{1/2}, q + 1 + 2q^{1/2}]$ as the *Hasse–Weil interval*.

We use $\oplus$ to denote the group operation. For example, $Q \oplus \mathcal{O} = Q$ for any point $Q \in \mathbf{E}(\mathbb{F}_q)$.

It is well-known that the group of $\mathbb{F}_q$-rational points $\mathbf{E}(\mathbb{F}_q)$ is of the form

$$\mathbf{E}(\mathbb{F}_q) \cong \mathbb{Z}/L\mathbb{Z} \times \mathbb{Z}/M\mathbb{Z}, \tag{3}$$

where the integers L and M are uniquely determined by the condition $M \mid L$. In particular, L is the *exponent* of $\mathbf{E}(\mathbb{F}_q)$, which we denote by $\ell_{\mathbf{E}}(p)$. Thus $\ell_{\mathbf{E}}(p)P = \mathcal{O}$ for any point $P \in \mathbf{E}(\mathbb{F}_q)$.

We also recall that the *Weil pairing* implies that

$$M \mid q - 1.$$

For a point $Q \in \mathbf{E}(\mathbb{F}_q)$ we use $x(Q)$ and $y(Q)$ to denote its components, that is, $Q = (x(Q), y(Q))$.

We refer to [142] for precise definition of several important classes of elliptic curves that occur in this paper, such as *ordinary* and *supersingular curves* and also *CM* and *non-CM* curves.

2 Structure of Elliptic Curves over Finite Fields

2.1 Isogeny and Isomorphism Classes in Various Families

We recall that two curves $\mathbf{E}_{r,s}$ and $\mathbf{E}_{u,v}$ (given by a Weierstraß equation of the form (1) where $\gcd(q, 6) = 1$) are *isomorphic* over $\mathbb{F}_q$ if for some $\lambda \in \mathbb{F}_q^*$ we have

$$r\lambda^4 = u \quad \text{and} \quad s\lambda^6 = v.$$

Note that this is different from the isomorphism of the corresponding groups of points. Furthermore, $\mathbf{E}_{r,s}$ and $\mathbf{E}_{u,v}$ are *isogenous* over $\mathbb{F}_q$ if

$$\#\mathbf{E}_{r,s}(\mathbb{F}_q) = \#\mathbf{E}_{u,v}(\mathbb{F}_q).$$

It is well-known that there are $q/2 + O(1)$ distinct isomorphism classes of elliptic curves over $\mathbb{F}_q$ (note that this is different from the isomorphism of the corresponding groups of points), see [83, 101].

Lenstra [101, Sect. 1] uses the link between the number of elliptic curves over $\mathbb{F}_p$ with a given number of $\mathbb{F}_p$-points and the Kronecker class numbers. McKee [117, Theorem 2] has given a more precise form of the upper bounds of Lenstra [101, Proposition 1.9], which implies that the number $W_p(t)$ of pairs $(a, b) \in \mathbb{F}_p^2$ for which the Weierstraß equation (1) defines a curve $\mathbf{E}$ with $\#\mathbf{E}(\mathbb{F}_p) = p + 1 - t$ can be estimated as

$$W_p(t) \ll \frac{d}{\varphi(d)} p^{3/2} \log p \ll p^{3/2} \log p \log\log p,$$

where d is the largest positive integer with $d^2 \mid t^2 - 4p$ (and a little stronger bound under the GRH). It is also shown by Lenstra [101, Proposition 1.9] that

$$W_p(t) \gg p^{3/2} (\log p)^{-1}$$

for all but at most two values of $t \in [p + 1 - 2p^{1/2}, p + 1 + 2p^{1/2}]$.

Obtaining tighter bounds on $W_p(t)$ and its average values over various sets of t (see, for example, [108] for one of such results on average), is of ultimate interest

for the refined analysis of this algorithm. In turn, such an analysis may require new analytic results about the size of L-functions of quadratic fields.

There are several works which count the number of curves $\mathbf{E}_{r,s}$ that are isomorphic over $\mathbb{F}_p$ to a given curve $\mathbf{E}_{r,s}$ with coefficients in r, s in a given box $(r, s) \in [R+1, R+K] \times [S+1, S+L]$, see [18, 72]. In particular, if for

$$KL \ge p^{3/2+\varepsilon} \quad \text{and} \quad \min\{K, L\} \ge p^{1/2+\varepsilon} \tag{4}$$

with fixed $\varepsilon > 0$, using the exponential sum technique, Fouvry and Murty [72] have obtained an asymptotic formula for every pair $(a, b) \in \mathbb{F}_p^2$ with $4a^3 + 27b^2 \ne 0$. In [18], using bounds of multiplicative character sums, for almost all (a, b), this condition (4) has been relaxed as

$$KL \ge p^{1+\varepsilon} \quad \text{and} \quad \min\{K, L\} \ge p^{1/4+\varepsilon}.$$

Furthermore, it is shown in [18] that

$$KL \ge p^{1+\varepsilon} \quad \text{and} \quad \min\{K, L\} \ge p^{1/4e^{1/2}+\varepsilon}$$

one can get a lower bound on the right order of magnitude (again for almost all (a, b)). On average over p, such results are established for even smaller boxes [18] and also for more general families of elliptic curves [137, 138].

In [32] much smaller boxes have been considered. Let $I(R, S; M)$ be the number of nonisomorphic curves $\mathbf{E}_{r,s}$ over $\mathbb{F}_p$ with coefficients r, s in a box $(r, s) \in [R+1, R+M] \times [S+1, S+M]$. Using the method of [30], which in turn has been further developed in [31], in [32] a lower bound of an asymptotically correct order of magnitude is given on $I(R, S; M)$. The method is related to several recent results on the distribution of solutions to polynomial congruences in very small boxes, see [30–32] for further references. Both the bound of [32] on $I(R, S; M)$ and some estimates of [30–32] on the density of solutions of polynomial congruences have been improved and generalised in [28].

For several other aspects of the distribution of the cardinalities of the reduction of a elliptic curve $\mathbf{E}$ over $\mathbb{Q}$ modulo consecutive primes, that lead to interesting number theoretic questions, see [6–8, 12, 39, 40, 43, 50, 89, 90, 113] and references therein.

Finally we notice one more important aspect of this problem. Instead of limiting the range of the coefficients one can consider some other forms of representing elliptic curves, such as Edwards

$$X^2 + Y^2 = a^2\big(1 + X^2Y^2\big), \tag{5}$$

Hessian

$$X^3 + Y^3 + 1 = 3aXY$$

Jacobi

$$Y^2 = X^4 + 2aX^2 + 1$$

Legendre

$$Y^2 = X(X-1)(X-a),$$

and other forms. There is an extensive literature where the authors study the distribution of isomorphism and isogeny classes of elliptic curves in these and several other related families, see [4, 27, 62–64, 83] and references therein.

2.2 Finding the Group Structure

The algorithm of Schoof [129] computes $\#\mathbf{E}(\mathbb{F}_q)$ in deterministic polynomial time, see also [5, 22, 23] for more recent improvements (both theoretic and practical). However, computing the group structure (3) seems to be much more complicated.

The deterministic algorithm of [98] computes the group structure of $\mathbf{E}(\mathbb{F}_q)$ for any elliptic curve $\mathbf{E}$ over $\mathbb{F}_q$ in exponential time $q^{1/2+o(1)}$ which is too slow for practical applications.

A more efficient but probabilistic algorithm of Miller [121] runs in expected polynomial time plus the time needed to factor $\gcd(\#\mathbf{E}(\mathbb{F}_q), q-1)$.

It is shown in [76] that for a sufficiently large p and for almost all elliptic curves $\mathbf{E}$ over $\mathbb{F}_p$, the factorisation part of the algorithm is in fact less time consuming than the rest of the computation (since $\gcd(\#\mathbf{E}(\mathbb{F}_p), p-1)$ tends to be rather small). This result is based on various estimates of sums over divisors of $p-1$.

For elliptic curves over high degree extensions of $\mathbb{F}_q$ a different approach has been suggested in n [141]. Let $\mathbf{E}$ be a curve over $\mathbb{F}_{q^n}$. In [141] a small set of points of $\mathbf{E}(\mathbb{F}_{q^n})$ is constructed that contains a point of order L (that is, of the largest possible order). This is an analogue of the results of [131, 132] on small sets in finite fields containing primitive elements. It is based on a bound of very short exponential sums obtained in [141], namely of sums over the set points P of $\mathbf{E}(\mathbb{F}_{q^n})$ with $x(P) \in \mathbb{F}_q$.

2.3 Groups Represented by Elliptic Curves

Although possible group structure of elliptic curves over $\mathbb{F}_q$, that is, necessary and sufficient conditions on possible values of L and M in (3) have been fully described in [128, 144, 148], the area still contains several open questions. For example, one can ask about the number and frequency of pairs (L, M) that may appear in (3) for all elliptic curves

- over a given finite field $\mathbb{F}_q$, see [65];
- over all possible finite fields, see [17, 51].

Many of the numerical results of [17, 65] indicate some surprising and not fully explained phenomena, thus it still remains to provide some theoretic or even heuristic explanation to them. Furthermore, some theoretic results of [17, 65] are based

on some classical number theoretic tools such as asymptotic formulas for sums of multiplicative functions, distribution of primes in progressions, finiteness of solution sets to some Diophantine equations, and so on, including a recent striking result of Baker [16] (improving those of Baier and Zhao [9, 10]) on the Bombieri–Vinogradov theorem modulo squares.

The case when the group of points is cyclic is very appealing and indeed has been studied quite extensively

Vlăduţ [146] has considered the statistic of elliptic curves $\mathbf{E}$ over $\mathbb{F}_q$ (and also separately for ordinary and all supersingular curves) such that the group $\mathbf{E}(\mathbb{F}_q)$ is cyclic.

For instance, by [146, Theorem 6.1] the proportion of isomorphism classes of elliptic curves over $\mathbb{F}_q$ that contain cyclic curves is

$$c(q) = \prod_{\substack{\ell \mid q-1 \\ \ell \text{ prime}}} \left(1 - \frac{1}{\ell(\ell^2-1)}\right) + O\big(q^{-1/2+o(1)}\big). \tag{6}$$

Furthermore, a complete characterisation of the class $\mathcal{Q}$ of prime powers q for which all curves $\mathbb{F}_q$ are cyclic (that is, $c(q) = 1$) is given in of [146, Theorem 4.1]. However, it is still unknown whether the set $\mathcal{Q}$ is infinite although it is reasonable to expect so because it includes all powers $q = 2^r$ for which $2^r - 1$ is a *Mersenne prime* except for $q = 4$. Furthermore, $\mathcal{Q}$ contains some other families of prime powers but studying them may not be much easier than studying Mersenne primes. In particular, by [146, Theorem 4.1], $c(q) = 1$ is possible only for even q. Taking the sequence $q = 2^r$ with a prime rm and using that all prime divisors $\ell \mid 2^r - 1$ satisfy $\ell \equiv 1 \pmod r$ we easy see from (6) that $c(q) = 1 + o(1)$ for such prime powers.

For curves in extension fields, this question appears to be somewhat easier and has been satisfactorily answered by Vlăduţ [147]. For example given a curve $\mathbf{E}$ over $\mathbb{F}_q$, and a large real $x > 0$ one can define the set

$$\mathcal{C}_{\mathbf{E}}(x) = \big\{n \le x \mid \mathbf{E}(\mathbb{F}_{q^n}) \text{ is cyclic}\big\}.$$

Now we define the upper and lower densities

$$D(\mathbf{E}) = \limsup_{x\to\infty} \#\mathcal{C}_{\mathbf{E}}(x)/x \quad \text{and} \quad d(\mathbf{E}) = \liminf_{x\to\infty} \#\mathcal{C}_{\mathbf{E}}(x)/x.$$

Clearly $D(\mathbf{E}) = d(\mathbf{E}) = 0$ unless $\mathbf{E}(\mathbb{F}_q)$ is cyclic itself. We now put

$$\Delta(q) = \max\big\{D(\mathbf{E}) \mid \mathbf{E} \in \mathcal{E}(q), \mathbf{E}(\mathbb{F}_q) \text{ is cyclic}\big\},$$
$$\delta(q) = \min\big\{d(\mathbf{E}) \mid \mathbf{E} \in \mathcal{E}(q), \mathbf{E}(\mathbb{F}_q) \text{ is cyclic}\big\}.$$

It is shown in [147, Corollary 4.1] that $\Delta(q) \le 7/8$ if q is even, and $\Delta(q) \in \{1/2, 2/3\}$ if q is odd and each case $\Delta(q) = 1/2$ and $\Delta(q) = 2/3$ is fully characterised. Yet another result of Vlăduţ [147, Corollary 4.2] asserts that

$$\liminf_{p\to\infty} \delta(p) = 0$$

when $p \to \infty$ runs through the set of primes. Similar quantities are also studied separately for the families of ordinary and supersingular curves.

In [34, 35, 42] several asymptotic formulas are given (under various standard assumptions such as GRH and in fact even unconditionally in the CM case) for the number of primes $p \le x$ such that a reduction modulo p of a fixed elliptic curve over $\mathbb{Q}$ is cyclic.

Furthermore, Cojocaru and Murty [42] give very good bounds on the smallest prime such that the reduction modulo this prime is cyclic.

2.4 Exponent

For a fixed non-CM elliptic curve $\mathbf{E}$ which is defined over $\mathbb{Q}$, Schoof [130] has shown that

$$\ell_{\mathbf{E}}(p) \gg \frac{p^{1/2} \log p}{\log \log p}$$

for all sufficiently large primes p. It has also been shown in [130], that, under the GRH, for any curve $\mathbf{E}$ over $\mathbb{Q}$,

$$\liminf_{p \to \infty} \frac{\ell_{\mathbf{E}}(p)}{p^{7/8} \log p} < \infty \tag{7}$$

where p runs through all prime numbers. This bound rests on an explicit form of the Chebotarev Density Theorem given by Lagarias, Montgomery and Odlyzko [99] (which assumes the GRH). Accordingly, unconditional results of [99] lead to an unconditional, albeit much weaker, upper bound on $\ell_{\mathbf{E}}(p)$, but this has never been worked out.

Schoof [130] also notices that it is very plausible that for CM curves the exponent can be very small infinitely often. For example, the curve $\mathbf{E}$ is given by $Y^2 = X^3 - X$ (with complex multiplication over $\mathbb{Z}[i]$) we have $\ell_{\mathbf{E}}(p) = k \sim p^{1/2}$ for primes p of the form $p = k^2 + 1$. Certainly, it is not known yet whether there are infinitely many such primes, but it is shown by Matomäki [115] that there are infinitely many primes of the form $p = mk^2 + 1$ with $k = p^{1/4+o(1)}$. Also, using this result, Pappalardi [125] has shown that there are infinitely many prime powers q (in fact of the form $q = p^3$) such that for some elliptic curve $\mathbf{E}$ over $\mathbb{F}_q$ we have

$$\ell_{\mathbf{E}}(q) = \left(\frac{1}{3} + o(1)\right) q^{2/3}.$$

Duke [57] has shown, unconditionally for CM elliptic curves and under the GRH for non-CM elliptic curves, that for any function $f(x) \to \infty$ for $x \to \infty$, the lower bound $\ell_{\mathbf{E}}(p) \ge p/f(p)$ holds for almost all primes p (that is, for a set of primes of relative density one). For non-CM elliptic curves, the only unconditional result available is also in [57], and asserts that the weaker inequality $\ell_{\mathbf{E}}(p) \ge p^{3/4}/\log p$

holds for almost all primes p. Moreover, it follows from the proof of that result that there exists a set of primes of relative density one such that for any p from this set the bound $\ell_{\mathbf{E}}(p) \ge p^{3/4}/\log p$ holds for all elliptic curves modulo p. It is shown in [69] that this bound is tight and any fixed ε there is a positive proportion of primes p such that there exists an elliptic curve $\mathbf{E}$ over $\mathbb{F}_p$ with

$$\ell_{\mathbf{E}}(p) \le p^{3/4+\varepsilon}.$$

Furthermore, in [69], using various results about the distribution of divisors, similar results are also obtained for higher genus curves.

Freiberg and Kurlberg [74], have given an asymptotic formula for the average value of $\ell_{\mathbf{E}}(p)$ over primes $p \le x$ (unconditionally for a CM curve $\mathbf{E}$, and under the GRH for an arbitrary curve $\mathbf{E}$). The result of [74] has recently been improved by Wu [150], see also [92].

It has been shown in [107] that for an ordinary curve $\mathbf{E}$ is defined over $\mathbb{F}_q$, then for its exponent in $\ell_{\mathbf{E}}(q^n)$ in extension fields of $\mathbb{F}_q$ much stronger lower bounds on $\ell_{\mathbf{E}}(q^n)$ can be obtained. For example, by [107, Theorem 3.1] for any $\varepsilon > 0$, the inequality

$$\ell_{\mathbf{E}}\left(q^n\right) \ge q^{n(1-\varepsilon)} \tag{8}$$

holds for all but at most $2^{\vartheta\varepsilon^{-6}+O(\varepsilon^{-5})}$ values of n, where $\vartheta = 3^7 2^{-10} = 2.135\ldots$.

The proof of (8) is based on a recent version of the celebrated *Subspace Theorem* (see [59, 61] for most recent achievements) and an upper bound of van der Poorten and Schlickewei [127], on the number of zeros of linear recurrence sequences. It also uses several ideas from [45], see also [112]. Pappalardi [125] has proved an explicit and fully uniform bound which also applies to supersingular curves: either $m = 2r$ is even and

$$\mathbf{E}(\mathbb{F}_{p^m}) \cong \mathbb{Z}_{p^r\pm1} \times \mathbb{Z}_{p^r\pm1}$$

or

$$\ell_{\mathbf{E}}\left(p^m\right) \ge 2^{-46} p^{m/2} \frac{m^{1/3}}{(\log m)^{8/3}(\log\log m)^{1/3}}.$$

Because the proof of (8) is based on the Subspace Theorem the set of exceptional values cannot be effectively determined. Using very different arguments, Luca and Shparlinski [107] have also derived a much weaker but effective bound

$$\ell_{\mathbf{E}}\left(q^n\right) \ge q^{n/2+\vartheta(q)n/\log n}$$

which holds for all positive integers n, where $\vartheta(q) > 0$ is an effectively computable constant.

We note that the bound (8) means that no result of the same strength as (7) is possible for elliptic curves in extension fields.

Accordingly, Luca, McKee and Shparlinski [105] give a more modest bound which asserts that for some absolute constant $\eta > 0$

$$\liminf_{n\to\infty} \frac{\ell_{\mathbf{E}}(q^n)}{q^n \exp(-n^{\eta/\log\log n})} < \infty.$$

The proof is based on studying the degree of the field extension of $\mathbb{F}_q$ containing all *k-torsion points*, that is, the points $P \in \mathbf{E}(\overline{\mathbb{F}}_q)$ on $\mathbf{E}$ in the algebraic closure $\overline{\mathbb{F}}_q$ of $\mathbb{F}_q$ with $kP = \mathcal{O}$. It is known that for $\gcd(k, q) = 1$ these points form a group

$$\mathbf{E}[k] \cong \mathbb{Z}/k\mathbb{Z} \times \mathbb{Z}/k\mathbb{Z}$$

of cardinality k^2. Let $d(k)$ denote the degree of the field of definition of $\mathbf{E}[k]$ (that is, the field generated by the coordinates of all the k-torsion points, over $\mathbb{F}_q$). It is shown in [105] that if r is a prime with

$$\gcd\bigl(r, q\left(t_n^2 - 4q^n\right)\bigr) = 1$$

where

$$t_n = \#\mathbf{E}(\mathbb{F}_{q^n}) - q^n - 1$$

and such that $t_n^2 - 4q^n$ is a quadratic residue modulo r, then we have $d(r) \mid r - 1$. Then a modification of a result of [2] is shown which asserts that infinitely many integers n have exponentially many divisors of the form $r - 1$, where r is one of the above primes. For each such n one can easily conclude that $\mathbf{E}(\mathbb{F}_{q^n})$ contains the corresponding torsion subgroups $\mathbf{E}[r]$ which forces $\ell_{\mathbf{E}}(q^n)$ to be sufficiently small compared to q^n.

Finally, one can also study an apparently easier question about the distribution of $\ell_{\mathbf{E}}(q)$ "on average" over various families of elliptic curves over $\mathbb{F}_q$, see [133]. For example, it is shown in [133] that "on average" over all elliptic curves $\mathbf{E}$ over $\mathbb{F}_p$, we have

$$\ell_{\mathbf{E}} \geq \frac{p}{\log p(\log\log p)^3},$$

see also [135].

2.5 Prime Cardinalities

Studying how often $\#\mathbf{E}(\mathbb{F}_q)$ is prime is of great interest, although it appears to be very hard to get any rigorously proved results here. In both situations when $\mathbf{E}$ is a fixed curve over $\mathbb{Q}$ and p varies and when p is fixed prime and $\mathbf{E}$ runs through all elliptic curves over $\mathbb{F}_p$, any unconditional results about the primality of $\#\mathbf{E}(\mathbb{F}_p)$ appears to be out of reach. In fact, since we know how the cardinalities are distributed, when the field is fixed, it is essentially a question of the existence of prime numbers in very short intervals. We note that unfortunately even the Riemann Hypothesis is not powerful enough to ensure the existence of a prime in every Hasse–Weil interval $[q + 1 - 2q^{1/2}, q + 1 + 2q^{1/2}]$ corresponding to (2), see [85].

Even if rigorous results seem to be out reach, heuristically the situation is reasonably well understood (which is quite sufficient for cryptographic applications) thanks to work of Koblitz [94, 96] and then thanks to follow up developments due to Galbraith and McKee [78], Weng [149] and Zywina [152] (see also [26] for a generalisation to Jacobians of curves of genus $g = 2$).

Studying the primality of $\#\mathbf{E}(\mathbb{F}_{q^n})/\#\mathbf{E}(\mathbb{F}_q)$ when $\mathbf{E}$ is a fixed curve over $\mathbb{F}_q$ and n varies is probably even harder. Clearly, this question is no easier than the question of primality of *Mersenne numbers*.

The only scenario in which rigorous results have been obtained is when both p and $\mathbf{E}$ both vary. In this case Koblitz [95] shows that $\#\mathbf{E}(\mathbb{F}_p)$ is prime for the set of pairs $(p, \mathbf{E})$ of the right order of magnitude. The result of [95] follows from the prime number theorem and the aforementioned result of Lenstra [101], which asserts that every integer value, in the interval $[p+1-2p^{1/2}, p+1+2p^{1/2}]$, except maybe for at most two such integers, is taken by $\#\mathbf{E}(\mathbb{F}_p)$ about the same number of times when $\mathbf{E}$ runs through the set of all elliptic curves over $\mathbb{F}_p$. The result of [95] also plays a very important role in the argument of [13].

Probably one of the hardest questions in this area is proving that for a given torsion-free curve $\mathbf{E}$ over $\mathbb{Q}$ there are infinitely many primes such that the cardinality of the reduction of $\mathbf{E}$ modulo p, that is, $\#\mathbf{E}(\mathbb{F}_p)$ is prime. This problem is probably harder that the twin prime conjecture. Thus, the best one can hope for is to prove that there are infinitely many primes for which $\#\mathbf{E}(\mathbb{F}_p)$ has few prime divisors and indeed several results of this type have been obtained, see [33, 36, 53, 87] and references therein. All these results are based on deep analytic tools such as sieve methods and effective versions of the Chebotarev density theorem such as of Lagarias, Montgomery and Odlyzko [99]. Some of these results also rely on the GRH and other number theoretic conjectures. However, on average over a and b for the family of curves $\mathbf{E}_{a,b}$ with coefficients in $|a| \le A$, $|b| \le B$ quite strong unconditional results about the frequency of prime cardinalities have been obtained by Balog, Cojocaru and David [15].

Certainly obtaining upper bounds on the number of reductions of $\mathbf{E}$ with prime cardinalities is easier, see [33, 36, 53]. In fact one can even obtain rather strong bounds on the number of reductions of $\mathbf{E}$ with pseudoprime cardinalities, see [41, 52, 111].

Finally, using bounds of some double character sums, Shparlinski and Sutherland [140] have studied the distribution of so-called Atkin and Elkies primes (which are important for point counting algorithms). In turn, it has led to fast and rigorously analysed algorithms for finding elliptic curves over $\mathbb{F}_q$ for which $\#\mathbf{E}(\mathbb{F}_q)$ is prime, see [140] for details.

2.6 Smooth and Easily Factorable Cardinalities

Maurer [116] has presented an polynomial time algorithm, which for any ε, given an integer N, requires at most $\varepsilon \log N$ bits of information and factors N. The idea of the

algorithm is to use an elliptic curve $\mathbf{E}$ over a finite field $\mathbb{F}_p$ such that $\#\mathbf{E}(\mathbb{F}_p)$ satisfies certain rather stringent smoothness conditions. Unfortunately a rigorous analysis of this algorithm requires very precise results about the distribution of smooth numbers in short intervals, which currently seems to be beyond reach. Thus, the main result of [116] is conditional and relies on heuristic assumptions.

Motivated by the work of Maurer [116], Koblitz, Menezes and Shparlinski [97] have considered oracle assisted algorithms for the Diffie–Hellman problem and obtained some nontrivial algorithms under both the same conjecture as in [116] and also under the GRH. These algorithms are based on some results about the plentitude of "easily factorable" integers in short intervals; these are the integers which are products of a smooth number and a large prime. One of the main ideas of [97], using genus two curves instead of elliptic curves, stems from the work of Lenstra, Pila and Pomerance [103, 104] and actually the argument makes direct use of some of the results of [103, 104]. More precisely, the algorithm of [103, 104] can be used to find all small prime factors of a given integer n, and then the remaining factor can be tested for primality with, for example, the celebrated AKS algorithm [3].

Amongst several other number theoretic tools, the approach of [97] is also based on estimates of Soundararajan [143] for smooth numbers in short intervals, bounds of Matomäki [114] for the number of large gaps between consecutive primes.

This direction is certainly open to further improvements and ramifications.

We also mention a recent result of Islam [84] which shows that an interval of the form $[x, x + x^{1/2}]$ contains a product of three relatively prime integers, in fact one of them can be chosen to prime, in prescribed ranges. For example, all of them can be taken in the interval $[0.5x^{1/3}, 2x^{1/3}]$. These results has allowed one to make rigorous some of the heuristic result of [20, 124] that give some reductions between the discrete logarithm and Diffie–Hellman problems on elliptic curves.

2.7 Endomorphism Rings

For a elliptic curve $\mathbf{E}$ over $\mathbb{F}_p$ the endomorphism ring is contained in the quadratic field $\mathbb{Q}(\sqrt{-\Delta_p(\mathbf{E})})$, where $\Delta_p(\mathbf{E}) = \#\mathbf{E}(\mathbb{F}_p) - p - 1$, which is called the *complex multiplication field*. Certainly this field depends on $D_p(\mathbf{E})$, which is a square-free integer defined by the relation $\Delta_p(\mathbf{E}) = D_p(\mathbf{E})d_p(\mathbf{E})^2$ for some integer $d_p(\mathbf{E})$, that is, $\mathbb{Q}(\sqrt{-\Delta_p(\mathbf{E})}) = \mathbb{Q}(\sqrt{-D_p(\mathbf{E})})$.

Luca and Shparlinski [108] have studied the distribution of $D_p(\mathbf{E})$ and $d_p(\mathbf{E})$ when $\mathbf{E}$ runs through the set of all elliptic curves over $\mathbb{F}_p$. It is shown in [108, Theorem 1] that for any positive

$$\delta \ll p^{1/6}(\log p)^{1/3}(\log\log p)^{-2/3} \tag{9}$$

there are at most $p^2\delta^{-1}(\log p)^2$ pairs $(a, b) \in \mathbb{F}_p^2$ for which the Weierstraß equation (1) defines an elliptic curve $\mathbf{E}$ with $d_p(\mathbf{E}) \geq \delta$. This bound is based on estimates

for the number of square divisors of the expressions $t^2 - 4p$. It seems quite plausible that using a result of Friedlander and Iwaniec [75, Bound (4.1)], one can extend the range of δ in (9).

Furthermore, by [108, Corollary 3] the set of all elliptic curves over $\mathbb{F}_p$ defines $2p^{1/2} + O(p^{1/3+o(1)})$ distinct quadratic fields $\mathbb{Q}(\sqrt{-\Delta_p(\mathbf{E})})$. Since any curve $\mathbf{E}$ and its twist $\overline{\mathbf{E}}$ given by (14) have the same value of $\Delta_p(\mathbf{E}) = \Delta_p(\overline{\mathbf{E}})$ and the cardinality $\mathbf{E}$ takes $4p^{1/2} + O(1)$ possible values, this result means that essentially all distinct (up to a twist) curves have distinct complex multiplication fields. The proof of [108, Corollary 3] is based on a improvement [108, Theorem 2] in the error term of the asymptotic formula for the number of solutions to $f(u) = f(v)$, $1 \le u, v \le Z$ for a quadratic polynomial $f(X) \in \mathbb{Z}[X]$ due to Cutter, Granville and Tucker [47]. Improving the error term in this formula and also extending it to polynomials of higher degree is of great interest as it is related to many other number theory problems, see [47].

A related question about the size of the *Tate–Shafarevich group* is considered in [33, 38, 136] and also reduces to the question on the distribution of the largest integer square dividing $t_p^2 - 4p$ when the prime p varies, where $\mathbf{E}$ is defined over $\mathbb{Q}$ (provided that p is large enough so that the reduction of $\mathbf{E}$ modulo p defines an elliptic curve over $\mathbb{F}_p$, otherwise we set $t_p = 0$). The result of Cojocaru and Duke [38] is based on the square sieve (and thus bounds of sums of quadratic characters), and has been improved in [136, Theorem 1.1] via some additional argument. Furthermore, using the same method, in [136, Theorem 1.2] an improvement of a result of Cojocaru and David [37] is given. Unfortunately all the above results are conditional and based on the GRH. No nontrivial unconditional results are known for these problems and obtaining such results certainly requires new ideas.

Square-free values of $t_p^2 - 4p$ have been studied by David and Jiménez Urroz [49] (using some results of [15, 18]) on average over primes p and also over the family of curves $\mathbf{E}_{a,b}$ with coefficients in $|a| \le A$, $|b| \le B$.

2.8 Ranks

We consider the parametric family of curves

$$\mathbf{E}_d: \quad y^2 + xy = x^3 - t^d$$

over the function field $\mathbb{F}_q(t)$, where d is a positive integer. Among other results, Ulmer [145, Proposition 6.4] has shown that the conjecture of Birch and Swinnerton-Dyer holds for each $\mathbf{E}_d$ when d is not divisible p.

For integers a and $m \ge 2$ with $\gcd(a, m) = 1$ we use $t_a(m)$ be the multiplicative order of a modulo m. We also denote by $\mathcal{U}_p$ the set of positive integers which divide some member of the sequence $p^n + 1$, for $n = 1, 2, \ldots$. Ulmer [145, Theorem 9.2] has also shown that for every $d \in \mathcal{U}_p$, the rank $R_q(d)$ of $\mathbf{E}_d$ over $\mathbb{F}_q(t)$ is given by

$$R_q(d) = I_q(d) - C_q(d), \tag{10}$$

where

$$I_q(d) = \sum_{e|d} \frac{\varphi(e)}{t_q(e)}$$

and $C_q(d)$ is an explicit correction term that always satisfies $0 \le C_q(d) \le 4$. (Note that $d \in \mathcal{U}_p$ implies that $\gcd(e, q) = 1$ for each $e \mid d$, so that $I_q(d)$ is defined.) Since members of $\mathcal{U}_p$ are relatively prime to p, the Birch and Swinnerton-Dyer conjecture holds for $\mathbf{E}_d$ for $d \in \mathcal{U}_p$, so that (10) holds as well for the analytic rank.

Ulmer [145] considers the specific case $d = p^n + 1$ and $q = p$. Then $t_p(d) = 2n$, and each $t_p(e) \mid 2n$, so that

$$I_p\left(p^n + 1\right) \ge \sum_{e|p^n+1} \frac{\varphi(e)}{2n} = \frac{p^n + 1}{2n}.$$

Thus,

$$R_p(d) \ge \frac{d \log p}{2 \log d} - 4,$$

which almost achieves the upper bound

$$R_p(d) \le \frac{d \log p}{2 \log d} + O\left(\frac{d(\log p)^2}{(\log d)^2}\right)$$

(uniformly over d and p) due to Brumer [25].

It is shown in [126, Theorem 1] that there exists an absolute constant $\alpha > 1/2$ such that for all finite fields $\mathbb{F}_q$ and all sufficiently large values of x (depending only on the characteristic p of $\mathbb{F}_q$),

$$\frac{1}{x} \sum_{d \le x} R_q(d) \ge x^{\alpha}.$$

Furthermore, by [126, Theorem 2] for any fixed p and $\varepsilon > 0$, as $x \to \infty$ and arbitrary q, we have

$$R_q(d) \ge (\log d)^{(1/3-\varepsilon) \log\log\log d}$$

for all but $o(x)$ values of $d \le x$.

These results depend on a variety of deep number theoretic tools such as the Chebotarev density theorem and the Bombieri–Vinogradov theorem.

The quantity $I_q(d)$ also appears in some other questions such as counting the number of irreducible factors of $T^d - 1$ in $\mathbb{F}_q[T]$ which has been investigated by Moree and Solé [123].

It is noticed in [126] that there are several more families of elliptic curves to which a similar technique can be applied, for example those introduced by Darmon [48].

3 Cryptographic Applications of Elliptic Curves over Finite Fields

3.1 Embedding Degree

Given an elliptic curve $\mathbf{E}$ over $\mathbb{F}_q$, the *embedding degree of a subgroup* $\mathcal{G}$ *of* $\mathbf{E}(\mathbb{F}_q)$ is defined as the smallest positive integer k with

$$\#\mathcal{G} \mid q^k - 1;$$

see [5, 23, 73]. Typically, only subgroups $\mathcal{G}$ of prime order ℓ of $\mathbf{E}(\mathbb{F}_q)$ are of interest.

The importance of the embedding degree for elliptic curve cryptography has been shown by Menezes, Okamoto and Vanstone [118] who noticed that if k is small that the discrete logarithm problem in the group $\mathcal{G} \subseteq \mathbf{E}(\mathbb{F}_q)$ can be reduced to the discrete logarithm problem in the group $\mathbb{F}_{q^k}^*$, where much faster algorithms are known, see [46]. Thus in order to preserve the hardness of the original discrete logarithm problem, this algorithmic speed-up has to be compensated for by the increase in the groups size, that is, by the size of k. Balasubramanian and Koblitz [13] have shown that the critical value of k is around $(\log q)^2$. It is also shown that, for a sufficiently large x, the probability that for a randomly chosen pair $(p, \mathbf{E})$ of a prime $p \in [x/2, x]$ and an elliptic curve $\mathbf{E}$ over $\mathbb{F}_p$, such $\#\mathbf{E}(\mathbb{F}_p)$ is prime (so that we always have $\mathcal{G} = \mathbf{E}(\mathbb{F}_p)$) the embedding degree satisfies $k \le (\log p)^2$ is $O(x^{-1}(\log x)^9(\log\log x)^2)$. Thus for a random elliptic curve with a prime number of points, the approach of [118] succeeds only with negligible probability.

In [106] several similar (albeit weaker) probability estimates have been obtained for every fixed p (that is, without the randomisation over primes in a given interval) and also without limiting to curves $\mathbf{E}$ of prime cardinalities $\#\mathbf{E}(\mathbb{F}_p)$. The results of [106] are, in particular, based on an upper bound:

$$N_p(x, y, z) = \#\big\{N \in [x - y, x + y] \mid \exists k \le z : N \mid p^k - 1\big\}.$$

It is shown in [106] that if $x \ge y \ge 1$, $\log x \asymp \log y \asymp \log p$ and $\log z \ll \log\log p$, then

$$N_p(x, y, z) \le y^{1-1/(2\kappa+3)+o(1)}$$

where

$$\kappa = \frac{\log z}{\log\log p}.$$

This bound is certainly open to further improvements that may lead to some interesting number theoretic questions.

Several more counting results about curves with small embedding degree are given in [44] and [107].

The dual situation when a fixed curve $\mathbf{E}$ defined over $\mathbb{Q}$ gets reduced modulo a random prime has been considered in [44]. More precisely, for such a curve and

positive real numbers T, K, L, let us define

$$R_{\mathbf{E}}(x; y, z) = \#\big\{p \le x : \#\mathbf{E}(\mathbb{F}_p) \ne p \pm 1 \text{ and } \mathbf{E}(\mathbb{F}_p) \text{ has a subgroup } \mathcal{G} \text{ of order } \#\mathcal{G} \ge y \text{ with embedding degree at most } z\big\}.$$

In [44], using some simple number theoretic argument, the bound

$$R_{\mathbf{E}}(x; y, z) \ll x^{3/2} z^2 y^{-1},$$

where $x \ge y > 0$ and $z > 0$ be arbitrary real numbers, has been derived.

The embedding degree for yet another family of curves has been studied in [107]. Namely, given a curve $\mathbf{E}$ over $\mathbb{F}_q$ we denote by $k(q^n)$ the number embedding degree of the group of $\mathbb{F}_{q^n}$-rational points on $\mathbf{E}$. Using an explicit form of the *Subspace Theorem* and estimates on the zeros of linear recurrence sequences, it is shown in [107, Theorem 5.1], that $k(q^n) > (\log q)^{1/6}$ for all $n \le x$ except maybe $o(n)$ of them.

On the other hand, we recall that there are classes of elliptic curves of small embedding degree and in some applications this a very desirable property, see [5, 23, 73].

3.2 Pairing-Friendly Curves

One of the first heuristic constructions of the so-called "pairing-friendly" curves, that is, curves of prime cardinalities and with a reasonably small embedding degree, is due to Miyaji, Nakabayashi and Takano [122].

The construction of [122] produces a family of parameters of elliptic curves of prime cardinalities and of small embedding degrees (namely of embedding degrees $k = 3, 4, 6$) and has been analysed (also heuristically) in [88, 91, 109]. In particular, for $k = 6$ it is shown in [109] that the number of such curves with the CM discriminant at most z is expected (by the order of magnitude) to be given by the sum

$$S(z) = \sum_{\substack{s \le z \\ s \text{ square-free}}} \sum_{n=1}^{\infty} \frac{1}{(\log x_n(s))^2}$$

where $x_n(s)$ be the x-component of the nth largest solution to the Pell equation

$$x^2 - 3sy^2 = -8. \tag{11}$$

It is known that the solutions to the associated *Pell equation*

$$x^2 - 3sy^2 = 1$$

(which actually defines the growth of the solutions to (11)) grow exponentially as a geometric progression $\alpha(s)^n$. Furthermore, it is also known that

$$\alpha(s) \gg s^{1/2} \tag{12}$$

and it is widely believed (see, for example, [102]), that for most s we have a much stronger bound,

$$\alpha(s) \ge \exp\bigl(s^{1/2+o(1)}\bigr). \tag{13}$$

We note that the bound (13) may suggest that

$$S(z) \ll \sum_{s \le z} \frac{1}{\alpha(s)^2} = z^{o(1)}.$$

However, Karabina and Teske [91] have shown that rather rare values of s that achieve the bound (12) provide enough contribution to guarantee that in fact

$$S(z) \gg \frac{z^{1/2}}{(\log z)^2}.$$

Under certain standard number theoretic hypothesis, a slightly stronger bound,

$$S(z) \gg \frac{z^{1/2}}{\log z},$$

has been derived in [88]. Understanding (even heuristically) the true growth of $S(z)$ leads to rather deep questions about the distribution of solutions of the Pell equation. It is possible that recent results of Fouvry [71] may help to shed some light on this question.

We note that there are also many other constructions of "pairing-friendly" curves. Unfortunately all of them are heuristic and based on various heuristic assumptions, see [73] for a detailed survey. It is certainly interesting to give a detailed analysis of these constructions and remove at least some of the heuristic assumptions.

Furthermore, some statistical results about the frequency of pairing friendly elliptic curves are given in [88, 109, 110].

3.3 Elliptic Twins

For a prime power q with $\gcd(q, 6) = 1$ and an elliptic curve $\mathbf{E}$ given by an projective Weierstraß equation (1) over $\mathbb{F}_q$ of q elements, we denote by $\overline{\mathbf{E}}$ its quadratic twist given by the following equation:

$$\overline{\mathbf{E}}: \quad \lambda y^2 = x^3 + ax + b, \tag{14}$$

where $\lambda \in \mathbb{F}_q$ is a quadratic non-residue (all non-residues lead to the same curve), see [142]. It is easy to see that

$$\#\mathbf{E}(\mathbb{F}_q) + \#\overline{\mathbf{E}}(\mathbb{F}_q) = 2(q+1).$$

It has been shown in [29] that curves $\mathbf{E}$ for which the numbers $\#\mathbf{E}(\mathbb{F}_q)$ and $\#\overline{\mathbf{E}}(\mathbb{F}_q)$ of $\mathbb{F}_q$-rational points on both curves are primes can be used for designing efficient cryptographic hash functions.

Accordingly, we call a pair (ℓ, r) of primes *elliptic twins*, if for some prime power q with $\gcd(q, 6) = 1$ and an integer $t \in [-2\sqrt{q}, 2\sqrt{q}]$ we have

$$\ell = q + 1 - t \quad \text{and} \quad r = q + 1 + t.$$

Heuristically, it is natural to believe that there are infinitely many elliptic twins. For example, the *Dickson prime s-tuplets conjecture*, taken with $s = 3$, implies that there are infinitely many prime triples $(q + 1 - t, q, q + 1 + t)$ for any odd t. Although this conjecture has received substantial computational support, and also Balog [14] has given some theoretic results towards this conjecture, still, obtaining a rigorous proof seems out of reach at present.

More precisely, let $W(x)$ denote the total number of pairs (q, t), where $q \le x$ is a prime power and $t \in [-2\sqrt{q}, 2\sqrt{q}]$ is an integer such that $q + 1 - t$ and $q + 1 + t$ are primes.

Friedlander and Shparlinski [77], assuming a rather uniform version of the *Bateman and Horn Conjecture* (see [19]), have derived the following asymptotic formula:

$$W(x) \sim A \frac{x^{\frac{3}{2}}}{(\log x)^3}, \tag{15}$$

where

$$A = 6 \prod_{\substack{\ell \ge 5 \\ \ell \text{ prime}}} \left(1 + \frac{1}{(\ell - 1)^3}\right) = 6.135897\ldots.$$

Clearly $W(x)$ counts isogeny classes corresponding to elliptic twins. Shparlinski and Sutantyo [139] have used a different approach to count, also under the Bateman and Horn Conjecture, isomorphism classes corresponding to elliptic twins. Furthermore, the same approach has also been used in [139] to give an alternative derivation of (15) that is based on a slightly different assumption (thus it can be considered as one more supporting evidence of the validity of (15)).

There are certainly many interesting number theoretic questions associated with this problem; some of them can probably be treated unconditionally (or under some standard hypothesis such the GRH). For example, one can prove that there are sufficiently many $t \in [-2\sqrt{q}, 2\sqrt{q}]$ such that both $P(q + 1 - t)$ and $P(q + 1 + t)$ are large.

3.4 Pseudorandom Number Generators and Hash Functions

There are several constructions which exploit the group structure of elliptic curves to design various pseudorandom number generators, see [134] for a survey and

also [21, 66, 86] for most recent results. Estimating the quality of these constructions is usually based on bounds of various character sums, that use of the estimate of [98] of character sums over the points of a subgroup $\mathcal{G}$ of the group $\mathbf{E}(\mathbb{F}_q)$. In turn, the bound of [98] is based on some results of Bombieri [24] and is nontrivial for subgroups of order $\#\mathcal{G} > q^{1/2+\varepsilon}$ for any fixed $\varepsilon > 0$. Obtaining nontrivial estimates over smaller subgroups is an important and challenging question.

Studying the distribution of the values of some hash functions to elliptic curves has also been based on various bounds of character sums over points of elliptic curves, see [67, 68, 70].

4 Concluding Remarks

The interplay between the theory of elliptic curves, their cryptographic applications and number theory is a new but very exciting area. Certainly, from the cryptographic point of view, the ultimate goal is finding a "better-than-generic" algorithm for the discrete logarithm problem on elliptic curves. This goal is maybe unreachable nowadays, yet its pursuit may lead to many significant theoretic results and also practical attacks, see [1, 54–56, 79, 82].

There are also many other ways where number theory and cryptography can enrich each other. It is enough to recall that the pioneering discovery of Edwards [58] of the family (5) has changed the whole landscape of the area of fast arithmetic on elliptic curves and at the same time provided a protection against timing/power attacks. It is natural to anticipate many more exciting discoveries in this direction.

Finally, we hope that this survey can help number theorists to see a new important direction and a source of new problems. On the other hand, the cryptographers may get a better feeling of what the modern number theory can (or cannot) achieve with respect to their demands.

References

1. L. M. Adleman, J. DeMarrais and M.-D. Huang, 'A subexponential algorithm for discrete logarithms over the rational subgroup of the Jacobians of large genus hyperelliptic curves over finite fields', *Lect. Notes in Comp. Sci.*, vol. **877**, Springer, Berlin, 1994, pp. 28–40.
2. L. M. Adleman, C. Pomerance and R. S. Rumely, 'On distinguishing prime numbers from composite numbers', *Ann. Math.*, **117** (1983), 173–206.
3. M. Agrawal, N. Kayal and N. Saxena, 'PRIMES is in P', *Ann. Math.*, **160** (2004), 781–793.
4. O. Ahmadi and R. Granger, 'On isogeny classes of Edwards curves over finite fields', *J. Number Theory*, **132** (2012), 1337–1358.
5. R. Avanzi, H. Cohen, C. Doche, G. Frey, T. Lange, K. Nguyen and F. Vercauteren, *Handbook of Elliptic and Hyperelliptic Curve Cryptography*, CRC Press, Boca Raton, 2005.
6. S. Baier, 'The Lang–Trotter conjecture on average', *J. Ramanujan Math. Soc.*, **22** (2007), 299–314.
7. S. Baier, 'A remark on the Lang–Trotter conjecture', *Proc. Conf. "New Directions in the Theory of Universal Zeta- and L-Functions"*, Würzburg, Oct. 2008, Ber. Math., Shaker Verlag, Aachen, 2009, pp. 11–18.

8. S. Baier and N. Jones, 'A refined version of the Lang–Trotter conjecture', *Int. Math. Res. Not.*, **3** (2009), 433–461.
9. S. Baier and L. Zhao, 'Bombieri–Vinogradov type theorems for sparse sets of moduli', *Acta Arith.*, **125** (2006), 187–201.
10. S. Baier and L. Zhao, 'An improvement for the large sieve for square moduli', *J. Number Theory*, **128** (2008), 154–174.
11. S. Baier and L. Zhao, 'On the low-lying zeros of Hasse–Weil L-functions for elliptic curves', *Adv. Math.*, **219** (2008), 952–985.
12. S. Baier and L. Zhao, 'The Sato–Tate conjecture on average for small angles', *Trans. Am. Math. Soc.*, **361** (2009), 1811–1832.
13. R. Balasubramanian and N. Koblitz, 'The improbability that an elliptic curve has subexponential discrete log problem under the Menezes–Okamoto–Vanstone algorithm', *J. Cryptology*, **11** (1998), 141–145.
14. A. Balog, 'The prime k-tuplets conjecture on average', *Analytic Number Theory*, Progress in Mathematics, vol. **85**, Birkhäuser, Boston, 1990, pp. 47–75.
15. A. Balog, A. Cojocaru and C. David, 'Average twin prime conjecture for elliptic curves', *Am. J. Math.*, **133** (2011), 1179–1229.
16. R. C. Baker, 'Primes in arithmetic progressions to spaced moduli', *Acta Arith.*, **153** (2012), 133–159.
17. W. D. Banks, F. Pappalardi and I. E. Shparlinski, 'On group structures realized by elliptic curves over arbitrary finite fields', *Exp. Math.*, **21** (2012), 11–25.
18. W. D. Banks and I. E. Shparlinski, 'Sato–Tate, cyclicity, and divisibility statistics on average for elliptic curves of small height', *Israel J. Math.*, **173** (2009), 253–277.
19. P. T. Bateman and R. A. Horn, 'A heuristic asymptotic formula concerning the distribution of prime numbers', *Math. Comput.*, **16** (1962), 363–367.
20. K. Bentahar, 'The equivalence between the DHP and DLP for elliptic curves used in practical applications, revisited', *Lect. Notes in Comp. Sci.*, vol. **3796**, Springer, Berlin, 2005, pp. 376–391.
21. S. R. Blackburn, A. Ostafe and I. E. Shparlinski, 'On the distribution of the subset sum pseudorandom number generator on elliptic curves', *Unif. Distrib. Theory*, **6** (2011), 127–142.
22. I. Blake, G. Seroussi and N. Smart, *Elliptic Curves in Cryptography*, London Math. Soc., Lecture Note Series, vol. **265**, Cambridge University Press, Cambridge, 1999.
23. I. Blake, G. Seroussi and N. Smart, *Advances in Elliptic Curves in Cryptography*, London Math. Soc., Lecture Note Series, vol. **317**, Cambridge University Press, Cambridge, 2005.
24. E. Bombieri, 'On exponential sums in finite fields', *Am. J. Math.*, **88** (1966), 71–105.
25. A. Brumer, 'The average rank of elliptic curves I', *Invent. Math.*, **109** (1992), 445–472.
26. W. Castryck, A. Folsom, H. Hubrechts and A. V. Sutherland, 'The probability that the number of points on the Jacobian of a genus 2 curve is prime', *Proc. Lond. Math. Soc.*, **104** (2012), 1235–1270.
27. W. Castryck and H. Hubrechts, 'The distribution of the number of points modulo an integer on elliptic curves over finite fields', *Preprint*, 2009 (available from http://arxiv.org/abs/0902.4332).
28. M.-C. Chang, J. Cilleruelo, M. Z. Garaev, J. Hernandez, I. E. Shparlinski and A. Zumalacárregui, 'Points on curves in small boxes en applications', *Preprint*, 2011 (available from http://arxiv.org/abs/1111.1543).
29. O. Chevassut, P.-A. Fouque, P. Gaudry and D. Pointcheval, 'The twist-AUgmented technique for key exchange', *Lect. Notes in Comp. Sci.*, vol. **3958**, Springer, Berlin, 2006, pp. 410–426.
30. J. Cilleruelo and M. Z. Garaev, 'Concentration of points on two and three dimensional modular hyperbolas and applications', *Geom. Funct. Anal.*, **21** (2011), 892–904.
31. J. Cilleruelo, M. Z. Garaev, A. Ostafe and I. E. Shparlinski, 'On the concentration of points of polynomial maps and applications', *Math. Z.* (to appear).
32. J. Cilleruelo, I. E. Shparlinski and A. Zumalacárregui, 'Isomorphism classes of elliptic curves over a finite field in some thin families', *Math. Res. Lett.*, **19** (2012), 335–343.

33. A. Cojocaru, 'Questions about the reductions modulo primes of an elliptic curve', *Proc. 7th Meeting of the Canadian Number Theory Association (Montreal, 2002)*, CRM Proceedings and Lecture Notes, vol. **36**, Am. Math. Soc., Providence, 2004, pp. 61–79.
34. A. Cojocaru, 'On the cyclicity of the group of $\mathbb{F}_p$-rational points of non-CM elliptic curves', *J. Number Theory*, **96** (2002), 335–350.
35. A. Cojocaru, 'Cyclicity of CM elliptic curves modulo p', *Trans. Am. Math. Soc.*, **355** (2003), 2651–2662.
36. A. Cojocaru, 'Reductions of an elliptic curve with almost prime orders', *Acta Arith.*, **119** (2005), 265–289.
37. A. C. Cojocaru and C. David, 'Frobenius fields for elliptic curves', *Am. J. Math.*, **130** (2008), 1535–1560.
38. A. Cojocaru and W. Duke, 'Reductions of an elliptic curve and their Tate–Shafarevich groups', *Math. Ann.*, **329** (2004), 513–534.
39. A. Cojocaru, É. Fouvry and M. R. Murty, 'The square sieve and the Lang–Trotter conjecture', *Canadian J. Math.*, **57** (2005), 1155–1177.
40. A. Cojocaru and C. Hall, 'Uniform results for Serre's theorem for elliptic curves', *Int.. Math. Res. Not.*, **2005** (2005), 3065–3080.
41. A. C. Cojocaru, F. Luca and I. E. Shparlinski, 'Pseudoprime reductions of elliptic curves', *Math. Proc. Cambridge Philos. Soc.*, **146** (2009), 513–522.
42. A. Cojocaru and M. R. Murty, 'Cyclicity of elliptic curves modulo p and elliptic curve analogues of Linnik's problem', *Math. Ann.*, **330** (2004), 601–625.
43. A. Cojocaru and I. E. Shparlinski, 'Distribution of Farey fractions in residue', *Proc. Am. Math. Soc.*, **136** (2008), 1977–1986.
44. A. C. Cojocaru and I. E. Shparlinski, 'On the embedding degree of reductions of an elliptic curve', *Inf. Proc. Lett.*, **109** (2009), 652–654.
45. P. Corvaja and U. Zannier, 'A lower bound for the height of a rational function at S-unit points', *Monatsh. Math.*, **144** (2005), 203–224.
46. R. Crandall and C. Pomerance, *Prime Numbers: A Computational Perspective*, Springer, New York, 2004.
47. P. Cutter, A. Granville and T. J. Tucker, 'The number of fields generated by the square root of values of a given polynomial', *Can. Math. Bull.*, **46** (2003), 71–79.
48. H. Darmon, 'Heegner points and elliptic curves of large rank over function fields', *Heegner Points and Rankin L-Series*, Math. Sci. Res. Inst. Publ., vol. **49**, Cambridge University Press, Cambridge, 2004, pp. 317–322.
49. C. David and J. Jiménez Urroz, 'Square-free discriminants of Frobenius rings', *Int. J. Number Theory*, **6** (2010), 1391–1412.
50. C. David and E. Smith, 'Elliptic curves with a given number of points over finite fields', *Compos. Math.* (to appear).
51. C. David and E. Smith, 'A Cohen–Lenstra phenomenon for elliptic curves', *Preprint*, 2012 (available from http://arxiv.org/abs/1206.1585).
52. C. David and J. Wu, 'Pseudoprime reductions of elliptic curves', *Can. J. Math.*, **64** (2012), 81–101.
53. C. David and J. Wu, 'Almost-prime orders of elliptic curves over finite fields', *Forum Math.*, **24** (2012), 99–120.
54. C. Diem, 'The GHS Attack in odd characteristic', *J. Ramanujan Math. Soc.*, **18** (2003), 1–32.
55. C. Diem, 'On the discrete logarithm problem in elliptic curves', *Compos. Math.*, **147** (2011), 75–104.
56. C. Diem, 'On the discrete logarithm problem in elliptic curves', *Preprint*, 2011.
57. W. Duke, 'Almost all reductions modulo p of an elliptic curve have a large exponent', *C. R. Math.*, **337** (2003), 689–692.
58. H. M. Edwards, 'A normal form for elliptic curves', *Bull. Am. Math. Soc.*, **44** (2007), 393–422.
59. J.-H. Evertse, 'An improvement of the quantitative subspace theorem', *Compos. Math.*, **101** (1996), 225–311.

60. J. S. Ellenberg and A. Venkatesh, 'Reflection principles and bounds for class group torsion', *Int. Math. Res. Not.*, **2007** (2007), 1–18. Article ID rnm002
61. J.-H. Evertse and H. P. Schlickewei, 'A quantitative version of the absolute subspace theorem', *J. Reine Angew. Math.*, **548** (2002), 21–127.
62. R. R. Farashahi, 'On the number of distinct Legendre, Jacobi, Hessian and Edwards curves', *Preprint*, 2011 (available from http://arxiv.org/abs/1112.5714).
63. R. R. Farashahi, D. Moody and H. Wu, 'Isomorphism classes of Edwards curves over finite fields', *Finite Fields Their Appl.*, **18** (2012), 597–612.
64. R. R. Farashahi and I. E. Shparlinski, 'On the number of distinct elliptic curves in some families', *Designs Codes Cryptogr.*, **54** (2010), 83–99.
65. R. R. Farashahi and I. E. Shparlinski, 'On group structures realized by elliptic curves over a fixed finite field', *Exp. Math.*, **21** (2012), 1–10.
66. R. R. Farashahi and I. E. Shparlinski, 'Pseudorandom bits from points on elliptic curves', *IEEE Trans. Inf. Theory*, **58** (2012), 1242–1247.
67. R. R. Farashahi, P.-A. Fouque, I. E. Shparlinski, M. Tibouchi and J. F. Voloch, 'Indifferentiable deterministic hashing to elliptic and hyperelliptic curves', *Math. Comput.* (to appear).
68. R. R. Farashahi, I. E. Shparlinski and J. F. Voloch, 'On hashing into elliptic curves', *J. Math. Cryptol.*, **3** (2009), 353–360.
69. K. Ford and I. E. Shparlinski, 'On finite fields with Jacobians of small exponent', *Int. J. Number Theory*, **4** (2008), 819–826.
70. P.-A. Fouque and M. Tibouchi, 'Estimating the size of the image of deterministic hash functions to elliptic curves', *Lect. Notes in Comp. Sci.*, vol. **6212**, Springer, Berlin, 2010, pp. 81–91.
71. É. Fouvry, 'On the size of the fundamental solution of Pell equation', *Preprint*, 2011.
72. É. Fouvry and M. R. Murty, 'On the distribution of supersingular primes', *Can. J. Math.*, **48** (1996), 81–104.
73. D. Freeman, M. Scott and E. Teske, 'A taxonomy of pairing-friendly elliptic curves', *J. Cryptol.*, **23** (2010), 224–280.
74. T. Freiberg and P. Kurlberg, 'On the average exponent of elliptic curves modulo p', *Preprint*, 2012 (available from http://arxiv.org/abs/1203.4382).
75. J. B. Friedlander and H. Iwaniec, 'Square-free values of quadratic polynomials', *Proc. Edinburgh Math. Soc.*, **53** (2010), 385–392.
76. J. B. Friedlander, C. Pomerance and I. E. Shparlinski, 'Finding the group structure of elliptic curves over finite fields', *Bull. Aust. Math. Soc.*, **72** (2005), 251–263.
77. J. B. Friedlander and I. Shparlinski, 'Elliptic twin prime conjecture', *Lect. Notes in Comp. Sci.*, vol. **5557**, Springer, Berlin, 2009, pp. 77–81.
78. S. D. Galbraith and J. McKee, 'The probability that the number of points on an elliptic curve over a finite field is prime', *J. Lond. Math. Soc.*, **62** (2000), 671–684.
79. P. Gaudry, F. Hess, and N. Smart, 'Constructive and destructive facets of Weil descent', *J. Cryptol.*, **15** (2002), 19–46.
80. D. R. Heath-Brown, 'The average analytic rank of elliptic curves', *Duke Math. J.*, **122** (2004), 591–623.
81. H. A. Helfgott and A. Venkatesh, 'Integral points on elliptic curves and 3-torsion in class groups', *J. Am. Math. Soc.*, **19** (2006), 527–550.
82. F. Hess, 'Computing relations in divisor class groups of algebraic curves over finite fields', *Preprint*, 2005.
83. E. W. Howe, 'On the group orders of elliptic curves over finite fields', *Compos. Math.*, **85** (1993), 229–247.
84. A. Islam, 'Products of three pairwise coprime integers in short intervals', *LMS J. Comput. Math.*, **15** (2012) 59–70.
85. H. Iwaniec and E. Kowalski, *Analytic Number Theory*, Am. Math. Soc., Providence, 2004.
86. D. Jao, D. Jetchev and R. Venkatesan, 'On the bits of the elliptic curve Diffie–Hellman secret keys', *Lect. Notes in Comp. Sci.*, vol. **4859**, Springer, Berlin, 2007, pp. 33–47.

87. J. Jiménez Urroz, 'Almost prime orders of CM elliptic curves modulo p', *Lect. Notes in Comp. Sci.*, vol. **5011**, Springer, Berlin, 2008, pp. 74–87.
88. J. Jiménez Urroz, F. Luca and I. E. Shparlinski, 'On the number of isogeny classes of pairing-friendly elliptic curves and statistics of MNT curves', *Math. Comput.*, **81** (2012), 1093–1110.
89. N. Jones, 'Averages of elliptic curve constants', *Math. Ann.*, **345** (2009), 685–710.
90. N. Jones, 'Almost all elliptic curves are Serre curves', *Trans. Am. Math. Soc.*, **362** (2010), 1547–1570.
91. K. Karabina and E. Teske, 'On prime-order elliptic curves with embedding degrees $k = 3, 4$ and 6', *Lect. Notes in Comput. Sci.*, vol. **5011**, Springer, Berlin, 2008, pp. 102–117.
92. S. Kim, 'On the average exponent of CM elliptic curves modulo p', *Preprint*, 2012 (available from http://arxiv.org/abs/1207.6652).
93. N. Koblitz, 'Elliptic curve cryptosystems', *Math. Comput.*, **48** (1987), 203–209.
94. N. Koblitz, 'Primality of the number of points on an elliptic curve over a finite field', *Pacific J. Math.*, **131** (1988), 157–166.
95. N. Koblitz, 'Elliptic curve implementation of zero-knowledge blobs', *J. Cryptol.*, **4** (1991), 207–213.
96. N. Koblitz, 'Almost primality of group orders of elliptic curves defined over small finite fields', *Exp. Math.*, **10** (2001), 553–558.
97. N. Koblitz, A. Menezes and I. E. Shparlinski, 'Discrete logarithms, Diffie–Hellman, and reductions', *Vietnam J. Math.*, **39** (2011), 267–285.
98. D. R. Kohel and I. E. Shparlinski, 'Exponential sums and group generators for elliptic curves over finite fields', *Lect. Notes in Comp. Sci.*, vol. **1838**, Springer, Berlin, 2000, pp. 395–404.
99. J. C. Lagarias, H. L. Montgomery and A. M. Odlyzko, 'A bound for the least prime ideal in the Chebotarev density theorem', *Invent. Math.*, **54** (1979), 271–296.
100. S. Lang, *Elliptic Curves: Diophantine Analysis*, Springer, Berlin, 1978.
101. H. W. Lenstra, 'Factoring integers with elliptic curves', *Ann. Math.*, **126** (1987), 649–673.
102. H. W. Lenstra, 'Solving the Pell equation', *Not. Am. Math. Soc.*, **49** (2002), 182–192.
103. H. W. Lenstra, J. Pila, and C. Pomerance, 'A hyperelliptic smoothness test. I', *Phil. Trans. R. Soc. Lond. (A)*, **345** (1993), 397–408.
104. H. W. Lenstra, Jr., J. Pila, and C. Pomerance, 'A hyperelliptic smoothness test. II', *Proc. Lond. Math. Soc.*, **84** (2002), 105–146.
105. F. Luca, J. McKee and I. E. Shparlinski, 'Small exponent point groups on elliptic curves', *J. Théor. Nr. Bordx.*, **17** (2005), 859–870.
106. F. Luca, D. J. Mireles and I. E. Shparlinski, 'MOV attack in various subgroups on elliptic curves', *Illinois J. Math.*, **48** (2004), 1041–1052.
107. F. Luca and I. E. Shparlinski, 'On the exponent of the group of points on elliptic curves in extension fields', *Int. Math. Res. Not.*, **2005** (2005), 1391–1409.
108. F. Luca and I. E. Shparlinski, 'Discriminants of complex multiplication fields of elliptic curves over finite fields', *Can. Math. Bull.*, **50** (2007), 409–417.
109. F. Luca and I. E. Shparlinski, 'Elliptic curves with low embedding degree', *J. Cryptol.*, **19** (2006), 553–562.
110. F. Luca and I. E. Shparlinski, 'On finite fields for pairing based cryptography', *Adv. Math. Commun.*, **1** (2007), 281–286.
111. F. Luca and I. E. Shparlinski, 'On the counting function of elliptic Carmichael numbers', *Canad. Math. Bull.* (to appear).
112. C. Magagna, 'A lower bound for the r-order of a matrix modulo N', *Monats. Math.*, **153** (2008), 59–81.
113. G. Martin, P. Pollack and E. Smith, 'Averages of the number of points on elliptic curves ', *Preprint*, 2012 (available from http://arxiv.org/abs/1208.0919).
114. K. Matomäki, 'Large differences between consecutive primes', *Quart. J. Math.*, **58** (2007), 489–518.
115. K. Matomäki, 'A note on primes of the form $p = aq^2 + 1$', *Acta Arith.*, **137** (2009), 133–137.
116. U. M. Maurer, 'On the oracle complexity of factoring integers', *Comput. Complex.*, **5** (1996), 237–247.

117. J. McKee, 'Subtleties in the distribution of the numbers of points on elliptic curves over a finite prime field', *J. Lond. Math. Soc.*, **59** (1999), 448–460.
118. A. Menezes, T. Okamoto and S. A. Vanstone, 'Reducing elliptic curve logarithms to logarithms in a finite field', *IEEE Trans. Inform. Theory*, **39** (1993), 1639–1646.
119. L. Merel, 'Bornes pour la torsion des courbes elliptiques sur les corps de nombres', *Invent. Math.*, **124** (1996), 437–449.
120. V. S. Miller, 'Uses of elliptic curves in cryptography', *Lect. Notes in Comp. Sci.*, vol. **218**, Springer, Berlin, 1986, pp. 417–426.
121. V. S. Miller, 'The Weil pairing, and its efficient calculation', *J. Cryptol.*, **17** (2004), 235–261.
122. A. Miyaji, M. Nakabayashi and S. Takano, 'New explicit conditions of elliptic curve traces for FR-reduction', *IEICE Trans. Fundam.*, **E84-A** (2001), 1234–1243.
123. P. Moree and P. Solé, 'Around Pelikán's conjecture on very odd sequences', *Manuscr. Math.*, **117** (2005), 219–238.
124. A. Muzereau, N. P. Smart and F. Vercauteren, 'The equivalence between the DHP and DLP for elliptic curves used in practical applications', *LMS J. Comput. Math.*, **7** (2004) 50–72.
125. F. Pappalardi, 'On the exponent of the group of points of an elliptic curve over a finite field', *Proc. Am. Math. Soc.*, **139** (2011), 2337–2341.
126. C. Pomerance and I. E. Shparlinski, 'Rank statistics for a family of elliptic curves over a function field', *Pure Appl. Math. Quart.*, **6** (2010), 21–40.
127. A. J. van der Poorten and H. P. Schlickewei, 'Zeros of recurrence sequences', *Bull. Aust. Math. Soc.*, **44** (1991), 215–223.
128. H.-G. Rück, 'A note on elliptic curves over finite fields', *Math. Comput.*, **49** (1987), 301–304.
129. R. Schoof, 'Elliptic curves over finite fields and the computation of square roots mod p', *Math. Comput.*, **44** (1985), 483–494.
130. R. Schoof, 'The exponents of the group of points on the reduction of an elliptic curve', *Arithmetic Algebraic Geometry*, Progr. Math., vol. **89**, Birkhäuser, Boston, 1991, pp. 325–335.
131. V. Shoup, 'Searching for primitive roots in finite fields', *Math. Comput.*, **58** (1992), 369–380.
132. I. Shparlinski, 'On primitive elements in finite fields and on elliptic curves', *Mat. Sb.*, **181** (1990), 1196–1206 (in Russian).
133. I. E. Shparlinski, 'Orders of points on elliptic curves', *Affine Algebraic Geometry*, Am. Math. Soc., Providence, 2005, pp. 245–252.
134. I. E. Shparlinski, 'Pseudorandom number generators from elliptic curves', *Recent Trends in Cryptography*, Contemp. Math., vol. **477**, Am. Math. Soc., Providence, 2009, pp. 121–141.
135. I. E. Shparlinski, 'Exponents of modular reductions of families of elliptic curves', *Rev. Unión Mat. Argent.*, **50** (2009), 69–74.
136. I. E. Shparlinski, 'Tate–Shafarevich groups and Frobenius fields of reductions of elliptic curves', *Quart. J. Math.*, **61** (2010), 255–263.
137. I. E. Shparlinski, 'On the Sato–Tate conjecture on average for some families of elliptic curves', *Forum Math.* (to appear).
138. I. E. Shparlinski, 'On the Lang–Trotter and Sato–Tate conjectures on average for polynomial families of elliptic curves', *Preprint*, 2012 (available from http://arxiv.org/abs/1203.6500).
139. I. E. Shparlinski and A. V. Sutantyo, 'Distribution of elliptic twin primes in isogeny and isomorphism classes', *Preprint*, 2012.
140. I. E. Shparlinski and A. V. Sutherland, 'On the distribution of Atkin and Elkies primes', *Preprint*, 2011 (available from http://arxiv.org/abs/1112.3390).
141. I. E. Shparlinski and F. Voloch, 'Generators of elliptic curves over finite fields', *Preprint*, 2011.
142. J. H. Silverman, *The Arithmetic of Elliptic Curves*, Springer, Berlin, 1995.
143. K. Soundararajan, 'Smooth numbers in short intervals', *Preprint*, 2010 (available from http://arxiv.org/abs/1009.1591).

144. M. A. Tsfasman and S. G. Vlăduţ, *Algebraic-Geometric Codes*, Kluwer Academic, Dordrecht, 1991.
145. D. Ulmer, 'Elliptic curves with large rank over function fields', *Ann. Math.*, **155** (2002), 295–315.
146. S. G. Vlăduţ, 'Cyclicity statistics for elliptic curves over finite fields', *Finite Fields Their Appl.*, **5** (1999), 13–25.
147. S. G. Vlăduţ, 'On the cyclicity of elliptic curves over finite field extensions', *Finite Fields Their Appl.*, **5** (1999), 354–363.
148. J. F. Voloch, 'A note on elliptic curves over finite fields', *Bull. Soc. Math. Fr.*, **116** (1988), 455–458.
149. A. Weng, 'On group orders of rational points of elliptic curves', *Quaest. Math.*, **25** (2002), 513–525.
150. J. Wu, 'The average exponent of elliptic curves modulo p', *Preprint*, 2012 (available from http://arxiv.org/abs/1206.5928).
151. M. Young, 'Low-lying zeros of families of elliptic curves', *J. Am. Math. Soc.*, **19** (2006), 205–250.
152. D. Zywina, 'A refinement of Koblitz's conjecture', *Int. J. Number Theory*, **7** (2011), 739–769.

Random Matrix Theory and Its Innovative Applications

Alan Edelman and Yuyang Wang

Abstract Recently more and more disciplines of science and engineering have found Random Matrix Theory valuable. Some disciplines use the limiting densities to indicate the cutoff between "noise" and "signal." Other disciplines are finding eigenvalue repulsions a compelling model of reality. This survey introduces both the theory behind these applications and MATLAB experiments allowing a reader immediate access to the ideas.

1 Random Matrix Theory in the Press

Since the beginning of the 20th century, Random matrix theory (RMT) has been finding applications in number theory, quantum mechanics, condensed matter physics, wireless communications, etc., see [7, 12, 15, 16]. Recently more and more disciplines of science and engineering have found RMT valuable (Fig. 1). New applications in RMT are being found every day, some of them surprising and innovative when compared with the older applications.

For newcomers to the field, it may be reassuring to know that very little specialized knowledge of random matrix theory is required for applications, and therefore the "learning curve" to become a user is not at all steep. Two methodologies are worth highlighting.

1. *Distinguishing "signal" from "noise"*: Generate a matrix of data specific to your application (e.g. a correlation matrix or a sample covariance matrix) and perhaps normalize the data to have mean 0 and variance 1. Compare the answer to the known example of singular values or eigenvalues of random matrices. (Usually

Note to our readers: This survey paper is in large part a precursor to a book on Random Matrix Theory that will be forthcoming. We reserve the right to reuse materials in the book.

A. Edelman
Department of Mathematics, MIT, Cambridge, USA
e-mail: edelman@math.mit.edu

Y. Wang (✉)
Department of Computer Science, United States Tufts University, Medford, USA
e-mail: ywang02@cs.tufts.edu

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5_5,

Fig. 1 New Scientist cover story entitled *Entering the matrix: the simple idea at the root of reality*. Quoting Raj Rao Nadakuditi: "It really does feel like the ideas of random matrix theory are somehow buried deep in the heart of nature"

it is enough to know the quarter circle law, the semi-circle law, and the Marčenko-Pastur Laws, see Sect. 2 for details.)

(1a) "*No correlation*": If the answer obtained is similar enough to one of the known laws in RMT, one might declare the data to be all noise. In one example (Fig. 2), Popoff et al. [20] use the fact that the distribution of the singular values of the transmission matrix follows a quarter circle law (see Sect. 2) to show that the matrix elements are not significantly correlated, thereby justifying the fact that their experimental procedure does not introduce spurious correlations.

(1b) "*Correlations*": If there are singular values/eigenvalues larger than those predicted by RMT, one considers those as indicative of "signal" or correlations in the data. Most recently, Arup Chakraborty, a Chemistry and Chemical Engineering professor at MIT and his colleagues used RMT (Fig. 3) to find sectors of HIV that rarely undergo multiple mutations [8]. The mathematical procedure is relatively easy to explain and apply. It is known that if we take a matrix `A = randn(m, n)` (in MATLAB notation), which is an $m \times n$ random matrix with independent and identically distributed (i.i.d.) standard normal, then the eigenvalues of the Wishart matrix $A^T A/m$ in the limit as $m/n = r$ and $m, n \to \infty$ are almost surely in the interval $[(1-\sqrt{r})^2, (1+\sqrt{r})^2]$. Thus if we have a correlation matrix with eigenvalues larger than $(1+\sqrt{r})^2$ we consider these as signal rather than noise. Theoretical understanding of the "signal" eigenvalues may be found in [19].

2. *Spacing Distributions*: Generate a statistic that one considers likely to be represented by the spacings, or the largest eigenvalue, or the smallest eigenvalue of

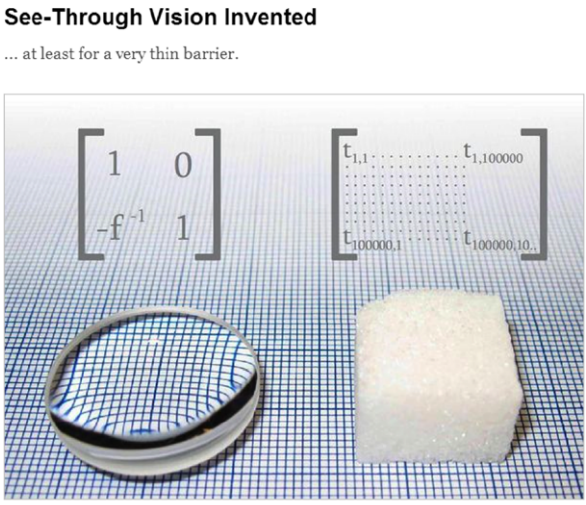


Fig. 2 Comparing the singular values of a transmission matrix to that of a random matrix suggests that there are no spurious correlations

a random matrix. These might be recognized respectively by a statistic that is feeling repulsions from two sides, that is feeling pushed outward from behind, or pushed inward towards a "hard edge."

(2a) *Repulsion from two sides*: The repulsion from two edges is given by what is known as the bulk distribution spacings (Fig. 4) and is described in [21].

(2b) *Repulsion from "behind" with no barrier*: The repulsion from behind is given by the Tracy-Widom distributions ($\beta = 2$ is the most common application, but $\beta = 1$ also shows up). In a remarkable first of a kind experiment [23], Kazumasa Takeuchi and Masaki Sano have measured the interface in turbulent liquid crystal growth. Two kinds of experiments are reported, flat and curved interfaces. Careful measurements of the first four moments of the random radii match those of the largest eigenvalues of real and complex matrices respectively. These are the Tracy-Widom laws. (See Sect. 2 for the complex Tracy-Widom law.) The use of the third and fourth moments, in the form of skewness and kurtosis, indicate that the Tracy-Widom law really appears to be applying in practice. In other results where

Fig. 3 Random matrix techniques in a recent study featured in the Wall Street Journal

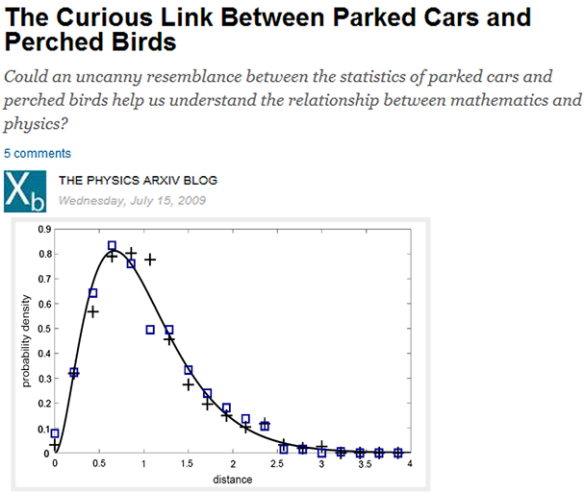


Fig. 4 Random matrix theory spacings shows itself in the gaps between parked cars, the distances between perched birds (illustrated above), the distribution peaks that neutron scatter off heavy nuclei, etc.

this same law is conjectured, statistics sometimes seem indistinguishable from the normal distribution.

(2c) *Repulsion with barrier*: The repulsion towards the hard edge has been less commonly applied, but corresponds to a "Bessel kernel."

The rest of this paper is organized as follows. In Sect. 2, we introduce two classical random matrix ensembles, namely, Hermite ensembles and Laguerre ensembles. Further, we describe the limiting eigenvalue densities. Section 3 starts with a numerical question of how to calculate the eigenvalues of a random matrix efficiently. Then, we discuss theoretical implications of the computational trick.

2 Famous Laws in RMT with MATLAB Experiments

We introduce the classic random matrix ensembles and then we will provide four famous laws in RMT with corresponding MATLAB experiments. Notice that although measure-theoretical probability is not required to understand and appreciate the beauty of RMT in this paper, the extension of probabilistic measure-theoretical tools to matrices is nontrivial, we refer interested readers to [2, 6].

While we expect our readers to be familiar with real and complex matrices, it is reasonable to consider quaternion matrices as well. Let us start with the Gaussian random matrices $G_1(m, n)$ (`G1 = randn(m, n)`), which is an $m \times n$ matrix with i.i.d. standard real random normals. In general, we use the parameter β to denote the number of standard real normals and thus $\beta = 1, 2, 4$ correspond to real, complex and quaternion respectively. $G_\beta(m, n)$ can be generated by the MATLAB command shown in Table 1. Notice that since quaternions do not exist in MATLAB they are "faked" using 2×2 complex matrices.

Table 1 Generating the Gaussian random matrix $G_\beta(m,n)$ in MATLAB

β	MATLAB command
1	`G = randn(m, n)`
2	`G = randn(m, n) + i*randn(m, n)`
4	`X = randn(m, n) + i*randn(m, n);` `Y = randn(m, n) + i*randn(m, n);G = [X Y; -conj(Y) conj(X)]`

Table 2 Hermite and Laguerre ensembles

Ensemble	Matrices	Weight function	Equilibrium measure	Numeric	MATLAB
Hermite	Wigner	$e^{-x^2/2}$	Semi-circle	`eig`	`g = G(n,n);` `H=(g+g')/2`
Laguerre	Wishart	$x^{v/2-1}e^{-x/2}$	Marcenko-Pastur	`svd`	`g = G(m,n);` `L=(g'*g)/m;`

If A is an $m \times n$ Gaussian random matrix $G_\beta(m,n)$ then its joint element density is given by

$$\frac{1}{(2\pi)^{\beta mn/2}} \exp\left(-\frac{1}{2}\|A\|_F^2\right), \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

The most important property of G_β, be it real, complex, or quaternion, is its *orthogonal invariance*. This makes the distribution impervious to multiplication by an orthogonal (unitary, symplectic) matrix, provided that the two are independent. This can be inferred from the joint element density in (1) since its Frobenius norm, $\|A\|_F$, is unchanged when A is multiplied by an orthogonal (unitary, symplectic) matrix. The orthogonal invariance implies that no test can be devised that would differentiate between Q_1A, A, and AQ_2, where Q_1 and Q_2 are non-random orthogonal and A is Gaussian. Readers will later see that this simple property leads to wonderful results both in practice and in theory.

The most well-studied random matrices have names such as Gaussian, Wishart, MONOVA, and circular. We prefer Hermite, Laguerre, Jacobi, and perhaps Fourier. In a sense, they are to random matrix theory as Poisson's equation is to numerical methods. Of course, we are thinking in the sense of the problems that are well-tested, well-analyzed, and well-studied because of nice fundamental analytic properties. These matrices play a prominent role because of their deep mathematical structure. There are four channels of structure lurking underneath numeric analysis, graph theory, multivariate statistics [17] and operator algebras [18]. In this paper, we will focus on the Hermite and Laguerre ensembles, which is summarized in Table 2. The other random matrix ensembles are discussed in details in [10].

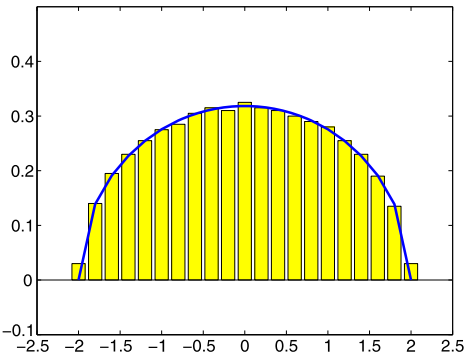


Fig. 5 Semi-circle law with one 1000×1000 matrix. Plotted is the histogram of the 1000 eigenvalues and the semi-circle on the *blue line*

2.1 The Most Famous Semi-circle Law

In [24], Wigner originally showed that the limiting eigenvalue distribution of simple random symmetric $n \times n$ matrices $X = (A + A^T)/2$ where $A = G_1(n, n)$, follow a semi-circle distribution (Fig. 5), which is given by

$$p(x) = \frac{1}{2\pi}\sqrt{4 - x^2}. \tag{2}$$

When properly normalized, the curve looks like a semi-circle of radius 2. This distribution depicts the histogram of the n eigenvalues of a symmetric random $n \times n$ matrix obtained by symmetrizing a matrix of random normals. X constructed in this way is called the β-*Hermite ensemble* or Gaussian ensemble, more specifically *Gaussian orthogonal ensemble (GOE)* ($\beta = 1$), *Gaussian unitary ensemble (GUE)* ($\beta = 2$) and *Gaussian symplectic ensemble (GSE)* ($\beta = 4$). Code 1 histograms the random eigenvalues and plots the semi-circle. The mathematical theorem requires only one matrix $t = 1$ and $n \to \infty$, though the computer is happier with much smaller values for n.

Theorem 1 (Wigner 1955) *Let X_n be a sequence of random symmetric $n \times n$ matrices ($n = 1, 2, \ldots$), satisfying*

1. *Independent elements (up to matrix symmetry): The elements x_{ij} for $i \le j$ of each X_n are independent random variables.*[1]
2. *Zero Mean: The elements x_{ij} of each X_n satisfy $\mathbb{E}(x_{ij}) = 0$.*
3. *Unit off diagonal variance (for normalization): The elements x_{ij} of each X_n satisfy $\mathbb{E}(x_{ij}^2) = 1$.*
4. *Bounded moments: For $k = 1, 2, \ldots$, there is some bound*[2] *B_k, independent of n, such that for all $m \le k$, $\mathbb{E}(|x_{ij}|^m) \le B_k$.*

[1] Strictly speaking, the random variables should be written $x_{ij}(n)$.

[2] If the $x_{ij}(n)$ are identically distributed, a very common assumption, then it is sufficient to require finite moments.

Code 1 Semicircle Law (Random symmetric matrix eigenvalues)

```
%Experiment:  Gaussian Random Symmetric Eigenvalues
%Plot:        Histogram of the eigenvalues
%Theory:      Semicircle as n->infinity
%% Parameters
n=1000;        %matrix size
t=1;           %trials
v=[];          %eigenvalue samples
dx=.2;         %binsize
%% Experiment
for i=1:t,
    a=randn(n);     % random nxn matrix
    s=(a+a')/2;     % symmetrized matrix
    v=[v;eig(s)];   % eigenvalues
end
v=v/sqrt(n/2);      % normalized eigenvalues
%% Plot
[count,x]=hist(v,-2:dx:2);
cla reset
bar(x, count/(t*n*dx),'y');
hold on;
%% Theory
plot(x,sqrt(4-x.^2)/(2*pi),'LineWidth',2)
axis([-2.5 2.5 -.1 .5]);
```

Under these assumptions, the distribution of the eigenvalues of such $\frac{1}{\sqrt{n}}X_n$ *asymptotically approaches a semi-circle distribution* (2) *in the following sense. As* $n \to \infty$, $\mathbb{E}[\lambda^k]$ *matches the moments of the semicircle distribution for a randomly chosen* λ *from n eigenvalues.*

Wigner's original proof is combinatorial; he showed that the significant terms in $\mathbb{E}[\mathbf{Tr}(X^{2k})]$ count the ordered trees on $k+1$ vertices. This count is well-known today as the *Catalan numbers*, where the nth Catalan number is

$$C_n = \frac{1}{n+1}\binom{2n}{n}. \tag{3}$$

The moments of a semi-circle distribution are the very same Catalan numbers.

Furthermore, the eigenvalue density of the β-*Hermite ensemble* ($\beta = 1, 2, 4$) almost surely converges to the semi-circle distribution. The proof can be found in Chap. 2 of [5], which also discusses the state of the art knowledge concerning the assumption of bounded moments. Roughly speaking, if the independent elements are i.i.d., the finite moment condition can be dropped. If not, there is a weaker condition by Girko that is claimed to be an if and only if.

The semi-circle law acts like a central limit theorem for (infinitely) large symmetric random matrices. If we average a large number of independent and identi-

cally distributed random matrices, the classical central limit theorem says that the elements become Gaussians.

Another important problem is the convergence rate of the (empirical) eigenvalue density, which was answered by Bai in [3, 4]. We refer interested readers to Chap. 8 of [5].

On the other hand, in the finite case where n is given, one may wonder how are the eigenvalues of an $n \times n$ symmetric random matrix A distributed? Fortunately, for the Hermite ensemble, the answer is known explicitly and the density is called the *level density* in the Physics literature [15]. It is worth mentioning that the joint element density of an $n \times n$ matrix A_β from the Hermite ensemble is [9]

$$\frac{1}{2^{n/2}} \frac{1}{\pi^{n/2+n(n-1)\beta/4}} \exp\left(-\frac{1}{2}\|A\|_F^2\right), \tag{4}$$

and the joint eigenvalue probability density function is

$$f_\beta(\lambda_1, \ldots, \lambda_n) = c_H^\beta \prod_{i<j} |\lambda_i - \lambda_j|^\beta \exp\left(-\sum_{i=1}^n \frac{\lambda_i^2}{2}\right), \tag{5}$$

with

$$c_H^\beta = (2\pi)^{-n/2} \prod_{j=1}^n \frac{\Gamma(1+\frac{\beta}{2})}{\Gamma(1+\frac{\beta}{2}j)}.$$

The level density ρ_n^A for an $n \times n$ ensemble A with real eigenvalues is the distribution of a random eigenvalue chosen from the ensemble. More precisely, the level density can be written in terms of the marginalization of the joint eigenvalue density. For example, in the Hermite ensemble case,

$$\rho_{n,\beta}^A(\lambda_1) = \int_{\mathbb{R}^{n-1}} f_\beta(\lambda_1, \ldots, \lambda_n) \mathrm{d}\lambda_2 \cdots \mathrm{d}\lambda_n. \tag{6}$$

In the following part, we will show the exact semi-circle for the GUE case and give numerical approaches that calculate the level density efficiently. Notice that such formulas also exist for the finite GOE and GSE, we refer interested readers to [15].

If A is a $n \times n$ complex Gaussian and we take $(A + A^T)/2$ which is an instance of the GUE ensemble, the eigenvalue density was derived by Wigner in 1962 as $\sum_{j=0}^{n-1} \phi_j^2(x)$, where

$$\phi_j(x) = \left(2^j j! \sqrt{\pi}\right)^{-\frac{1}{2}} \exp\left(-x^2/2\right) H_j(x) \tag{7}$$

and $H_j(x)$ is the jth Hermite polynomial, which is defined as

$$H_j(x) = \exp\left(x^2\right)\left(-\frac{\mathrm{d}}{\mathrm{d}x}\right)^j \exp\left(-x^2\right) = j! \sum_{i=0}^{j/2} (-1)^i \frac{(2x)^{j-2i}}{i!(j-2i)!}.$$

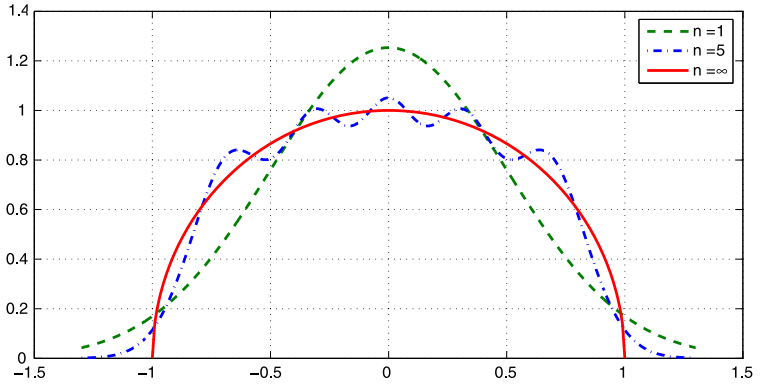


Fig. 6 Level density of the GUE ensemble ($\beta = 2$) for different values of n. The limiting result when $n \to \infty$ is Wigner's famous semi-circle law

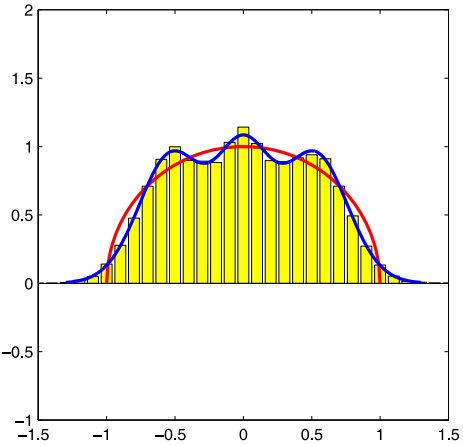


Fig. 7 The exact semicircular law with ten thousand matrices. Plotted is the histogram (30000 eigenvalues) on the real line. The unit semi-circle and the exact density function which is defined in terms of Hermite polynomials (Code 2)

Figure 6 compares the normalized level density of the GUE for different values of n.

Most useful for computation is the three term recurrence

$$H_{j+1}(x) = 2xH_j(x) - 2jH_{j-1}(x), \tag{8}$$

starting with $H_{-1} = 0$ and $H_0 = 1$ so that $H_1(x) = 2x$. Therefore, ignoring the normalization term $(\sqrt{\pi})^{-\frac{1}{2}} \exp(-x^2/2)$ in $\phi_j(x)$, we define

$$\widetilde{\phi}_j(x) = \left(2^j j!\right)^{-\frac{1}{2}} H_j(x).$$

From (8), we can get the three term recurrence for $\widetilde{\phi}_j(x)$ as follows

$$\sqrt{j} \cdot \widetilde{\phi}_j(x) = \sqrt{2}x \cdot \widetilde{\phi}_{j-1}(x) - \sqrt{j-1} \cdot \widetilde{\phi}_{j-2}(x). \tag{9}$$

Based on (9), one can do the direct calculation of summing $\phi_n(x)$ for each x. But there are two better ways to calculate the level density.

Code 2 Calculating the level density for the finite GUE

```
%Experiment:   Eigenvalues of GUE matrices
%Plot:         Histogram of eigenvalues
%Theory:       Semicircle and finite semicircle
%% Parameters
n=3;           % size of matrix
s=10000;       % number of samples
d=.1;          % bin size
e=[];          % eigenvalue samples
%% Experiment
for i=1:s
    a=randn(n)+sqrt(-1)*randn(n);
    a=(a+a')/(2*sqrt(4*n));
    v=eig(a);
    e=[e v];
end
[m x]=hist(e,-1.5:d:1.5);
bar(x,m*pi/(2*d*n*s),'y');
axis('square')
axis([-1.5 1.5 -1 2])
%% Theory
hold on
t=-1:.01:1;
plot(t,sqrt(1-t.^2),'r', 'Linewidth', 2) % Plot the semicircle
levels(n)                                % Plot the finite
    semicircle
hold off
```

1. The first approach is based on the following equation

$$\sum_{j=0}^{n-1} \widetilde{\phi}_j^2(x) = n\widetilde{\phi}_n^2(x) - \sqrt{n(n+1)}\widetilde{\phi}_{n-1}(x)\widetilde{\phi}_{n+1}(x). \tag{10}$$

   This formula comes from the famous Christoffel-Darboux relationship for orthogonal polynomials. Therefore, we can combine (10) with the three term recurrence for $\widetilde{\phi}$ and Code 3 realizes the idea.
2. The other way comes from the following interesting equivalent expression

$$\sum_{j=0}^{n-1} \phi_j^2(x) = \left\| (\sqrt{\pi})^{-\frac{1}{2}} \exp\left(-x^2/2\right) \cdot v \right\|^2, \tag{11}$$

   where

$$v = \frac{u}{u_1}, \quad u = (T - \sqrt{2}x \cdot \mathbb{I})^{-1} e_{n-1}, \tag{12}$$

Code 3 Computing the level density (GUE) using Christoffel-Darboux

```
function z=levels(n)
%Plot exact semicircle formula for GUE
x=[-1:.001:1]*sqrt(2*n)*1.3;
pold = 0*x;                          % -1st Hermite polynomial
p= 1+0*x;                            %  0th Hermite polynomial
k=p;
for j=1:n;                           % Three term recurrence
    pnew = (sqrt(2)*x.*p-sqrt(j-1)*pold)/sqrt(j);
    pold = p; p=pnew;
end
pnew = (sqrt(2)*x.*p-sqrt(n)*pold)/sqrt(n+1);
k = n*p.^2-sqrt(n*(n+1))*pnew.*pold; % Use p.420 of Mehta
% Multiply the correct normalization
k=k.*exp(-x.^2)/sqrt(pi);
% Rescale so that "semicircle" is on [-1,1] and area is pi/2
plot(x/sqrt(2*n),k*pi/sqrt(2*n),'b', 'Linewidth', 2);
```

where u_1 is the first element of u and e_{n-1} is the column vector where only the $n-1$st entry is 1. T is a tridiagonal matrix that is related to the three term recurrence such that

$$T = \begin{bmatrix} 0 & \sqrt{1} & 0 & 0 & \cdots & 0 \\ \sqrt{1} & 0 & \sqrt{2} & 0 & \cdots & 0 \\ 0 & \sqrt{2} & 0 & \sqrt{3} & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{n-1} & 0 \end{bmatrix}.$$

To see this, from (9), we have the following relation

$$\begin{bmatrix} -\sqrt{2}x & \sqrt{1} & 0 & \cdots & 0 \\ \sqrt{1} & -\sqrt{2}x & \sqrt{2} & \cdots & 0 \\ 0 & \sqrt{2} & -\sqrt{2}x & \sqrt{3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sqrt{n-1} & -\sqrt{2}x \end{bmatrix} \begin{bmatrix} \widetilde{\phi}_0(x) \\ \widetilde{\phi}_1(x) \\ \widetilde{\phi}_2(x) \\ \vdots \\ \widetilde{\phi}_{n-1}(x) \end{bmatrix} = C \times \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

where C can be determined easily by the initial condition $\widetilde{\phi}_0(x) = 1$ which justifies (12).

Though one can easily use (11) to compute the density at x, we can avoid inverting $T - \sqrt{2}x \cdot \mathbb{I}$ for every x. Given the eigendecomposition of $T = H\Lambda H^T$, we have

$$\begin{aligned} u &= (T - \sqrt{2}x \cdot \mathbb{I})^{-1} e_{n-1} \\ &= H(\Lambda - \sqrt{2}x \cdot \mathbb{I})^{-1} H^T e_{n-1}. \end{aligned}$$

Code 4 Computing the level density (GUE) using the tridiagonal matrix

```
function z = levels2(n)
%% Plot exact semicircle formula for GUE
xfull = [-1:.001:1] * sqrt(2*n) * 1.3;
% Form the Tridiagonal matrix
T = diag(sqrt(1:n-1),1);
T = T+T';
% Do the eigendecomposition of T, T = UVU'
[U, V] = eig(T);
% extract the eigenvalues
V = diag(V);
% precompute U'*e_n
% tmp_en = U' * ((0:n-1) == n-1)';
tmp_en = U(end, :)';
for i = 1:length(xfull),
    x = xfull(i);
    % generate the v vector as in (2.5)
    v = U * (tmp_en./(V - sqrt(2)*x));
    % multiply the normalization term
    y(i) = norm((sqrt(pi))^(-1/2) * exp(-x^2/2) * v/v(1))^2;
end
% Rescale so that "semicircle" is on [-1,1] and area is pi/2
plot(xfull/sqrt(2*n), y*pi/sqrt(2*n), 'r—', 'Linewidth', 2);
```

Thus, for each x, we only need to invert the diagonal matrix $\Lambda - \sqrt{2}x \cdot \mathbb{I}$, provided that H is stored beforehand. Code 4 gives the corresponding implementation.

2.2 Marčenko-Pastur Law (Special Case: Quarter Circle Law)

Laguerre ensemble consists of random matrices $L = A^T A/m$ where $A = G_\beta(m, n)$ and the following theorem gives the limiting eigenvalue distribution of its eigenvalues [14] when $\beta = 1$ (real case).

Theorem 2 (Marčenko, Pastur 1967) *Let X_n be a sequence of random symmetric $m \times n$ matrices $(n = 1, 2, \ldots)$, with $m \geq n$, satisfying*

1. *Independence*: *The elements x_{ij} of each $X_n$are independent random variables.*
2. *Zero Mean*: *The elements x_{ij} of each X_n satisfy $\mathbb{E}(x_{ij}) = 0$.*
3. *Unit variance*: *The elements x_{ij} of each X_n satisfy $\mathbb{E}(x_{ij}^2) = 1$.*
4. *Bounded moments*: *There is some bound B, independent of n, such that $\forall n, \mathbb{E}(|x_{ij}|^k) \leq B$.*
5. *Asymptotic Aspect Ratio*: *m depends on n in such a way that $n/m \to r \leq 1$ as $n \to \infty$.*

Under these assumptions, the distribution of the eigenvalues of $\frac{1}{m}X_n^T X_n$ *asymptotically approaches the Marčenko-Pastur law as* $n \to \infty$,

$$f(x) = \frac{\sqrt{(x-a)(b-x)}}{2\pi xr},$$

where $a = (1-\sqrt{r})^2$ *and* $b = (1+\sqrt{r})^2$.

According to the Marčenko-Pastur Law, we have the density of the singular values of $X/\sqrt{m}$ as

$$f(s) = \frac{\sqrt{(s^2-a^2)(b^2-s^2)}}{\pi sr}.$$

When $r = 1$, we get the special case that

$$f(s) = \frac{1}{\pi}\sqrt{4-s^2},$$

on $[0, 2]$. This is the famous *quarter circle law*. The singular values of a normally distributed square matrix lie on a quarter circle. The moments are Catalan numbers. We provide the code of the eigenvalue formulations in Code 5 and the singular value formulation in Code 6 with figures shown in Fig. 8.

2.3 Circular Law

The eigenvalues of Hermite and Laguerre ensembles are distributed on the real line. In general, an interesting question is that when properly normalized, how are the eigenvalues `randn(n)` distributed on the complex plain. The following theorem provides the answer [13].

Theorem 3 (Girko 1984) *The complex eigenvalues divided by* $\sqrt{n}$ *of an* $n \times n$ *random matrix with independent elements of mean* 0 *and variance* 1 *converge* (*under reasonable conditions*) *to the uniform distribution on the unit disk in the complex plane.*

The "Saturn effect" (Fig. 9 Right): Notice the concentration of eigenvalues on the real line and the gap near the real line. The real line serves as an attractor to some of the eigenvalues. Two things are worth mentioning:

1. The Saturn effect is consistent with the circular law. As $n \to \infty$, the $O(\sqrt{n})$ eigenvalues on the real line do not matter. Also the diminishing repulsion is consistent with the circular law.
2. There are $O(\sqrt{n})$ real eigenvalues sometimes clashes with our intuition. After all, the real line is a set of measure 0. Why should there be so many eigenvalues on the real line?

Code 5 Marčenko-Pastur Law

```
%Experiment: Gaussian Random
%Plot: Histogram of the eigenvalues of X'X/m
%Theory: Marcenko-Pastur as n->infinity
%% Parameters
t=1; %trials
r=0.1;     %aspect ratio
n=2000; %matrix column size
m=round(n/r);
v=[]; %eigenvalue samples
dx=.05; %binsize
%% Experiment
for i=1:t,
    X=randn(m,n); % random mxn matrix
    s=X'*X; %sym pos def  matrix
    v=[v;eig(s)]; % eigenvalues
end
v=v/m; % normalized eigenvalues
a=(1-sqrt(r))^2;b=(1+sqrt(r))^2;
%% Plot
[count,x]=hist(v,a:dx:b);
cla reset
bar(x, count/(t*n*dx),'y');
hold on;
%% Theory
x=linspace(a,b);
plot(x,sqrt((x-a).*(b-x))./(2*pi*x*r),'LineWidth',2)
axis([0 ceil(b) -.1 1.5]);
```

2.4 Tracy-Widom Distribution (Law)

Tracy-Widom law provides the limiting density for the largest eigenvalue of the Hermite ensemble (Fig. 10). The probability density for the Tracy-Widom distribution is given by the formula

$$p(x) = \frac{\mathrm{d}}{\mathrm{d}x} \exp\left(- \int_x^\infty (t-x) q(t)^2 \mathrm{d}t \right),$$

where $q(t)$, is defined as the solution of a so-called Painlevé II differential equation:

$$\ddot{q}(t) = t q(t) + 2 q(t)^3,$$

with the boundary condition that as $t \to \infty$, $q(t)$ is asymptotic to the Airy function $\mathrm{Ai}(t)$.

While this may seem more formidable than the normal and semi-circle distributions, there are codes that may be used as black boxes for accurately calculating the Tracy-Widom distribution. This distribution depicts the histogram of the largest eigenvalue of a complex version of the random symmetric matrices. The distribution

Code 6 Quarter circle law

```
%Experiment: Gaussian Random
%Plot: Histogram singular values
%Theory: Quatercircle Law
%% Parameters
t=1; %trials
r=1;     %aspect ratio
n=2000; %matrix column size
m = n;
v=[]; %eigenvalue samples
dx=.05; %binsize
a = 0; b = 2;
%% Experiment
for i=1:t,
  v=[v;svd(randn(n))]; % singular values
end
v=v/sqrt(m); % normalized
close all;
[count, x]=hist(v,(a-dx/2):dx:b); cla reset
bar(x, count/(t*n*dx),'y'); hold on;
%% Theory
x=linspace(a,b);
plot(x,sqrt(4 - x.^2)/pi,'LineWidth',2)
axis square
axis([0 2 0 2/3]);
```

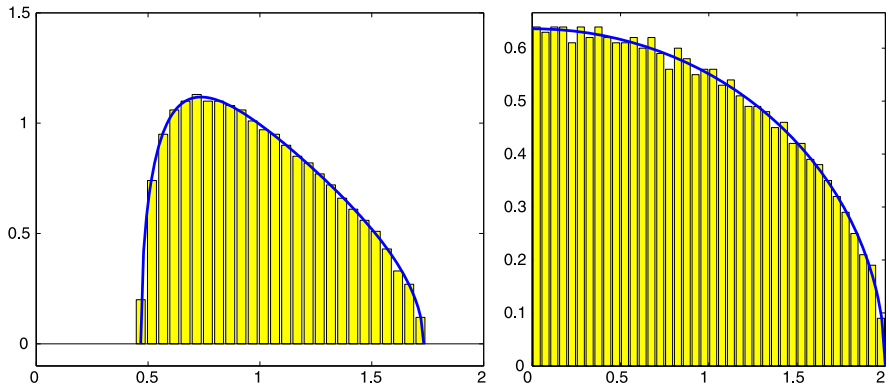


Fig. 8 *Left*: Marčenko-Pastur Law ($r = 0.1$) with a 20000×2000 matrix X. Plotted is the histogram of the 2000 eigenvalues of $X^T X/20000$; *Right*: Quarter Circle Law ($r = 1$) with a 2000×2000 matrix. Plotted are histograms of its singular values

has also been showing up in many other applications. We show in Code 7 that even the formidable is but a few lines of MATLAB. It is based on solving the following

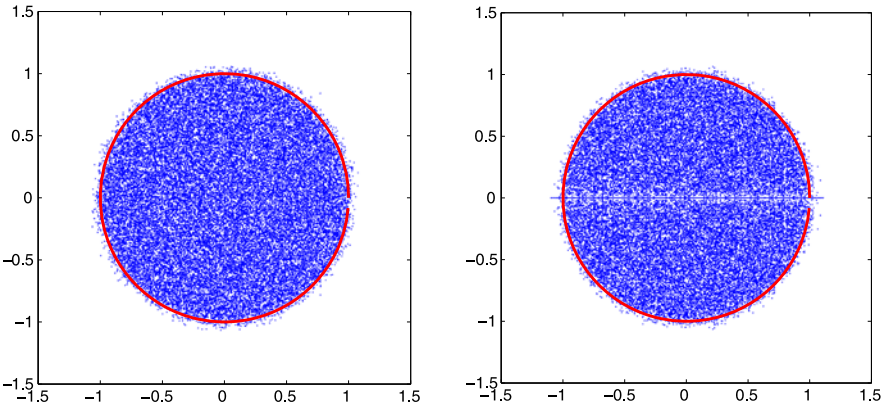


Fig. 9 Circular Law for the real and the complex case with 200 200 × 200 random matrices. Plotted is eigenvalues in the complex plain (40000 eigenvalues)

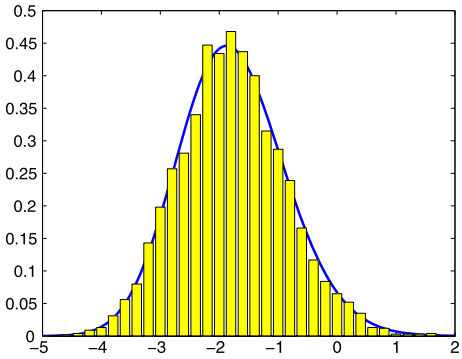


Fig. 10 Tracy-Widom law with 5000 matrices. Plotted is the histogram of the 500000 eigenvalues and the Tracy-Widom distribution on the *blue line*

differential equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{Bmatrix} q \\ q' \\ I \\ I' \end{Bmatrix} = \begin{Bmatrix} q' \\ tq + 2q^3 \\ I' \\ q^2 \end{Bmatrix},$$

where $I(x) = \int_x^\infty (t-x)q(t)^2 \mathrm{d}t$. This has the advantage of evolving the needed quantity $I(x)$, as we go rather than post-processing. In Code 8, we calculate the largest eigenvalue of an instance from the 2-Hermite ensemble (GUE). We normalize the eigenvalues by subtracting them by $2\sqrt{n}$ and multiplying them by $n^{1/6}$.

Code 7 Calculating the Tracy-Widom Distribution

```
%Theory: Compute and Plot the Tracy-Widom Distribution
%%Parameters
t0=5;     %right endpoint
tn=-8;    %left endpoint
dx=.005; %discretization
%%Theory: The differential equation solver
deq=@(t,y) [y(2); t*y(1)+2*y(1)^3; y(4); y(1)^2];
opts=odeset('reltol',1e-12,'abstol',1e-15);
y0=[airy(t0);airy(1,t0);0;airy(t0)^2]; % boundary conditions
[t,y]=ode45(deq,t0:-dx:tn,y0,opts); %solve
F2=exp(-y(:,3));     % the distribution
f2=gradient(F2,t);  % the density
%% Plot
%cla reset
plot(t,f2,'LineWidth',2)
axis([-5 2 0 .5]);
```

Code 8 Largest Eigenvalues of a random Hermitian matrix

```
%Experiment:  Largest Eigenvalue of Random Hermitian Matrices
%Plot:        Histogram of the normalized largest eigenvalues
%Theory:      Tracy-Widom as n->infinity
%% Parameters
n=100;        %matrix size
t=5000;        %trials
v=[];         %eigenvalue samples
dx=.2;        %binsize
%% Experiment
for i=1:t,
    a=randn(n)+sqrt(-1)*randn(n);     % random nxn complex matrix
    s=(a+a')/2;           % Hermitian matrix
    v=[v;max(eig(s))];  % Largest Eigenvalue
end
v=n^(1/6)*(v-2*sqrt(n));    % normalized eigenvalues
%% Plot
hold on
[count,x]=hist(v,-5:dx:2);
bar(x, count/(t*dx),'y');
```

3 Random Matrix Factorization

A computational trick can also be a theoretical trick. Therefore do not dismiss an efficient computation as a mere “implementation detail”, it may be where the next theory comes from.

Direct random matrix experiments usually involve `randn(n)`. Since many linear algebra computations require $\mathcal{O}(n^3)$ operations, it seems more feasible to take n relatively small, and take a large number of Monte Carlo instances. This has been our strategy in the example codes so far.

In fact, matrix computations involve a series of reductions. With normally distributed matrices, the most expensive reduction steps can be avoided on the computer as they can be done with mathematics! All of a sudden $\mathcal{O}(n^3)$ computations become $\mathcal{O}(n^2)$ or even better.

3.1 The Chi-Distribution and Orthogonal Invariance

There are two key facts to know about a vector of independent standard normals. Let v_n denote such a vector. In MATLAB this would be `randn(n,1)`. Mathematically, we say that the n elements are independent and i.i.d. standard normals (mean 0, variance 1).

- *Chi distribution*: the Euclidean length $\|v_n\|$, which is the square root of the sum of the n squares of Gaussians, has what is known as the χ_n distribution.
- *Orthogonal invariance*: for any fixed orthogonal matrix Q, or if Q is random and independent of v_n, the distribution of Qv_n is identical to that of v_n. In other words, it is impossible to tell the difference between a computer-generated v_n or Qv_n upon inspecting only the output.

We shall see that these two facts allow us to very powerfully transform matrices involving standard normals to simpler forms. For reference, we mention that the χ_n distribution has the probability density

$$f(x) = \frac{x^{n-1}e^{-x^2/2}}{2^{n/2-1}\Gamma(n/2)}. \tag{13}$$

There is no specific requirement that n be an integer, despite our original motivation as the length of a Gaussian vector. The square of χ_n is the distribution that underlies the well known *Chi-squared test*. It can be seen that the mean of χ_n^2 is n. (For integers, it is the sum of the n standard normal variables.) We have that v_n is the product of the random scalar χ_n, which serves as the length, and an independent vector that is uniform on the sphere, which serves as the direction.

3.2 The QR Decomposition of `randn(n)`

Given a vector v_n, we can readily construct an orthogonal reflection or rotation H_n such that $H_n v_n = \pm\|v_n\|e_1$, where e_1 denotes the first column of the identity. In

matrix computations, there is a standard technique known as constructing a *Householder transformation* which is a reflection across the external angle bisector of these two vectors.

Therefore, if v_n follows a multivariate standard normal distribution, $H_n v_n$ yields a Chi distribution for the first element and 0 otherwise. Furthermore, let `randn(n)` be an $n \times n$ matrix of i.i.d. standard normals. It is easy to see now that through successive Householder reflections of size $n, n-1, \ldots$ we can orthogonally transform `randn(n)` into the upper triangular matrix

$$H_1 H_2 \cdots H_{n-1} H_n \times \texttt{randn(n)}$$
$$= R_n = \begin{pmatrix} \chi_n & G & G & \ldots & G & G & G \\ & \chi_{n-1} & G & \ldots & G & G & G \\ & & \chi_{n-2} & \ldots & G & G & G \\ & & & \ddots & \vdots & \vdots & \vdots \\ & & & & \chi_3 & G & G \\ & & & & & \chi_2 & G \\ & & & & & & \chi_1 \end{pmatrix}.$$

Here all elements are independent and represent a distribution and the "G"'s are all i.i.d. standard normals. It is helpful to watch a 3×3 real Gaussian matrix ($\beta = 1$) matrix turn into R:

$$\begin{pmatrix} G & G & G \\ G & G & G \\ G & G & G \end{pmatrix} \to \begin{pmatrix} \chi_3 & G & G \\ 0 & G & G \\ 0 & G & G \end{pmatrix} \to \begin{pmatrix} \chi_3 & G & G \\ 0 & \chi_2 & G \\ 0 & 0 & G \end{pmatrix} \to \begin{pmatrix} \chi_3 & G & G \\ 0 & \chi_2 & G \\ 0 & 0 & \chi_1 \end{pmatrix}.$$

The "G"'s as the computation progresses are not the same numbers, merely indicating the distribution. One immediate consequence is the following interesting fact

$$\mathbb{E}\big[\mathbf{det}\big(\texttt{randn(n)}^2\big)\big] = n! \tag{14}$$

This could also be obtained for any $n \times n$ matrix with independent entries with mean 0 and variance 1, by squaring the "big formula" for the determinant, noting that cross terms have expectation 0, and the $n!$ squared terms each have expectation 1.

3.2.1 Haar Measure on Orthogonal Matrices

Let Q be a random orthogonal matrix, such that one can not tell the difference between the distribution of AQ and Q for any fixed orthogonal matrix A. We say that Q has the uniform or *Haar distribution* on orthogonal matrices.

From the previous part of this section, with a bit of care we can say that `randn(n)` = (orthogonal uniform with Haar measure)(R_n) is the QR decomposition of `randn(n)`. Therefore, code for generating Q can be as simple as `[Q,~]=qr(randn(n))`.

Code 9 Sample a random unitary matrix

```
%Experiment:  Generate random orthogonal/unitary matrices
%Plot:        Histogram eigenvalues
%Theory:      Eigenvalues are on unit circle
%% Parameters
t=5000;          %trials
dx=.05;          %binsize
n=10;            %matrix size
v=[];            %eigenvalue samples
%% Experiment
for i=1:t
    % Sample random unitary matrix
    [X ~]=qr(randn(n)+sqrt(-1)*randn(n));
    % If you have non-uniformly sampled eigenvalues, you may
        need this fix
    X=X*diag(sign(randn(n,1)+sqrt(-1)*randn(n,1)));
    v=[v;eig(X)];
end
%% Plot
x=(-(1+dx/2):dx:(1+dx/2))*pi;
h1=rose(angle(v),x);
set(h1,'Color','black')
%% Theory
hold on
h2=polar(x, t*n*dx/2*x.^0);
set(h2,'LineWidth',2)
hold off
```

Similarly, `[Q,~]=qr(randn(n)+sqrt(-1)*randn(n))` gives a random unitary matrix Q. For unitary matrix Q, its eigenvalues will be complex with a magnitude of 1, i.e. they will be distributed on the unit circle in the complex plane. Code 9 generates a random unitary matrix and histograms the angles of its eigenvalues.

3.2.2 Longest Increasing Subsequence

There is an interesting link between the moments of the eigenvalues of Q and the number of permutations of length with longest increasing subsequence k. For example, the permutation (3 1 8 4 5 7 2 6 9 10) has (1 4 5 7 9 10) or (1 4 5 6 9 10) as the longest increasing subsequence of length 6. For $n = 4$, there are 24 possible permutations listed in Table 3. We underline the fourteen permutations with longest increasing subsequence of length 2. Of these, one permutation (4 3 2 1) has length 1 and the other thirteen have length 2.

Given a permutation of the numbers from 1 through n, the longest increasing subsequence may be found with the following admittedly cryptic algorithm "Patience sort" (Code 10).

Table 3 Permutations for $n = 4$

1 2 3 4	2 1 3 4	3 1 2 4	4 1 2 3
1 2 4 3	2 1 4 3	3 1 4 2	4 1 3 2
1 3 2 4	2 3 1 4	3 2 1 4	4 2 1 3
1 3 4 2	2 3 4 1	3 2 4 1	4 2 3 1
1 4 2 3	2 4 1 3	3 4 1 2	4 3 1 2
1 4 3 2	2 4 3 1	3 4 2 1	4 3 2 1

A remarkable result from random matrix theory is that the number of permutations of length n with longest increasing subsequence less than or equal to length k is given by

$$\mathbb{E}_{Q_k} \left| \mathbf{Tr}(Q_k) \right|^{2n}, \tag{15}$$

where Q_k is a $k \times k$ random unitary matrix (Code 11). Who would have thought the moments of a random unitary matrix would count such a combinatorial object? Notice that the length of the permutation is n. Permutations of size 10 indicate the 20th moment of the absolute trace. The size of the matrix is the length of the longest increasing subsequence. Sometimes this may seem backwards, but it is correct.

3.3 The Tridiagonal Reductions of GOE

Eigenvalues are usually defined early in one's education as the roots of the characteristic polynomial. Many people just assume that this is the definition that is used during a computation, but it is well established that this is not a good method for computing eigenvalues. Rather, a matrix factorization is used. In the case that S is symmetric, an orthogonal matrix Q is found such that $Q^T S Q = \Lambda$ is diagonal. The columns of Q are the eigenvectors and the diagonal of Λ are the eigenvalues.

Code 10 Patience sort

```
function z=patiencesort(p)
% Patience sort
%
% Parameters
%   p : Permutation
% Returns
%   z : Length of longest increasing subsequence
piles=[];
for i=1:length(p)
    piles(1+sum(p(i)>piles))=p(i);
end
z=length(piles);
```

Code 11 Random Orthogonal matrices and the Longest increasing sequence

```
%Experiment:  Counts longest increasing subsequence statistics
t=200000;     % Number of trials
n=4;          % permutation size
k=2;          % length of longest increasing subsequence
v=zeros(t,1); % samples
for i=1:t
    [X,DC]=qr(randn(k)+sqrt(-1)*randn(k));
    X=X*diag(sign(randn(k,1)+sqrt(-1)*randn(k,1)));
    v(i)=abs(trace(X))^(2*n);
end
z = mean(v);
p = perms(1:n); c = 0;
for i=1:factorial(n)
    c = c + (patiencesort(p(i,:)) <= k);
end
[z c]
```

Mathematically, the construction of Q is an iterative procedure, requiring infinitely many steps to converge. In practice, S is first tridiagonalized through a finite process which usually takes the bulk of the time. The tridiagonal is then iteratively diagonalized. Usually, this takes a negligible amount of time to converge in finite precision.

If $X =$ `randn(n)` and $S = (X + X^T)/\sqrt{2}$, then the eigenvalues of S follow the semi-circle law while the largest one follows the Tracy-Widom law. We can tridiagonalize S with the finite Householder procedure. The result is

$$T_n = \begin{pmatrix} G\sqrt{2} & \chi_{n-1} & & & & & & \\ \chi_{n-1} & G\sqrt{2} & \chi_{n-2} & & & & & \\ & \chi_{n-2} & G\sqrt{2} & \chi_{n-3} & & & & \\ & & \chi_{n-3} & \ddots & \ddots & & & \\ & & & \ddots & G\sqrt{2} & \chi_2 & & \\ & & & & \chi_2 & G\sqrt{2} & \chi_1 & \\ & & & & & \chi_1 & G\sqrt{2} \end{pmatrix}, \tag{16}$$

where $G\sqrt{2}$ refers to a Gaussian with mean 0 and variance 2. The superdiagonal and diagonal are independent, as the matrix is symmetric. The matrix T_n has the same eigenvalue distribution as S, but numerical computation of the eigenvalues is considerably faster when the right software is used.

A dense eigensolver requires $\mathcal{O}(n^3)$ operations and will spend nearly all of its time constructing T_n. Given that we know the distribution for T_n a priori, this is wasteful. The eigenvalues of T_n require $\mathcal{O}(n^2)$ time or better. In addition, dense matrix requires $\mathcal{O}(n^2)$ in storage while the tridiagonal matrix only needs $\mathcal{O}(n)$.

In a similar fashion, we can compute the singular values of a rectangular $m \times n$ matrix considerably faster by reducing it to bidiagonal form (shown here for $n > m$), as follows

$$B_n = \begin{pmatrix} \chi_m & \chi_{n-1} & & & & & & & \\ & \chi_{m-1} & \chi_{n-2} & & & & & & \\ & & \chi_{m-2} & \chi_{n-3} & & & & & \\ & & & \ddots & \ddots & & & & \\ & & & & \chi_3 & \chi_{n-m+1} & & & \\ & & & & & \chi_2 & \chi_{n-m} & & \\ & & & & & & \chi_1 & \chi_{n-m-1} & \end{pmatrix}.$$

The story gets better. Random matrix experiments involving complex numbers or even over the quaternions reduce to real matrices even before they need to be stored on a computer. For general G_β, the R_n, tridiagonal and bidiagonal reduction have the following extensions

$$R_n = \begin{pmatrix} \chi_{n\beta} & G_\beta & G_\beta & \dots & G_\beta & G_\beta & G_\beta \\ & \chi_{(n-1)\beta} & G_\beta & \dots & G_\beta & G_\beta & G_\beta \\ & & \chi_{(n-2)\beta} & \dots & G_\beta & G_\beta & G_\beta \\ & & & \ddots & \vdots & \vdots & \vdots \\ & & & & \chi_{3\beta} & G_\beta & G_\beta \\ & & & & & \chi_{2\beta} & G_\beta \\ & & & & & & \chi_\beta \end{pmatrix},$$

$$T_n = \begin{pmatrix} G\sqrt{2} & \chi_{(n-1)\beta} & & & & & & \\ \chi_{(n-1)\beta} & G\sqrt{2} & \chi_{(n-2)\beta} & & & & & \\ & \chi_{(n-2)\beta} & G\sqrt{2} & \chi_{(n-3)\beta} & & & & \\ & & \chi_{(n-3)\beta} & \ddots & \ddots & & & \\ & & & \ddots & G\sqrt{2} & \chi_{2\beta} & & \\ & & & & \chi_{2\beta} & G\sqrt{2} & \chi_\beta & \\ & & & & & \chi_\beta & G\sqrt{2} & \end{pmatrix},$$

$$B_n = \begin{pmatrix} \chi_{m\beta} & \chi_{(n-1)\beta} & & & & & & \\ & \chi_{(m-1)\beta} & \chi_{(n-2)\beta} & & & & & \\ & & & \ddots & \ddots & & & \\ & & & & & \chi_{2\beta} & \chi_{(n-m)\beta} & \\ & & & & & & \chi_\beta & \chi_{(n-m-1)\beta} \end{pmatrix}.$$

Of interest is that T_n and B_n are real matrices whose eigenvalue and singular value distributions are exactly the same as the original complex and quaternion matrices. This leads to even greater computational savings because only real numbers need to be stored or computed with. Table 4 summaries how to generate instances from Hermite and Laguerre ensemble efficiently.

Table 4 Generating the β-Hermite and β-Laguerre ensembles efficiently

Ensemble	MATLAB commands

Hermite	`% Pick n, beta`
	`d = sqrt(chi2rnd(beta * [n:-1:1]))';`
	`H = spdiags(d, 1, n, n) + spdiags(randn(n, 1), 0, n, n);`
	`H = (H + H') / sqrt(2);`

Laguerre	`% Pick m, n, beta`
	`% Pick a > beta * (n - 1)/2`
	`d = sqrt(chi2rnd(2 * a - beta * [0:1:n-1]))';`
	`s = sqrt(chi2rnd(beta * [n:-1:1]))';`
	`B = spdiags(s, -1, n, n) + spdiags(d, 0, n, n);`
	`L = B * B';`

There are many extremely important practical steps we can take at this point. We outline two interesting practical points.

Sturm sequences can be used with T_n for the computation of histograms [1]. This is particularly valuable when there is interest in a relatively small number of histogram intervals (say 20 or 30) and n is very large. This is an interesting idea, particularly because most people think that histogramming eigenvalues first requires that they compute the eigenvalues, then sort them into bins. The Sturm sequence idea gives a count without computing the eigenvalues at all. This is a fine example of not computing more than is needed: if you only need a count, why should one compute the eigenvalues at all?

For the largest eigenvalue, the best trick for very large n is to only generate the upper left $10\,n^{1/3} \times 10\,n^{1/3}$ of the matrix. Because of what is known as the "Airy" decay in the corresponding eigenvector, the largest eigenvalue—which technically depends on every element in the tridiagonal matrix—numerically depends significantly only on the upper left part. This is a huge savings in a Monte Carlo generation. Further savings can be obtained by using the Lanczos "shift and invert" strategy given an estimate for the largest eigenvalue. We refer interested reads to Sect. 10 of [10]. Code 12 provides an example of how we succeed to compute the largest eigenvalue of a billion by billion matrix in the time required by naive methods for a hundred by hundred matrix.

3.4 Generalization Beyond Complex and Quaternion

There is little reason other than history and psychology to only consider the reals, complexes, and quaternions $\beta = 1, 2, 4$. The matrices given by T_n and B_n are well defined for any β, and are deeply related to generalizations of the Schur polynomials knows as the Jack Polynomials of parameter $\alpha = 2/\beta$. Much is known, but much remains to be known. Edelman [11] proposes in his method of "Ghosts and Shadows" that even G_β exists and has a meaning upon which algebra might be doable.

Code 12 Compute the largest eigenvalues of a billion by billion matrix

```
%% This code requires statistics toolbox
beta = 1; n = 1e9; opts.disp = 0; opts.issym = 1;
alpha = 10; k = round(alpha * n^(1/3)); % cutoff parameters
d = sqrt(chi2rnd(beta * n: -1: (n - k - 1)))';
H = spdiags(d, 1, k, k) + spdiags(randn(k, 1), 0, k, k);
H = (H + H')/sqrt(4 * n * beta); % Scale so largest eigenvlaue is
    near 1
eigs(H, 1, 1, opts);
```

Another interesting story comes from the fact that the reduced forms connect random matrices to the continuous limit, stochastic operators, which these authors believe represents a truer view of whys random matrices behave as they do [22].

4 Conclusion

In this paper, we give a brief summary of recent innovative applications in random matrix theory. We introduce the Hermite and Laguerre ensembles and give four famous laws (with MATLAB demonstration) that govern the limiting eigenvalue distributions of random matrices. Finally, we provide the details of matrix reductions that do not require a computer and give an overview of how these reductions can be used for efficient computation.

Acknowledgements The first author was supported in part by DMS 1035400 and DMS 1016125.

We acknowledge many of our colleagues and friends, too numerous to mention here, whose work has formed the basis of Random Matrix Theory. We particularly thank Raj Rao Nadakuditi for always bringing the latest applications to our attention.

References

1. J.T. Albrecht, C.P. Chan, and A. Edelman. Sturm sequences and random eigenvalue distributions. *Foundations of Computational Mathematics*, 9(4):461–483, 2009.
2. G.W. Anderson, A. Guionnet, and O. Zeitouni. *An Introduction to Random Matrices*. Cambridge Studies in Advanced Mathematics, vol. 118, 2010.
3. Z.D. Bai. Convergence rate of expected spectral distributions of large random matrices. Part II. Sample covariance matrices. *Annals of Probability*, 21:649–672, 1993.
4. Z.D. Bai. Convergence rate of expected spectral distributions of large random matrices. Part I. Wigner matrices. *The Annals of Probability*, 625–648, 1993.
5. Z. Bai and J. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices, 2nd edn.* Science Press, Beijing, 2010.
6. O.E. Barndorff-Nielsen and S. Thorbjørnsen. Lévy laws in free probability. *Proceedings of the National Academy of Sciences*, 99(26):16568, 2002.

7. R. Couillet and M. Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, Cambridge, 2011.
8. V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T.M. Allen, M. Altfeld, M. Carrington, D.J. Irvine, et al. Coordinate linkage of hiv evolution reveals regions of immunological vulnerability. *Proceedings of the National Academy of Sciences*, 108(28):11530, 2011.
9. I. Dumitriu. *Eigenvalue statistics for beta-ensembles*. Ph.D. thesis, Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, 2003.
10. A. Edelman and N.R. Rao. Random matrix theory. *Acta Numerica*, 14(233–297):139, 2005.
11. A. Edelman. The random matrix technique of ghosts and shadows. *Markov Processes and Related Fields*, 16(4):783–790, 2010.
12. P.J. Forrester. *Log-Gases and Random Matrices*, vol. 34. Princeton University Press, Princeton, 2010.
13. V.L. Girko. Circular law. *Teoriya Veroyatnostei I Ee Primeneniya*, 29:669–679, 1984.
14. V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1:457, 1967.
15. M.L. Mehta. *Random Matrices*, vol. 142. Academic Press, San Diego, 2004.
16. F. Mezzadri and N.C. Snaith. *Recent Perspectives in Random Matrix Theory and Number Theory*. Cambridge University Press, Cambridge, 2005.
17. R.J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1982.
18. A. Nica and R. Speicher. *Lectures on the Combinatorics of Free Probability*. London Mathematical Society Lecture Note Series, vol. 335. Cambridge University Press, New York, 2006.
19. D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica*, 17(4):1617, 2007.
20. S.M. Popoff, G. Lerosey, R. Carminati, M. Fink, A.C. Boccara, and S. Gigan. Measuring the transmission matrix in optics: an approach to the study and control of light propagation in disordered media. *Physical Review Letters*, 104(10):100601, 2010.
21. P. Šeba. Parking and the visual perception of space. *Journal of Statistical Mechanics: Theory and Experiment*, 2009:L10002, 2009.
22. B.D. Sutton. *The stochastic operator approach to random matrix theory*. Ph.D. thesis, Massachusetts Institute of Technology, 2005.
23. K.A. Takeuchi and M. Sano. Universal fluctuations of growing interfaces: evidence in turbulent liquid crystals. *Physical review letters*, 104(23):230601, 2010.
24. E.P. Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *Annals of Mathematics*, 62:548–564, 1955.

Boundary Closures for Sixth-Order Energy-Stable Weighted Essentially Non-Oscillatory Finite-Difference Schemes

Mark H. Carpenter, Travis C. Fisher, and Nail K. Yamaleev

Abstract A general strategy was presented in 2009 by Yamaleev and Carpenter (J. Comput. Phys. 228(11):4248–4272, 2009; J. Comput. Phys. 228(8):3025–3047, 2009), for constructing Energy Stable Weighted Essentially Non-Oscillatory (ESWENO) finite-difference schemes on *periodic* domains. Fisher et al. (J. Comput. Phys. 230(10):3727–3752, 2011) provided boundary closures for the fourth-order ESWENO scheme that maintain, the WENO stencil biasing properties and satisfy the summation-by-parts (SBP) operator convention, thereby ensuring stability in an L_2 norm. Herein, the general capability of finite-domain schemes is extended by providing closures for the sixth-order case. Third-order and fifth-order boundary closures are developed that are stable in diagonal and block norms, respectively, and achieve fourth- and sixth-order global accuracy for hyperbolic systems. A novel set of nonuniform flux interpolation points is necessary near the boundaries to simultaneously achieve (1) accuracy, (2) the SBP convention, and (3) WENO stencil biasing mechanics. Complete implementation details for the diagonal-norm sixth-order operator are provided as well as examples that demonstrate shock capturing and multi-domain capabilities.

1 Introduction

High-order *W*eighted *E*ssentially *N*on-*O*scillatory (WENO) methods are well suited for high fidelity simulations of complex physics that contain discontinuities. An example of a typical application is a canonical simulation of sound that is generated by a shock-vortex interaction [9, 28]. The high-order nature of WENO efficiently

M.H. Carpenter (✉) · T.C. Fisher
NASA Langley Research Center, Hampton, VA, 23681, USA
e-mail: mark.h.carpenter@nasa.gov

T.C. Fisher
e-mail: travis.fisher@nasa.gov

N.K. Yamaleev
Department of Mathematics, North Carolina A&T State University, Greensboro, NC, 27411, USA
e-mail: nkyamale@ncat.edu

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5_6,

resolves the subtle details of the sound generation and propagation, while the stencil biasing mechanics ensures robust, high resolution properties in the vicinity of the shock.

Conventional WENO formulations are well suited for structured, tensor-product meshes in multiple dimensions, efficiently implemented using one-dimensional operators in each of the three coordinate directions. They encounter serious challenges, however, at nodes that are near domain boundaries (or domain interfaces in a multi-domain implementation). To illustrate this, first recall that all high-order finite-difference formulations use "inward biased" stencils near boundaries that maintain the accuracy and stability of the interior scheme. Next, note that WENO schemes invoke stencil biasing mechanics throughout the domain, including nodes that adjoin the boundaries (e.g., sixth-order WENO schemes test four candidate stencils at each node). Thus, not only must the boundary closures that are used for WENO schemes be (1) inward biased, (2) stable, and (3) accurate, they also must be stable for any possible combination of candidate stencils that *occur anywhere in the domain*. Simultaneously satisfying these constraints is a remote possibility if left strictly to chance.

Many ad hoc boundary closures have been proposed for conventional WENO schemes. The simplest strategy is to discard stencils that extend outside the domain [8]. This strategy, however, reduces the local accuracy from $2p$ to $p + 1$. Alternatively, ghost points can be used in combination with physical boundary conditions to specify data in stencils that extend outside the domain [8]. This reduces the generality of the boundary treatment. In general, most implementations still rely on low-order boundary closures to achieve stability and robustness.

Fourth-order boundary closures for the "finite-domain" operator are described in a recent article by Fisher et al. [10]. These closures and near-wall biasing mechanics complement the *periodic domain* ESWENO finite-difference methodology reported in Refs. [26] and [27]. The new interior/boundary ESWENO schemes retain all of the salient features of the original periodic schemes. The finite-domain ESWENO schemes are constructed by adding a "special" nonlinear artificial dissipation term to a conventional WENO scheme (e.g. those found in Refs. [1, 13, 17]). The additional term is design-order accurate for smooth solutions including extrema, and is constructed such that the resulting ESWENO scheme is stable in the L_2-energy norm.

Recent experiences with the fourth-order ESWENO scheme indicate that its accuracy and computational efficiency are comparable to that achieved with the conventional fifth-order upwind-biased WENO schemes (e.g. [13]). Nevertheless, its resolving efficiency appears to be suboptimal relative to centered schemes of sixth-order accuracy. To this end, herein a complete description of the sixth-order class of ESWENO schemes is given. The design strategy for these schemes is equivalent to that used to build the fourth-order ESWENO schemes:

- Develop a sixth-order finite-domain target scheme that is stable, conservative, and accurate for smooth flows. This task is accomplished using the summation-by-parts (SBP) [15, 26] matrix operator framework.

- Recast the target scheme in the "dual grid" framework of the conventional WENO scheme, whereby the solution is stored and advanced at the grid points, while the interface fluxes that are constructed at the "half points" to ensure conservation. A special set of flux points and interpolants are required near the boundaries to accomplish this task.
- Develop a finite-domain WENO biasing strategy that allows all stencil weights to deviate from their target values. Precise control of the biasing mechanics ensures design-order accuracy for smooth solutions and essentially non-oscillatory properties at discontinuities. Any existing WENO stencil biasing strategy will suffice, e.g. WENO-Z [1].
- Test the stability of the finite-domain WENO scheme. If it is unstable, add a design-order artificial dissipation term that ensures an L_2-energy estimate for the combined operator.

The sixth-order ESWENO boundary closures are (1) conservative and L_2-energy stable for constant coefficient (linear) hyperbolic systems, (2) design-order accurate throughout the domain, including regions near the boundaries or near smooth extrema, and (3) have full WENO stencil biasing mechanics at all possible points. Furthermore, they exhibit enhanced robustness relative to conventional WENO operators. Indeed, results are presented for several test cases (involving gradients/discontinuities that cross physical boundaries) that demonstrate the enhanced stability of ESWENO schemes. Finally, as described next, the single-domain ESWENO operators fulfill an extremely important niche in the broader context of a generalized WENO formulation.

Stepping back to gain a "bird's-eye" perspective, all single-domain ESWENO operators (unlike conventional WENO operators) are elemental "building blocks" of a general multi-domain, complex geometry WENO methodology. The single-domain ESWENO operators generalize into multi-domain operators by coupling adjoining domains in a way that ensures conservation, accuracy and stability. (See Ref. [4] for coupling procedures used by multi-domain SBP operators.) Clearly, each subdomain retains the stencil biasing mechanics of the original single-domain ESWENO operators. Thus, simulations of complex geometries (e.g. chemically reacting internal flow through a scramjet combustor) can be performed using efficient and robust high-order ESWENO formulations. A multi-domain ESWENO capability does not provide the same geometric flexibility as a state-of-the-art finite element formulation [e.g. discontinuous Galerkin (DG)] because subdomains are generally larger than elements. Nevertheless, problems with geometries that are vastly more complex than canonical flow problems can be efficiently simulated, while achieving the state-of-the-art robustness of a structured grid WENO operator in each subdomain.[1]

The organization of the paper is as follows. Section 2 introduces relevant definitions, including a summary of finite-domain SBP operators, continuous, and

[1] Stencil biasing across block interfaces is still an area of active development. Maturity is approximately equivalent to that of state-of-the-are finite-element WENO formulations.

semidiscrete SBP stability proofs, and a summary of dual-grid finite-domain WENO matrix nomenclature. Section 3 then expresses the WENO target operator in SBP form, including a derivation that justifies the use of nonuniform flux points. Section 4 formulates the artificial dissipation term that guarantees the stability of the ESWENO operator, while Sect. 5 presents an analysis of the accuracy of the boundary closure terms. Section 6 presents a summary of numerical test cases. Conclusions and future directions are presented in Sect. 7. Detailed implementation mechanics are included in Appendices 8.1, 8.2 and 8.3.

2 Definitions

2.1 Summation-By-Parts Operators

Summation-by-parts matrix theory is used to prove the stability of the linear WENO target operator, as well as the general nonlinear ESWENO operator. Herein, the essential elements of the SBP matrix theory are presented.

2.1.1 Continuous Energy Estimate

Consider a linear, scalar wave equation

$$\begin{aligned} \frac{\partial v}{\partial t}+\frac{\partial f(v)}{\partial x}&=\frac{\partial^2 v}{\partial x^2}; \qquad f=av; \quad t\geq 0,\ -1\leq x\leq 1,\\ v(x,0)&=v_0(x), \quad t=0,\ -1\leq x\leq 1,\\ \gamma v(-1,t)-\varepsilon v_x(-1,t)&=g_{-1}(t)=L_{-1}(v), \quad t\geq 0,\\ \zeta v(+1,t)+\varepsilon v_x(+1,t)&=g_{+1}(t)=L_{+1}(v), \quad t\geq 0, \end{aligned} \tag{1}$$

where v is the continuous solution, a is a constant, $v_0(x)$ is a bounded piecewise continuous function in L_2, and ε satisfies $0<\varepsilon$. The boundary conditions that lead to a wellposed problem satisfy the following constraints [2, 6]

$$0\leq a+2\zeta; \qquad 0\leq -a+2\gamma. \tag{2}$$

Multiplying (1) by v and manipulating the resulting expression leads to the continuous energy equation

$$\frac{\partial}{\partial t}\|v\|_\Omega^2-2\varepsilon\|v_x\|_\Omega^2=\left[-av^2+2\varepsilon v v_x\right]\Big|_{x=-1}^{x=+1}. \tag{3}$$

Solving the boundary conditions in (1) for εv_x and substituting into (3) yields

$$\begin{aligned} \frac{\partial}{\partial t}\|v\|_\Omega^2-2\varepsilon\|v_x\|_\Omega^2=&+2v(+1,t)g_{+1}-(2\zeta+a)v(+1,t)^2\\ &+2v(-1,t)g_{-1}-(2\gamma-a)v(-1,t)^2. \end{aligned} \tag{4}$$

An energy estimate follows immediately by manipulating the boundary terms on the right-hand side of the equation (i.e. negative definite terms plus terms involving only boundary data).

2.1.2 Semidiscrete Energy Estimate

On the uniform solution point mesh $\mathbf{x}$ (i.e. $x_j = (j-1)/(N-1)$, $1 \le j \le N$), define the *discrete* solution $\mathbf{u}$, flux $\mathbf{f}$, and flux derivative $\mathbf{f}_x$ as

$$\mathbf{u} = \left[u_1(t), \ldots, u_N(t)\right]^T; \qquad \mathbf{f} = a\mathbf{u};$$
$$\mathbf{f}_x = \mathscr{D}_{SBP}\mathbf{f} = \left[(au)_x(x_1,t), \ldots, (au)_x(x_N,t)\right]^T$$

and the projections of the *continuous* solution v flux $f(v)$ and its derivative $\frac{\partial f(v)}{\partial x}$ onto $\mathbf{x}$ as

$$\mathbf{v} = \left[v(x_1,t), \ldots, v(x_N,t)\right]^T; \qquad \mathbf{F} = a\mathbf{v};$$
$$\mathbf{F}_x = \left[(av)_x(x_1,t), \ldots, (av)_x(x_N,t)\right]^T.$$

Define $\mathscr{D}_{SBP}$ to be a sixth-order approximation to the first-order derivative term in (1) that accounts for k points of accuracy z_{bc} at each boundary ($z_{bc} = 3$, or $z_{bc} = 5$) as

$$\mathbf{f}_x = \mathscr{D}_{SBP}\mathbf{f}; \qquad \mathbf{F}_x = \mathscr{D}_{SBP}\mathbf{F} + T_e \tag{5}$$

with the truncation error

$$|T_e| = \left[O\left(\delta x_1^{z_{bc}}\right), \ldots, O\left(\delta x_k^{z_{bc}}\right), O\left(\delta x^6\right), \ldots, O\left(\delta x^6\right), O\left(\delta x_{N+1-k}^{z_{bc}}\right), \ldots, O\left(\delta x_N^{z_{bc}}\right)\right]^T$$

with the grid spacing defined by $\delta x = 1/(N-1)$.

Place a mild restriction on the generality of the derivative operator (see Ref. [4] or Ref. [26]), and define the matrix $\mathscr{D}_{SBP}$ to be a SBP derivative operator of the form

$$\begin{aligned} &\mathscr{D}_{SBP} = \mathscr{P}^{-1}[\mathscr{Q} + \mathscr{R}]; \qquad \mathscr{Q} + \mathscr{Q}^T = \mathscr{B} = \mathrm{Diag}[-1, 0, \ldots, 0, 1]; \\ &\mathscr{R} = \mathscr{R}^T; \qquad \zeta^T \mathscr{R} \zeta \ge 0; \\ &\mathscr{P} = \mathscr{P}^T; \qquad \zeta^T \mathscr{P} \zeta > 0; \quad \zeta \ne \mathbf{0} \end{aligned} \tag{6}$$

with the matrix $\mathscr{R}$ defined by the expression

$$\begin{aligned} \mathscr{R} = {} & \Lambda_0 + \Delta\Lambda_1[\Delta]^T + \Delta[\Delta]^T\Lambda_2\Delta[\Delta]^T + \Delta[\Delta]^T\Delta\Lambda_3[\Delta]^T\Delta[\Delta]^T \\ & + \Delta[\Delta]^T\Delta[\Delta]^T\Lambda_4\Delta[\Delta]^T\Delta[\Delta]^T + \Delta[\Delta]^T\Delta[\Delta]^T\Delta\Lambda_5[\Delta]^T\Delta[\Delta]^T\Delta[\Delta]^T \end{aligned} \tag{7}$$

and Δ given by the nonsquare matrix

$$\Delta = \begin{pmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix}. \tag{8}$$

Furthermore, demand that the diagonal matrices Λ_i be positive semidefinite and of the appropriate size. That is,

$$\begin{aligned} &\Lambda_e = \mathrm{Diag}[\lambda_1, \ldots, \lambda_N]; \quad \lambda_j \geq 0,\ j = 1, N;\ e = 2p,\ p = 0, 1, 2, \\ &\Lambda_o = \mathrm{Diag}[\lambda_0, \ldots, \lambda_N]; \quad \lambda_j \geq 0,\ j = 0, N;\ o = 2p+1,\ p = 0, 1, 2. \end{aligned} \tag{9}$$

Note that the matrix operator $\mathscr{D}_{SBP}$ is nothing more than a conventional non-dissipative SBP operator $\mathscr{D}$ plus an as yet arbitrary dissipation matrix $\mathscr{R}$. These definitions for the SBP derivative operator lead to the following theorem.

Theorem 1 *Consider the semidiscrete approximation of* (1)

$$\begin{aligned} \mathbf{u}_t + \mathscr{D}_{SBP}\mathbf{f} &= \varepsilon \mathscr{D}\mathscr{D}\mathbf{u} \\ &\quad - \mathscr{P}^{-1}\{\sigma_{-1}[L^D_{-1}(\mathbf{u}) - g_{-1}]\mathbf{e}_{-1}\} \\ &\quad - \mathscr{P}^{-1}\{\sigma_{+1}[L^D_{+1}(\mathbf{u}) - g_{+1}]\mathbf{e}_{+1}\}, \quad t \geq 0,\ -1 \leq x \leq +1, \\ \mathbf{u}(x_i, 0) &= v_0(x_i), \quad t = 0,\ -1 \leq x \leq +1 \end{aligned} \tag{10}$$

with discrete boundary operators and penalty vectors defined by

$$\begin{aligned} &L^D_{-1}(\mathbf{u}) = (\gamma\mathbf{u} - \varepsilon\mathscr{D}\mathbf{u})|_{-1}; \qquad L^D_{+1}(\mathbf{u}) = (\zeta\mathbf{u} + \varepsilon\mathscr{D}\mathbf{u})|_{+1}; \\ &\mathbf{e}_{-1} = [1, 0, \ldots, 0, 0]^T; \qquad \mathbf{e}_{+1} = [0, 0, \ldots, 0, 1]^T. \end{aligned} \tag{11}$$

The semidiscrete operator defined by (10), *and* (11) *is* $(z_{bc}+1)$*-order accurate for sufficiently smooth* $v_0(x)$ *and is* L_2 *stable for suitably chosen Simultaneous Approximation Term* (*SAT*) [2] *penalty parameters* $\sigma_{\pm 1}$.

Proof Only a sketch of the proof is presented. Interested readers can find more details in reference [5]. Accuracy follows immediately from the accuracy constraint that is imposed in (5) and the theorem of Gustafsson [12]. The energy method [11] is used to prove stability. That is, substitute (6) through (9) into (10), multiply the result by the vector $\mathbf{u}^T\mathscr{P}$, add the result to its transpose, and rearrange the expression using (6) to obtain the semidiscrete energy

$$\begin{aligned} \frac{d}{dt}\|\mathbf{u}\|^2_{\mathscr{P}} - 2\varepsilon\|\mathscr{D}\mathbf{u}\|^2_{\mathscr{P}} &= +\big[-a\mathbf{u}^T\mathscr{B}\mathbf{u} + 2\varepsilon\mathbf{u}^T\mathscr{D}\mathbf{u}\big]\big|_{x=-1}^{x=+1} \\ &\quad - 2\sigma_{-1}\mathbf{u}_{-1}\big[(\gamma\mathbf{u} - \varepsilon\mathscr{D}\mathbf{u})|_{-1} - g_{-1}\big] \end{aligned}$$

$$
\begin{aligned}
&- 2\sigma_{+1}\mathbf{u}_{+1}\left[(\zeta\mathbf{u} + \varepsilon\mathscr{D}\mathbf{u})|_{+1} - g_{+1}\right] \\
&- 2a\mathbf{u}^T\left[\Lambda_0 + \Delta\Lambda_1[\Delta]^T + \Delta[\Delta]^T\Lambda_2\Delta[\Delta]^T\right. \\
&+ \Delta[\Delta]^T\Delta\Lambda_3[\Delta]^T\Delta[\Delta]^T \\
&+ \Delta[\Delta]^T\Delta[\Delta]^T\Lambda_4\Delta[\Delta]^T\Delta[\Delta]^T \\
&\left.+ \Delta[\Delta]^T\Delta[\Delta]^T\Delta\Lambda_5[\Delta]^T\Delta[\Delta]^T\Delta[\Delta]^T\right]\mathbf{u}. \quad (12)
\end{aligned}
$$

Extensive algebraic manipulations of the boundary terms in (12) result in conditions for the SAT penalty parameters $\sigma_{\pm 1}$ that lead to strong stability [5]. Note that the diagonal matrices Λ_i must be positive semidefinite to ensure stability of the formulation. □

Remark The differentiation matrix $\mathscr{D}_{SBP}$ that is used in (10) is only constrained by accuracy and structure conditions. Note that (6) could be nonlinear (i.e., $\mathscr{Q} = \mathscr{Q}(\mathbf{u})$ and $\Lambda_i = \Lambda_i(\mathbf{u})$) despite the linearity of (1). The broad generality of $\mathscr{D}_{SBP}$ enables L_2-stable SBP schemes (i.e., ESWENO) to be constructed with all of the essential nonlinear biasing properties of WENO schemes.

Remark The previous single-domain result establishes the conditions required to achieve semi-discrete strong-stability, subject to consistent, wellposed initial and boundary conditions. This result naturally extends to (1) systems of equations via characteristic rotations (see Sect. 8.2), (2) multiple dimensions via a tensor product formulation [21, 22], (3) viscous equations [18], and (4) problems with geometric complexity that require a multiple domain formulation [6].

2.1.3 Complementary Grids

The implementation of the finite-domain WENO (and ESWENO) stencil biasing mechanics, as a matter of convenience, uses two complementary sets of grid points (or meshes) that are defined in each subdomain. For convenience of presentation we assume each can be represented on the nondimensional interval $0 \le x \le 1$. The relationship between the two meshes is illustrated in Fig. 1. The discrete sets differ in dimension by one and are expressed by using the vectors

$$\mathbf{x} = [x_1, x_2, \ldots, x_{N-1}, x_N]^T; \qquad \bar{\mathbf{x}} = [\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_{N-1}, \bar{x}_N]^T$$

with $x_1 = \bar{x}_0 = 0$ and $x_N = \bar{x}_N = 1$. The low-dimensional vector $\mathbf{x}$ contains a uniformly distributed set of points $x_j = (j-1)\delta x$, $j = 1, \ldots, N$, with $\delta x = 1/(N-1)$. Because the discrete solution u_j is stored and advanced in time at the points x_j, these points are referred to as "solution points." Surrounding each solution point x_j is a control volume. A second set of points $\bar{\mathbf{x}}$ is situated at the interfaces between adjoining control volumes, making the width of each control volume $(\Delta\bar{\mathbf{x}})_j = \bar{x}_j - \bar{x}_{j-1}$. Because the solution $\mathbf{u}$ and the flux $\mathbf{f}$ data that are interpo-

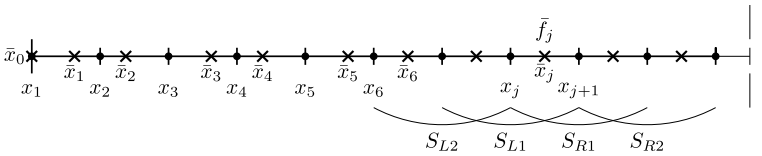


Fig. 1 Half computational domain which shows the relationship between the ESWENO solution and flux points. Also shown at flux point $\bar{f}_j$ are stencil dependencies of four smoothness indicators S_k, $k = [L2, L1, R1, R2]$. (Interpolation operators and smoothness indicators have identical dependencies)

lated from $\mathbf{x}$ are used to construct interface fluxes $\bar{\mathbf{f}}$ at $\bar{\mathbf{x}}$, points in this second set are referred to as the "flux points." (The overbar nomenclature is reserved herein for those quantities that are defined at the flux points $\bar{\mathbf{x}}$.) Note that the flux points $\bar{x}_j$ are uniformly distributed in the interior, but are nonuniform near each boundary.

2.1.4 Discrete Differentiation

On each subdomain, define a general class of finite-domain WENO schemes by using the following sequence of matrix operations:

$$\mathscr{D}_{weno} = \mathscr{P}^{-1} \Delta \sum_r \bar{w}^{(r)}{}_M \mathscr{I}_N^{(r)}. \tag{13}$$

Furthermore, define the action of $\mathscr{D}_{weno}$ on the flux $\mathbf{f}$ by using the following nomenclature:

$$\mathscr{D}_{weno}\mathbf{f} = \mathscr{P}^{-1}\Delta\bar{\mathbf{f}} = \mathscr{P}^{-1}\Delta \sum_r \bar{w}^{(r)}\bar{\mathbf{f}}^{(r)} = \mathscr{P}^{-1}\Delta \sum_r \bar{w}^{(r)}{}_M \mathscr{I}_N^{(r)}\mathbf{f}.$$

The discrete differentiation operator $\mathscr{D}_{weno}$ reflects the three distinct steps that are used by the WENO stencil biasing mechanics. Each of the three steps is now summarized.

First, data are interpolated from $\mathbf{x}$ to the flux points $\bar{\mathbf{x}}$ by using (r) nonsquare, second-order, interpolation operators ${}_M\mathscr{I}_N^{(r)}$

$$\bar{\mathbf{f}}^{(r)} = {}_M\mathscr{I}_N^{(r)}\mathbf{f}. \tag{14}$$

The dimensions of the matrix interpolants ${}_M\mathscr{I}_N^{(r)}$ reflect the dimensional discrepancy between $\mathbf{x}$ and $\bar{\mathbf{x}}$. All the candidate fluxes $\bar{\mathbf{f}}^{(r)}$ in the interior of the domain are of accuracy three, rather than the full sixth order.

Next, the nonlinear stencil biasing weights $\bar{w}^{(r)}$ are formed by using the expressions

$$\begin{aligned} &\bar{w}_j^{(r)} = \frac{\bar{\alpha}_j^{(r)}}{\sum_r \bar{\alpha}_j^{(r)}}; \qquad \bar{\alpha}_j^{(r)} = \bar{d}_j^{(r)}\left(1 + \frac{\bar{\tau}_j}{\varepsilon + \bar{\beta}_j^{(r)}}\right), \quad \forall r,\ 1 \le j \le N-1; \\ &\bar{w}^{(r)} = \mathrm{Diag}\big[\bar{w}_0^{(r)}, \ldots, \bar{w}_N^{(r)}\big], \quad \forall r \end{aligned} \tag{15}$$

where $\bar{d}_j^{(r)}$ is the target weight of the candidate stencil and $\bar{\beta}_j^{(r)}$ and $\bar{\tau}_j$ are the smoothness indicators. A single flux $\bar{\mathbf{f}}$ is then formed as a convex combination of the weights from the candidate fluxes as

$$\bar{\mathbf{f}} = \sum_r \bar{w}^{(r)} \bar{\mathbf{f}}^{(r)}. \tag{16}$$

Finally, the combined flux $\bar{\mathbf{f}}$ is differentiated from $\bar{\mathbf{x}}$ back onto the solution points $\mathbf{x}$ by using a telescoping, nonsquare difference operator

$$\mathbf{f}_x = \mathscr{P}^{-1} \Delta \bar{\mathbf{f}}. \tag{17}$$

The matrix $\mathscr{P}$ accounts for variable grid spacing δx_j throughout the domain (see Fig. 1), while the operator Δ ensures discrete conservation. The dimensions of Δ and ${}_M\mathscr{I}_N^C$ are $[N \times (N+1)]$ and $[(N+1) \times N]$, respectively, while their product $\Delta_M \mathscr{I}_N^C$ is square $[N \times N]$, as is the matrix $\mathscr{P}$.

2.1.5 Target Weights

The target weights $\bar{d}^{(r)}$ are derived by expressing a desirable *linear* scheme in terms of the two-grid WENO framework. Any consistent operator (in principle) could be used as the target for a WENO scheme. The target scheme provides an "anchor" for the nonlinear scheme, that is, the nonlinear weights $\bar{w}^{(r)}$ are designed to be small perturbations away from the target weights $\bar{d}^{(r)}$.

Essentially no stencil biasing occurs when the solution is well resolved (i.e., the limit $\delta x \to 0$ for smooth $\mathbf{u}$), and the weight functions $\bar{w}^{(r)}$ used in (15) reduce to

$$\bar{w}_j^{(r)} = \frac{\bar{d}_j^{(r)}}{\sum_r \bar{d}_j^{(r)}} = \bar{d}_j^{(r)}. \tag{18}$$

Equation (13) reduces in this smooth limit to the *target* operator

$$\mathscr{D}_{tarW} = \mathscr{P}^{-1} \Delta \sum_r \bar{d}^{(r)}\, {}_M\mathscr{I}_N^{(r)}. \tag{19}$$

Although both $\bar{w}^{(r)}$ and $\bar{d}^{(r)}$ are consistent and of design order, only the weights $\bar{w}^{(r)}$ minimize oscillations.

2.1.6 Sixth-Order WENO

Figure 1 is a schematic of a sixth-order central operator. For brevity the right computation boundary is not included. In the domain interior, four third-order fluxes $\bar{\mathbf{f}}^{(r)} = {}_M\mathscr{I}_N^{(r)}\mathbf{f}$, $r = \{L2, L1, R1, R2\}$, are constructed at the flux points $\bar{\mathbf{x}}$ by using four distinct stencils S^{L2}, S^{L1}, S^{R1} and S^{R2}.[2] Each flux is constructed from data that are obtained by using three points: $S^{L2} = \{x_{j-2}, x_{j-1}, x_{j-0}\}$, $S^{L1} = \{x_{j-1}, x_{j-0}, x_{j+1}\}$, $S^{R1} = \{x_{j-0}, x_{j+1}, x_{j+2}\}$ and $S^{R2} = \{x_{j+1}, x_{j+2}, x_{j+3}\}$. The interpolants $S^{L4} = \{x_{j-4}, x_{j-3}, x_{j-2}\}$, $S^{L3} = \{x_{j-3}, x_{j-2}, x_{j-1}\}$, $S^{R3} = \{x_{j+2}, x_{j+3}, x_{j+4}\}$, $S^{R4} = \{x_{j+3}, x_{j+4}, x_{j+5}\}$ are included near boundaries to increase the fidelity of the boundary closures.

The combined flux $\bar{f}_j$ is given by the expression

$$\bar{f}_j = +\bar{w}_j^{L2}\bar{f}_j^{L2} + \bar{w}_j^{L1}\bar{f}_j^{L1} + \bar{w}_j^{R1}\bar{f}_j^{R1} + \bar{w}_j^{R2}\bar{f}_j^{R2}, \quad 3 \le j \le N-3. \tag{20}$$

The expressions for $\bar{f}_j$ near the boundaries are more complicated but conform to the same matrix conventions. Complete definitions of the parameters $\bar{\tau}_j$, ε, $\bar{\beta}_j^{(r)}$, $r = L2, L1, R1, R2$, including the auxiliary formulae at the points near the boundaries (i.e. $r = L4, L3, R3, R4$) as well as implementation details, are included in Appendices 8.1 and 8.2.

Remark Note the relative proximity of the solution and flux points (i.e. x_j and $\bar{x}_j$) in the interior and near the boundaries. The flux points $\bar{x}_j$ are evenly spaced in the interior, but not near the boundary.

3 Expressing Summation-By-Parts Operators in WENO Form

3.1 Target Operator

The WENO target operators that are developed herein use non-dissipative, seven-point, sixth-order, central discretizations, which are expressed in the SBP matrix convention as

$$\mathscr{D}_{tarW} = \mathscr{P}^{-1}\mathscr{Q}. \tag{21}$$

Note that any dissipation that enters the WENO scheme comes solely from the nonlinear stencil biasing mechanics and from the explicit stabilization terms that are added to guarantee energy stability (i.e., $\mathscr{R} = 0$ in the general expression $\mathscr{D} = \mathscr{P}^{-1}(\mathscr{Q} + \mathscr{R})$).

[2] The $r = \{L2, L1, R1, R2\}$ nomenclature identifies the origin of the data relative to the interface position.

The matrices $\mathscr{P}$ and $\mathscr{Q}$ are composed of the conventional banded operators in the interior of the domain as

$$P_I = \delta x\, I; \qquad Q_I = \text{Heptadiagonal}\left[\frac{-1}{60}, \frac{+9}{60}, \frac{-45}{60}, 0, \frac{+45}{60}, \frac{-9}{60}, \frac{+1}{60}\right].$$

Sixth-order SBP boundary closures require auxiliary stencils at $N_b \geq 6$ adjoining boundary points [24] to satisfy the necessary accuracy, symmetry, and skew-symmetry constraints that are given by (6). Herein, $N_b = 6$ is assumed, whereby the matrices $\mathscr{P}$ and $\mathscr{Q}$ take the form

$$\mathscr{P} = \delta x \begin{pmatrix} P_0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & (P_0)^{PT} \end{pmatrix}; \qquad \mathscr{Q} = \begin{pmatrix} Q_0 & Q_d & 0 \\ -Q_d^T & Q_I & Q_d \\ 0 & -Q_d^T & -(Q_0)^{PT} \end{pmatrix} \tag{22}$$

with

$$P_0 = \begin{pmatrix} p_{11} & \cdots & p_{16} \\ \vdots & \ddots & \vdots \\ p_{16} & \cdots & p_{66} \end{pmatrix}; \qquad Q_0 = \begin{pmatrix} \frac{-1}{2} & q_{12} & \cdots & q_{16} \\ -q_{12} & 0 & \ddots & \vdots \\ \vdots & \ddots & & q_{56} \\ -q_{16} & \cdots & -q_{56} & 0 \end{pmatrix};$$

$$Q_d = \begin{pmatrix} 0 & \cdots & & \\ 0 & \cdots & & \\ 0 & \cdots & & \\ \frac{+1}{60} & 0 & \cdots & \\ \frac{-9}{60} & \frac{+1}{60} & 0 & \cdots \\ \frac{45}{60} & \frac{-9}{60} & \frac{+1}{60} & 0 \quad \cdots \end{pmatrix} \tag{23}$$

where PT denotes the per-symmetric transpose $a_{[i,j]} = a_{[N-(i-1),N-(j-1)]}$.

Two types of $\mathscr{P}$ norms are used to prove stability with SBP discretizations: diagonal, and block. The diagonal norm is frequently warranted because stability proofs extend to variable-coefficient problems [20], as well as to certain nonlinear formulations [14, 19]. These proofs are not shared by the block-norm boundary closures. Diagonal-norm boundary closures (unlike block-norm closures), however, cannot maintain the full design-order potential of the $(2p)$-order interior operator because the necessary accuracy/structure constraints in (6) limit the accuracy of boundary stencils at order (p). Thus, the global accuracy of diagonal-norm closures is only $(p+1)$-order despite $(2p)$-order interior stencil accuracy [12, 25].

3.2 Equating the Target WENO and SBP Discretization Matrices

Equating the target WENO and SBP matrix forms that are given in (19) and (21), respectively, yields the matrix equation

$$\mathscr{D}_{tarW} = \mathscr{P}^{-1}\Delta \sum_r \bar{d}^{(r)}\, {}_M\mathscr{I}_N^{(r)} = \mathscr{P}^{-1}\mathscr{Q}. \tag{24}$$

This equation provides the necessary conditions that the flux points $\bar{\mathbf{x}}$, the target weights $\bar{d}^r$, and the interpolation operators ${}_M\mathscr{I}_N^r$ must satisfy to express the SBP target operator $\mathscr{P}^{-1}\mathscr{Q}$ in WENO flux form.

Equation (24) is a nonlinear function of the flux points $\bar{\mathbf{x}}$. Indeed, if we assume uniform solution points $\mathbf{x}$, then the functional dependency of the target coefficients and interpolants is given by

$$\begin{aligned} &\left[\bar{\mathbf{d}}^{(r)} = \bar{\mathbf{d}}^{(r)}(\bar{\mathbf{x}});\ {}_M\mathscr{I}_N^{(r)} = {}_M\mathscr{I}_N^{(r)}(\bar{\mathbf{x}})\right], \quad r = L4, \ldots, R4; \\ &\mathscr{P} = \mathscr{P}(\vartheta_l), \qquad \mathscr{Q} = \mathscr{Q}(\vartheta_l), \quad l = 1, 2. \end{aligned}$$

The existence of a solution to (24) is not guaranteed when an arbitrary solution ($\mathbf{x}$) and flux ($\bar{\mathbf{x}}$) points are used. A solution does exist for a uniform solution and flux points and periodic operators (see reference [27] for details). The task at hand is to determine the points $\mathbf{x}$, $\bar{\mathbf{x}}$ and boundary operators ${}_M\mathscr{I}_N^r$, $\bar{d}^r$ that admit solutions to (24).

3.3 Nonuniform Flux Points

An analytic solution to (24) is not forthcoming for uniformly spaced solution ($\mathbf{x}$) and flux ($\bar{\mathbf{x}}$) points on the domain $0 \le x \le 1$:

$$\begin{aligned} &x_{j+1} = (j-1)\delta x, \quad j = 1, N,\ \delta x = 1/(N-1); \\ &\bar{x}_k = \tfrac{(x_k + x_{k+1})}{2}, \quad k = 1, N-1,\ x_0 = 0,\ \bar{x}_N = 1, \end{aligned}$$

for either the diagonal-norm or the block-norm SBP operators $\mathscr{P}^{-1}\mathscr{Q}$.

Multiple solutions exist for (24), for either the diagonal or block norms, if the second, through sixth flux points $\bar{\mathbf{x}}$ are shifted to a nonuniform set of points near the boundaries. (Figure 1 illustrates the relationship between the solution points $\mathbf{x}$ and the flux points $\bar{\mathbf{x}}$, including the shifted flux points near each boundary.) The location of the flux points is precisely related to the $\mathscr{P}$ norm that is used to prove stability for the SBP operator. We state without proof the following lemma which is an essential ingredient for proving stability of the ESWENO operators:

Lemma 1 *Assume N uniformly distributed solution points $\mathbf{x}$. An $N+1$ dimensional set of flux points $\bar{\mathbf{x}}$ exists for which the central sixth-order WENO derivative*

operators, can be expressed in SBP form. The discrete set of points $\bar{\mathbf{x}}$ *satisfies the following expression*:

$$\mathscr{P}^{-1}\Delta\bar{\mathbf{x}} = \mathbf{1} \tag{25}$$

with $\mathbf{1}$ *as the unit vector.*

A proof can be constructed for each individual case in question.

Remark The solution ($\mathbf{x}$) and flux ($\bar{\mathbf{x}}$) points are interdigitated, the exceptions being the collocated end points.

Given this set of flux points $\bar{\mathbf{x}}$, the interpolation stencils $S^{(r)}$, $r = \{L2, L1, R1, R2\}$, are consistently defined for all points $\bar{x}_j$, $j = 3, N-3$. The flux points next to each boundary $\bar{x}_j$, $j = 1, 2$ and $j = N-1, N-2$, do not have a complete set of interpolants $S^{(r)}$. To guarantee linear stability, additional "inward-biased" stencils are defined at points near boundaries. Specifically, the interpolation operators I^{L4}, I^{L3}, I^{R3}, I^{R4} and weights $\bar{w}^{L4}$, $\bar{w}^{L3}$, $\bar{w}^{R3}$, $\bar{w}^{R4}$ are added in (20) and (24), but only at the boundaries.

Two general solutions of (24) are determined. One solution is derived using a diagonal norm ($\text{WENO}_{3\text{-}6\text{-}3}$); the other, a block norm ($\text{WENO}_{5\text{-}6\text{-}5}$). All necessary operators and implementation details for the WENO/$\text{ESWENO}_{3\text{-}6\text{-}3}$ schemes are given in Appendices 8.1 through 8.3. Owing to the complexity of the numerical coefficients, the implementation operators for the WENO/$\text{ESWENO}_{5\text{-}6\text{-}5}$ schemes are not included herein. They can, however, be obtained from the authors.

4 ESWENO Energy Stabilization Terms

Equating the general SBP and nonlinear WENO matrix forms that are given in (6) and (13), respectively yields the matrix equation

$$\mathscr{D}_{weno} = \mathscr{P}^{-1}\Delta\sum_r \bar{w}^{(r)}{}_M\mathscr{I}_N^{(r)} = \mathscr{P}^{-1}(\mathscr{Q}_{weno} + \mathscr{R}_{weno}). \tag{26}$$

The baseline WENO scheme $\mathscr{D}_{weno}$ is not guaranteed to be stable for arbitrary combinations of the nonlinear weights $\bar{w}^{(r)}$. Stability can be achieved, however, by combining $\mathscr{D}_{weno}$ with an appropriate artificial dissipation term $\mathscr{P}^{-1}\mathscr{R}_{es}$ as

$$\mathscr{D}_{esweno} = \mathscr{D}_{weno} + \mathscr{P}^{-1}\mathscr{R}_{es}. \tag{27}$$

The new scheme $\mathscr{D}_{esweno}$ that is given by (27) is designated as the Energy Stable WENO (ESWENO) scheme because the dissipation term $\mathscr{P}^{-1}\mathscr{R}_{es}$ is designed such that $\mathscr{D}_{esweno}$ satisfies an L_2 energy estimate. The design of $\mathscr{P}^{-1}\mathscr{R}_{es}$ is as follows.

First, form the symmetric/skew-symmetric decomposition of the nonlinear WENO scheme as

$$\begin{aligned}\mathscr{D}_{weno} &= \mathscr{P}^{-1}(\mathscr{Q}_{weno} + \mathscr{R}_{weno});\\ \mathscr{Q}_{weno} &= \frac{1}{2}\big[(\mathscr{P}\mathscr{D}_{weno}) - (\mathscr{P}\mathscr{D}_{weno})^T + \mathscr{B}\big];\\ \mathscr{R}_{weno} &= \frac{1}{2}\big[(\mathscr{P}\mathscr{D}_{weno}) + (\mathscr{P}\mathscr{D}_{weno})^T - \mathscr{B}\big]\end{aligned} \tag{28}$$

and substitute (28) into (27) to obtain the expression

$$\mathscr{D}_{esweno} = \mathscr{P}^{-1}(\mathscr{Q}_{weno} + \mathscr{R}_{weno} + \mathscr{R}_{es}).$$

An L_2 energy estimate is obtained if $\mathscr{D}_{esweno}$ conforms to the general structural requirements of an SBP operator $\mathscr{P}^{-1}(\mathscr{Q}+\mathscr{R})$ that is given in (6) through (9).

Next, expand the symmetric matrix $\mathscr{R}_{weno}$ into its elemental components by using the decomposition[3]

$$\begin{aligned}\mathscr{P}^{-1}\mathscr{R}_{weno} = \mathscr{P}^{-1}\big[&\Lambda_0 + \Delta\Lambda_1[\Delta]^T + \Delta[\Delta]^T\Lambda_2\Delta[\Delta]^T + \Delta[\Delta]^T\Delta\Lambda_3[\Delta]^T\Delta[\Delta]^T\\ &+ \Delta[\Delta]^T\Delta[\Delta]^T\Lambda_4\Delta[\Delta]^T\Delta[\Delta]^T\\ &+ \Delta[\Delta]^T\Delta[\Delta]^T\Delta\Lambda_5[\Delta]^T\Delta[\Delta]^T\Delta[\Delta]^T\big]\end{aligned} \tag{29}$$

and assume the energy stabilization term to be of the same form as

$$\begin{aligned}\mathscr{P}^{-1}\mathscr{R}_{es} = \mathscr{P}^{-1}\big[&\hat{\Lambda}_0 + \Delta\hat{\Lambda}_1[\Delta]^T + \Delta[\Delta]^T\hat{\Lambda}_2\Delta[\Delta]^T + \Delta[\Delta]^T\Delta\hat{\Lambda}_3[\Delta]^T\Delta[\Delta]^T\\ &+ \Delta[\Delta]^T\Delta[\Delta]^T\hat{\Lambda}_4\Delta[\Delta]^T\Delta[\Delta]^T\\ &+ \Delta[\Delta]^T\Delta[\Delta]^T\Delta\hat{\Lambda}_5[\Delta]^T\Delta[\Delta]^T\Delta[\Delta]^T\big].\end{aligned} \tag{30}$$

The existence of these decompositions is established in Ref. [27]. Furthermore, the diagonal matrix $\hat{\Lambda}_0$ is identically zero for both the diagonal and block-norm schemes. Thus, (30) reduces to

$$\begin{aligned}\mathscr{P}^{-1}\mathscr{R}_{es} = \mathscr{P}^{-1}\big[&\Delta\hat{\Lambda}_1[\Delta]^T + \Delta[\Delta]^T\hat{\Lambda}_2\Delta[\Delta]^T + \Delta[\Delta]^T\Delta\hat{\Lambda}_3[\Delta]^T\Delta[\Delta]^T\\ &+ \Delta[\Delta]^T\Delta[\Delta]^T\hat{\Lambda}_4\Delta[\Delta]^T\Delta[\Delta]^T\\ &+ \Delta[\Delta]^T\Delta[\Delta]^T\Delta\hat{\Lambda}_5[\Delta]^T\Delta[\Delta]^T\Delta[\Delta]^T\big].\end{aligned} \tag{31}$$

[3]Note that the dimensions of the even indexed Λ_i are $(N \times N)$, while the odd-indexed Λ_i have dimensions $(N+1 \times N+1)$. This is the result of the difference matrix Δ having dimensions $[N \times (N+1)]$.

In general, $\mathscr{R}_{weno}$ will not be positive semidefinite because of its nonlinear dependence on $\bar{w}^{(r)}$. Thus, the individual components $[\lambda_j]_i$ of Λ_i in (29) may be negative.

Constructing the individual components $[\hat{\lambda}_j]_i$ in $\hat{\Lambda}_i$ as

$$[\hat{\lambda}_j]_i = \frac{1}{2}\Big(\sqrt{\big([\lambda_j]_i\big)^2 + \delta_i^2} - [\lambda_j]_i\Big), \quad i = 0, 5,\ j = 0, N,\ 0 < \delta_i \le O\big(\delta x^5\big) \tag{32}$$

smoothly enforces positivity

$$[\lambda_j]_i + [\hat{\lambda}_j]_i \ge 0, \quad \forall i, j$$

and guarantees that any nonzero vector ζ satisfies $\zeta^T(\mathscr{R}_{weno} + \mathscr{R}_{es})\zeta \ge 0$. (A constraint on δ_i is needed to achieve design accuracy and is derived in the following section.)

Implementation is facilitated by defining the vector

$$\begin{aligned}\bar{\boldsymbol{\psi}} = \big[&\hat{\Lambda}_1[\Delta]^T + [\Delta]^T\hat{\Lambda}_2\Delta[\Delta]^T + [\Delta]^T\Delta\hat{\Lambda}_3[\Delta]^T\Delta[\Delta]^T \\ &+ [\Delta]^T\Delta[\Delta]^T\hat{\Lambda}_4\Delta[\Delta]^T\Delta[\Delta]^T \\ &+ [\Delta]^T\Delta[\Delta]^T\Delta\hat{\Lambda}_5[\Delta]^T\Delta[\Delta]^T\Delta[\Delta]^T\big]\mathbf{f}.\end{aligned} \tag{33}$$

This definition of $\bar{\boldsymbol{\psi}}$ allows the energy stabilization term to be combined with the original flux $\bar{\mathbf{f}}$ to yield the conservative form:

$$\mathbf{f}_x = \mathscr{D}_{esweno}\mathbf{f} = \mathscr{P}^{-1}\Delta(\bar{\mathbf{f}}_{weno} + \bar{\boldsymbol{\psi}}). \tag{34}$$

5 Consistency Analysis

The consistency of periodic ESWENO schemes is derived in Refs. [26] and [27] and extends without modification to the interior portion of the finite-domain ESWENO schemes. Three additional clarifications are needed to ensure that the finite-domain ESWENO schemes retain the formal accuracy of the target SBP operator. First, sufficient conditions that govern the maximum allowable deviation of WENO weights from their target values $\bar{w}_j^{(r)} - \bar{d}_j^{(r)}$ are needed at boundary stencils. Second, the boundary stencil biasing mechanics must not violate these sufficient conditions at smooth extrema near the boundaries. Third, the energy stabilization terms that are added at unstable boundary stencils must not violate the sufficient conditions.

5.1 Necessary and Sufficient Conditions for Consistency

Define the implicit auxiliary functions $h_j(x)$ on each control volume as

$$f_j(x) = \frac{1}{\delta x_j}\int_{x-\frac{\delta x_j}{2}}^{x+\frac{\delta x_j}{2}} h_j(\eta)d\eta;$$
$$\frac{df_j(x)}{dx} = \frac{1}{\delta x_j}\left[h_j\left(x+\frac{\delta x_j}{2}\right) - h_j\left(x-\frac{\delta x_j}{2}\right)\right]. \tag{35}$$

See reference [23] for a detailed discussion on the definition of the implicit function $h_j(x)$. The elemental fluxes $\bar{\mathbf{f}}^{(r)} = {}_M\mathscr{I}_N^{(r)}\mathbf{f}$, $r = \{L4, L3, L2, L1, R1, R2, R3, R4\}$ are constructed from local interpolation operators throughout the domain, and are third-order approximations to the implicit flux functions $h_j(x)$ in the domain interior ($7 \le j \le N-6$).

Unevenly spaced flux points, $\bar{\mathbf{x}}$, near each boundary produce solution points that are not centered within each interval $\Delta\bar{\mathbf{x}}_j$. This nonuniformity violates the underlying assumptions used to construct highly accurate $h_j(x)$ functions from pointwise data on $\mathbf{x}$. Thus, a different paradigm is needed to determine accurate flux interpolants near boundaries.

All interpolants are constructed from data residing at three solution points $\mathbf{x}$, and satisfy the following accuracy constraints throughout the domain:

$$\mathbf{f}_x = \mathscr{P}\Delta_M\mathscr{I}_N^{(r)}\mathbf{f} + O\left[\delta x_1^2, \ldots, \delta x_7^2, \delta x_8^3, \ldots, \delta x_{N-7}^3, \delta x_{N-6}^2, \ldots, \delta x_N^2\right]^T, \quad \forall r. \tag{36}$$

Note that the interior interpolants satisfy both (36) as well as the implicit relation $\bar{\mathbf{f}}^{(r)} = h(\eta) + O(\delta x^3)$. The near-boundary interpolants satisfy at worst $\bar{\mathbf{f}}^{(r)} = h(\eta) + O(\delta x^1)$, and possibly the stronger condition $\bar{\mathbf{f}}^{(r)} = h(\eta) + O(\delta x^2)$, although no proof for this conjecture is available.

Expand the elemental fluxes as

$${}_M\mathscr{I}_N^{(r)}\mathbf{f} = \bar{\mathbf{f}}^{(r)} = \bar{\mathbf{h}}^{(r)}(\bar{\mathbf{x}}) + \sum_{l=s_l}^{s_h}(\Delta\bar{\mathbf{x}})^l\bar{\mathbf{c}}_l^{(r)} + O\left(\Delta\bar{\mathbf{x}}^{s_h+1}\right), \quad \forall r \tag{37}$$

with vectors of constants $\bar{\mathbf{c}}_l^{(r)}$ used in (37) not depending on $\Delta\bar{\mathbf{x}}$. Note that the general indices s_l and s_h allow one derivation to account for spatial variation in the accuracy of the fluxes $\bar{\mathbf{f}}_j^{(r)}$. For example, $s_l = 3$, $s_h = 6$ for the interior points $7 \le j \le N-6$ while $s_l = 1$, $s_h = 3$ for the near-boundary points of the diagonal norm WENO/ESWENO$_{3\text{-}6\text{-}3}$ scheme, and $s_l = 1$, $s_h = 5$ for the near-boundary points of the block norm WENO/ESWENO$_{5\text{-}6\text{-}5}$ scheme.

Subtracting (19) from (13) and multiplying by the flux $\mathbf{f}$ yields the expression

$$\mathscr{D}_{weno}\mathbf{f} - \mathscr{D}_{tarW}\mathbf{f} = \mathscr{P}^{-1}\Delta\sum_r\left(\bar{w}^{(r)} - \bar{d}^{(r)}\right){}_M\mathscr{I}_N^{(r)}\mathbf{f}. \tag{38}$$

Substituting (37) into (38) yields

$$\begin{aligned}[\mathscr{D}_{weno}\mathbf{f}] - [\mathscr{D}_{tarW}\mathbf{f}] &= \mathscr{P}^{-1}\Delta\sum_r\big(\bar{w}^{(r)} - \bar{d}^{(r)}\big)\bar{\mathbf{f}}^{(r)} \\ &= \mathscr{P}^{-1}\Delta\sum_r\big(\bar{w}^{(r)} - \bar{d}^{(r)}\big)\bar{\mathbf{h}}^{(r)}(\bar{\mathbf{x}}) \\ &\quad + \mathscr{P}^{-1}\Delta\sum_{l=s_l}^{s_h}\sum_r\big(\bar{w}^{(r)} - \bar{d}^{(r)}\big)(\Delta x)^l\bar{\mathbf{c}}_l^{(r)} \\ &\quad + O\big(\Delta\bar{\mathbf{x}}^{(s_h+1)}\big). \end{aligned} \tag{39}$$

Recognizing that $\mathscr{P} = O(\Delta\bar{\mathbf{x}})$, (39) shows that 6th-order accuracy is achieved if the weights of the WENO operator $\mathscr{D}_{weno}$ satisfy the following necessary and sufficient conditions:

$$\begin{aligned} \Delta\sum_r\big(\bar{w}^{(r)} - \bar{d}^{(r)}\big) &= O\big(\Delta\bar{\mathbf{x}}^{s_h+1}\big), \\ \Delta\sum_r\big(\bar{w}^{(r)} - \bar{d}^{(r)}\big)\bar{\mathbf{c}}_{s_l}^{(r)} &= O\big(\Delta\bar{\mathbf{x}}^{(s_h+1-s_l)}\big), \\ &\vdots \\ \Delta\sum_r\big(\bar{w}^{(r)} - \bar{d}^{(r)}\big)\bar{\mathbf{c}}_{s_h}^{(r)} &= O\big(\Delta\bar{\mathbf{x}}^{(1)}\big). \end{aligned} \tag{40}$$

By construction (i.e., consistency), both the target and WENO weights are normalized such that $\sum_r \bar{d}_j^{(r)} = 1$ and $\sum_r \bar{w}_j^{(r)} = 1$; thus, the first constraint in (40) is identically satisfied. The most restrictive of the constraints is $\Delta\sum_r(\bar{w}^{(r)} - \bar{d}^{(r)})\bar{\mathbf{c}}_{s_l}^{(r)} = O(\Delta\bar{\mathbf{x}}^{(s_h+1-s_l)})$, and the sufficient condition that $\bar{w}^{(r)}$ must satisfy to achieve the designed-order of accuracy is given by

$$\bar{w}^{(r)} - \bar{d}^{(r)} = O\big[\delta x_1^2, \ldots, \delta x_7^2, \delta x_8^4, \ldots, \delta x_{N-7}^4, \delta x_{N-6}^2, \ldots, \delta x_N^2\big]^T, \quad \forall r \tag{41}$$

and

$$\bar{w}^{(r)} - \bar{d}^{(r)} = O\big[\delta x_1^5, \ldots, \delta x_7^5, \delta x_8^4, \ldots, \delta x_{N-7}^4, \delta x_{N-6}^5, \ldots, \delta x_N^5\big]^T, \quad \forall r \tag{42}$$

for the WENO/ESWENO$_{3\text{-}6\text{-}3}$ scheme, and WENO/ESWENO$_{5\text{-}6\text{-}5}$ schemes, respectively.

5.2 Accuracy at Smooth Extrema

To achieve design accuracy for smooth solutions with an arbitrary number of vanishing derivatives, the small parameter ε that is used (uniformly throughout the domain)

in (15) to prevent division by zero, must be constrained by the relations

$$0 < O\left(\delta x^4\right) \leq \varepsilon \leq O\left(\delta x^2\right); \tag{43}$$

$$0 < O\left(\delta x^3\right) \leq \varepsilon \leq O\left(\delta x^2\right) \tag{44}$$

for the WENO/ESWENO$_{3\text{-}6\text{-}3}$ scheme, and WENO/ESWENO$_{5\text{-}6\text{-}5}$ schemes, respectively. This is demonstrated by using conventional truncation analysis to expand (15) in terms of the smoothness indicators $\bar{\beta}^{(r)}$ and $\bar{\tau}$ in order to study the behavior of the nonlinear weights $\bar{w}^{(r)}$ in the limit of any number of vanishing derivatives.

Inspection of (15) combined with the interior and near-wall biasing mechanics that are given in (71) through (74), reveals that identical mechanics ($r = \{L4, L3, L2, L1, R1, R2, R3, R4\}$) are used at the flux points $\bar{x}_j$, $1 \leq j \leq N-1$ (no biasing is used on the boundaries $j = 0$ and $j = N$). The biasing mechanics are structurally equivalent to leading order, despite the fact that the domain of dependence varies as the boundaries are approached. Note that all Taylor expansions at the nonuniform flux point $\bar{\mathbf{x}}_j$ are equivalent to leading order. Thus, the truncation analysis that is developed in (45) through (51) is uniformly valid for $1 \leq j \leq N-1$. As such, the formal accuracy of the finite-domain ESWENO scheme is not affected by the biasing mechanics at the points near each boundary, and a single derivation about an arbitrary grid point j is sufficient. Herein, we analyze only the WENO/ESWENO$_{3\text{-}6\text{-}3}$ scheme. The analysis of the WENO/ESWENO$_{5\text{-}6\text{-}5}$ scheme follows in a like manner.

Assume that all of the required derivatives of the solution $\mathbf{u}$ are continuous and expand $\bar{\beta}_j^{(r)}$ and $\bar{\tau}_j$ at $\bar{\mathbf{x}}_j$ in a Taylor series to obtain

$$\bar{\beta}_j^{(r)} = \left[f_x(\bar{x}_j)\right]^2 \delta x^2 + O\left(\delta x^4\right), \quad \forall r. \tag{45}$$

See Appendix 8.3 for more rigorous expansions of $\bar{\beta}_j^{(r)}$. The leading order truncation error for τ_4 is

$$\tau_4 = \left[f_{4x}(\bar{x}_j)\right]^2 \delta x^8 + O\left(\delta x^9\right) \tag{46}$$

where $[f_x(\bar{x}_j)]$ and $[f_{4x}(\bar{x}_j)]$ are the first- and fourth-order derivatives of f at $\bar{x}_j$.

We first consider a case with $f_x(\bar{x}_j) \neq 0$. Substituting (45) and (46) in (15) and accounting for (43) and the upper bound on $\varepsilon \leq O(\delta x^2)$ yields

$$\frac{\bar{\tau}_j}{\varepsilon + \bar{\beta}_j^{(r)}} = O\left(\delta x^6\right), \quad \forall r \tag{47}$$

which leads to

$$\bar{w}_j^{(r)} = \bar{d}_j^{(r)} + O\left(\delta x^6\right), \quad \forall r. \tag{48}$$

From (48), we can immediately conclude that the new weights satisfy the sufficient condition (41), thereby ensuring that the WENO scheme (and ESWENO) is design-order accurate when $f_x(\bar{x}_j) \neq 0$. The result is a much faster rate of convergence

to the corresponding target linear schemes, even on coarse and moderate grids, as compared to the conventional WENO scheme of Jiang and Shu [13].

Consider next the convergence of the WENO schemes near the critical points at which the first-order and higher order derivatives of the flux are equal to zero. Let $\bar{x}_c$ be a critical point at which the flux function is sufficiently smooth and at which its derivatives up to order n_{vd} are equal to zero. That is,

$$f_x(\bar{x}_c) = f_{2x}(\bar{x}_c) = \cdots = f_{(x_{n_{vd}})}(\bar{x}_c) = 0, \qquad f_{(x_{n_{vd}+1})}(\bar{x}_c) \neq 0.$$

In contrast to the previous case in which $f_x \neq 0$, the leading truncation error terms of the smoothness indicators at the critical point are not equal to each other; thus, no additional cancellation occurs. For any number of vanishing derivatives, the following inequalities always hold in the limit $\delta x \to 0$:

$$\bar{\tau}_j \leq O\left(\delta x^8\right) \ll O\left(\delta x^4\right) \leq \varepsilon$$

which yields

$$\frac{\bar{\tau}_j}{\varepsilon + \bar{\beta}_j^{(r)}} \ll 1, \quad \forall r. \tag{49}$$

By using (49), the weights can be recast as

$$\bar{w}_j^{(r)} = \frac{\bar{d}_j^{(r)} + \bar{d}_j^{(r)} \frac{\bar{\tau}_j}{\varepsilon + \bar{\beta}_j^{(r)}}}{1 + \sum_l \bar{\beta}_j^{(l)} \frac{\bar{\tau}_j}{\varepsilon + \bar{\beta}_j^{(l)}}} = \bar{d}_j^{(r)} + O\left(\frac{\bar{\tau}_j}{\varepsilon + \bar{\beta}_j^{(r)}}\right), \quad \forall r \tag{50}$$

where we use $\sum_r \bar{d}_j^{(r)} = 1$. For any number of vanishing derivatives, we have

$$\frac{\bar{\tau}_j}{\varepsilon + \bar{\beta}_j^{(r)}} \leq \frac{\bar{\tau}_j}{\varepsilon} \leq \frac{O(\delta x^8)}{O(\delta x^4)} = O\left(\delta x^4\right), \quad \forall r. \tag{51}$$

From (51), we see that $\bar{w}_j^{(r)}$ converges to $\bar{d}_j^{(r)}$ at the rate of $O(\delta x^4)$ or higher and satisfies the sufficient condition (41). In this case, the interior stencils determine the bounds on the deviation. This is not the case for the block norm operator WENO/ESWENO$_{5\text{-}6\text{-}5}$ because of the elevated accuracy constraints at the boundaries.

Remark If the only constraint that is imposed on ε is strict positivity, then the order of convergence of the sixth-order WENO scheme (and ESWENO) with the weights given in (15) and (73) through (74) may deteriorate from six to two.

5.3 Sufficient Conditions for Consistency of Energy Stabilization

The artificial dissipation term $\mathscr{P}^{-1}\mathscr{R}_{es}\mathbf{f}$ that is given in (31) must satisfy the condition

$$\begin{aligned}\mathscr{P}^{-1}\mathscr{R}_{es}\mathbf{f} &= \mathscr{P}^{-1}\big[\Delta\hat{\Lambda}_1[\Delta]^T + \Delta[\Delta]^T\hat{\Lambda}_2\Delta[\Delta]^T + \Delta[\Delta]^T\Delta\hat{\Lambda}_3[\Delta]^T\Delta[\Delta]^T \\ &\quad + \Delta[\Delta]^T\Delta[\Delta]^T\hat{\Lambda}_4\Delta[\Delta]^T\Delta[\Delta]^T \\ &\quad + \Delta[\Delta]^T\Delta[\Delta]^T\Delta\hat{\Lambda}_5[\Delta]^T\Delta[\Delta]^T\Delta[\Delta]^T\big]\mathbf{f} \\ &= O\big(\big[\delta x^{z_{bc}},\ldots,\delta x^6\ldots,\delta x^6\ldots,\delta x^{z_{bc}}\big]^T\big) \qquad (52)\end{aligned}$$

if the $\mathscr{D}_{esweno}$ and the $\mathscr{D}_{weno}$ schemes are to have the same formal accuracy.

In this section, we show that the conditions (41), combined with the following constraints on the smoothness parameters δ_i, $i = 1, 5$:

$$\delta_i \le O(\delta x)^5 \quad \forall i \qquad (53)$$

(used in (32) to ensure the smoothness of $[\hat{\lambda}_j]_i$) guarantee that the energy-stable modifications of the conventional sixth-order WENO scheme preserve the design order of the original WENO$_{3\text{-}6\text{-}3}$ scheme and WENO$_{5\text{-}6\text{-}5}$ schemes.

First, we present a lemma that establishes how $[\hat{\lambda}_j]_i$ depends on the parameter δ_i in the interior of the domain.

Lemma 2 *The local dissipation term $[\hat{\lambda}_j]_i$ smoothly depends on x and satisfies the following dependency on the smoothness parameter δ_i:*

$$[\hat{\lambda}_j]_i = \frac{1}{2}\Big(\sqrt{\big([\lambda_j]_i\big)^2 + \delta_i^2} - [\lambda_j]_i\Big) = \frac{1}{2}\delta_i + O\big(\delta x^4\big); \quad 7 \le j \le N-6;\ 0 \le i \le 5. \qquad (54)$$

Proof Smoothness follows immediately from the differentiability of all terms in the expression. Accuracy is shown as follows.

Each $[\lambda_j]_i$ is a linear function of the weights $\bar{w}_j^{(r)}$; that is, $[\lambda_j]_i = \sum_r c_j^{(r)}\bar{w}_j^{(r)}$. See (88) through (91) given in Appendix 8.3 for the precise values of $c_j^{(r)}$.

The dissipation matrix $\mathscr{R}$ (see (6) and (7)) vanishes for the target operator $\mathscr{D}_{tarW} = \mathscr{P}^{-1}\mathscr{Q}$; that is, $\bar{w}_j^{(r)} \to \bar{d}_j^{(r)}$, which immediately yields the expression $[\lambda_j]_i|_{tarW} = \sum_r c_j^{(r)}\bar{d}_j^{(r)} = 0$. Combining these two results gives the expression

$$[\lambda_j]_i = [\lambda_j]_i - [\lambda_j]_i|_{tarW} = \sum_r c_j^{(r)}\big(\bar{w}_j^{(r)} - \bar{d}_j^{(r)}\big).$$

The most restrictive deviation of the weights from their target values which are given by (41), is $\bar{w}_j^{(r)} - \bar{d}_j^{(r)} = O(\delta x^4)$, which leads immediately to the expression

$$[\lambda_j]_i = [\lambda_j]_i - [\lambda_j]_i|_{tarW} = \sum_r c_j^{(r)} \big(\bar{w}_j^{(r)} - \bar{d}_j^{(r)}\big) = O(\delta x)^4, \quad i = 1, 5, \ j = 0, N. \tag{55}$$

Combining (55) with the definition for $\hat{\Lambda}_i$ yields the desired result

$$[\hat{\lambda}_j]_i = \frac{1}{2}\sqrt{O\big(\delta x^8\big) + \delta_i^2} - O\big(\delta x^4\big) = \frac{1}{2}\delta_i + O\big(\delta x^4\big), \quad i = 1, 5. \tag{56}$$

□

The constraints needed for δ_i in the near-boundary stencils (see derivation resulting in (56)) are $[\hat{\lambda}_j]_i = \frac{1}{2}\delta_i + O(\delta x^2)$ and $[\hat{\lambda}_j]_i = \frac{1}{2}\delta_i + O(\delta x^5)$, for the ESWENO$_{3\text{-}6\text{-}3}$ and ESWENO$_{5\text{-}6\text{-}5}$ schemes, respectively. The constraint given in (53) is sufficiently accurate to achieve design accuracy in all three scenarios.

Consider now the stabilization terms defined by (88) through (92) (in Appendix 8.3). Note that the leading order truncation terms sum to zero in the interior of the domain, thereby leading to additional cancellation of errors in the matrices $\hat{\Lambda}_1$. For example, if the deviation of the nonlinear weights from their preferred values satisfies the expression $\bar{w}_j^{(r)} = \bar{d}_j^{(r)} + O(\delta x^4)$, $\forall r$, then

$$\begin{aligned}[\lambda_j]_1 &= +\big(\bar{w}_j^{L1} - \bar{w}_{j+1}^{L1}\big) - \big(\bar{w}_j^{R1} + \bar{w}_{j-1}^{R1}\big) \\ &\quad + 3\big(\bar{w}_j^{L2} - \bar{w}_j^{R2}\big) - 5\big(\bar{w}_{j+1}^{L2} + \bar{w}_{j-1}^{R2}\big) + 2\big(\bar{w}_{j+2}^{L2} - \bar{w}_{j-2}^{R2}\big) \\ &= O\big(\delta x^5\big).\end{aligned} \tag{57}$$

Substituting (56) into (88) through (92) (in Appendix 8.3) and accounting for additional cancellations of errors, yields the result Diag$(\hat{\Lambda}_i)$, $i = 1, 5$:

$$\hat{\Lambda}_i = O\big(\big[0, \delta x^3, \ldots, \delta x^5, \ldots, \delta x^5, \ldots, \delta x^3, 0\big]^T\big). \tag{58}$$

Combining the actions of the matrices $\mathscr{P}$, Δ, $[\Delta]^T$, with that of $\hat{\Lambda}_i$ given by (58) yields the result

$$\mathscr{P}^{-1}\mathscr{R}_{es}\mathbf{f} = O\big(\big[\delta x^{z_{bc}}, \ldots, \delta x^6, \ldots, \delta x^6, \ldots, \delta x^{z_{bc}}\big]^T\big). \tag{59}$$

Thus, the energy-stabilization terms maintain the global design accuracy of the ESWENO$_{3\text{-}6\text{-}3}$ and ESWENO$_{5\text{-}6\text{-}5}$ schemes.

Remark A general contravariant grid is considered in Ref. [26], and the constraints are given in terms of the distance $\delta\xi$ rather than δx. Similar constraints on $\delta\xi$ could be formulated here.

6 Numerical Tests

The accuracy and stability characteristics of the new finite-domain ESWENO scheme are tested by using the one-dimensional advection equation and the two-dimensional Euler equations. As a more practical test, the two-dimensional interaction of a shock and a vortex is simulated. All simulations are integrated in time with the same five-step, fourth-order Runge-Kutta scheme [3], using a timestep that ensures temporal error is negligible relative to the spatial error.

Theoretically, the additional energy stabilization terms are relatively inconsequential compared with the large number of floating point operations required to build the WENO weights. In practice, however the current implementation of ESWENO is approximately 30 % more costly (clock time) than conventional WENO. This largely results from the additional sweep through the domain (and the accompanying memory accesses) that is required to build the stabilization terms from the WENO weights, and from the additional communication that is required across processor boundaries (MPI). Further optimization to better utilize multi-layer memory structure of modern chip architectures is possible but has not been pursued at this time.

6.1 Linear Wave Equation

Numerical solutions of the linear advection equation are examined to test the accuracy and dissipation characteristics of the finite-domain ESWENO schemes. The convergence rates of the $\text{ESWENO}_{3\text{-}6\text{-}3}$ and $\text{ESWENO}_{5\text{-}6\text{-}5}$ operators are calculated for a smooth solution. To evaluate the robustness of the schemes, a wave solution with a time-dependent, discontinuous inlet condition is examined. Note that because the $\text{ESWENO}_{5\text{-}6\text{-}5}$ operator is constructed by using a block $\mathscr{P}$ norm, simply assigning the boundary solution to be the boundary data (injection boundary conditions) does not maintain energy stability. Instead, both ESWENO schemes utilize the SAT penalty method to enforce the boundary conditions [2].

6.1.1 Sine Wave

To test the global order of convergence, the ESWENO schemes are used to approximate the solution to the linear advection equation with the exact solution

$$u(x,t) = \sin^r(x-t), \quad r = 1, 6, \; x \in [-\pi, \pi], \; t \in [0, 4\pi]. \tag{60}$$

The L_2 and L_∞ errors are tabulated for the $\text{ESWENO}_{3\text{-}6\text{-}3}$ and $\text{ESWENO}_{5\text{-}6\text{-}5}$ schemes in Tables 1 through 4. Tables 1 and 2 show results for a parameter value $r = 1$, while Tables 3 and 4 show results for $r = 6$. The ESWENO schemes clearly deliver the design order of convergence for the linear advection equation provided that sufficient resolution is achieved.

Table 1 The L_2 and L_∞ error and rates for the 3-6-3 scheme applied to $u(x,t) = \sin^1(x-t)$

N	e_{L_2}	e_{L_2} rate	e_{L_∞}	e_{L_∞} rate
25	2.95e–02	–	3.35e–02	–
50	3.71e–04	6.31	2.55e–04	7.04
100	2.33e–05	3.99	1.26e–05	4.34
200	1.47e–06	3.99	7.93e–07	3.99
400	9.18e–08	4.00	4.95e–08	4.00

Table 2 The L_2 and L_∞ error and rates for the 5-6-5 scheme applied to $u(x,t) = \sin^1(x-t)$

N	e_{L_2}	e_{L_2} rate	e_{L_∞}	e_{L_∞} rate
25	3.10e–05	–	3.16e–05	–
50	3.64e–07	6.41	2.66e–07	6.89
100	5.96e–09	5.93	4.18e–09	5.99
200	9.50e–11	5.97	6.96e–11	5.90
400	6.88e–12	3.79	5.49e–12	3.66

Table 3 The L_2 and L_∞ error and rates for the 3-6-3 scheme applied to $u(x,t) = \sin^6(x-t)$

N	e_{L_2}	e_{L_2} rate	e_{L_∞}	e_{L_∞} rate
25	1.10e–01	–	1.03e–01	–
50	9.09–03	3.60	1.11e–02	3.21
100	2.41e–03	1.91	3.44e–02	1.69
200	9.82e–05	4.62	8.74e–05	5.30
400	5.79e–06	4.08	4.73e–06	4.21

Table 4 The L_2 and L_∞ error and rates for the 5-6-5 scheme applied to $u(x,t) = \sin^6(x-t)$

N	e_{L_2}	e_{L_2} rate	e_{L_∞}	e_{L_∞} rate
25	1.10e–01	–	1.31e–01	–
50	1.63e–02	2.76	1.61e–02	3.03
100	2.56e–05	9.31	2.51e–05	9.32
200	4.19e–07	5.93	4.31e–07	5.86
400	6.64e–09	5.98	6.82e–09	5.98

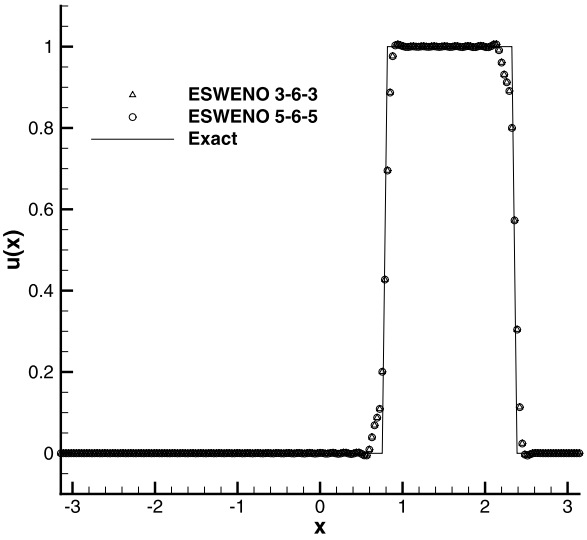


Fig. 2 The solutions from the sixth order ESWENO schemes at $t = \pi$ using $N = 200$ uniform cells are plotted along with the exact solution for the advected square wave

6.1.2 Square Wave

A square wave advection problem with the exact solution

$$u(x,t) = \frac{1}{2}\big[\tanh\big(b(x-t+3\pi/4)\big) - \tanh\big(b(x-t+\pi/4)\big)\big],$$

$$b = 2000,\ x \in [-\pi, \pi],\ t \in [0, 10] \tag{61}$$

is tested. The numerical solution is simulated up to $t = 10.0$ on a grid with $N = 200$ points. The initial condition is $u_0(x, 0) = 0$ in the domain. The wave is passed into the domain, advected through the interior, and then passed out of the domain.

The value $b = 2000$ creates a very sharp gradient in the solution that is not resolvable on the given grid. This problem tests the dissipation and oscillatory characteristics of the boundary treatment, as well as the interior treatment. A comparison of the numerical solution to the exact solution is shown for the two different boundary norms in Fig. 2. Both finite-domain ESWENO schemes are numerical stable and prevent oscillations and smearing of the discontinuity as the shock passes through the inflow boundary.

6.1.3 The Need for Energy Stabilization Terms

The conflicting requirements of accuracy, stability and smoothness can destabilize a conventional WENO formulation, especially near boundaries as the WENO biasing mechanics select inward-biased stencils that are not stable. Explosive growth of the solution can occur even for linear problems. The two previous studies using

Table 5 The stability results of the new WENO boundary closures with and without the energy stabilization terms are shown for different exact solutions and values of WENO stencil biasing parameter ε

Exact solution	$\mathscr{D}_{3\text{-}6\text{-}3}$		$\mathscr{D}_{5\text{-}6\text{-}5}$	
	WENO	ESWENO	WENO	ESWENO
$\sin(x-t), \varepsilon = \delta x^6$	**S**	**S**	**S**	**S**
$\sin(x-t), \varepsilon = \delta x^2$	**S**	**S**	**S**	**S**
$\sin^2(x-t), \varepsilon = \delta x^6$	$\mathscr{U}$	**S**	**S**	**S**
$\sin^2(x-t), \varepsilon = \delta x^2$	$\mathscr{U}$	**S**	**S**	**S**
$\sin^4(x-t), \varepsilon = \delta x^6$	$\mathscr{U}$	**S**	$\mathscr{U}$	**S**
$\sin^4(x-t), \varepsilon = \delta x^2$	$\mathscr{U}$	**S**	$\mathscr{U}$	**S**

the linear advection of the sine wave and square wave, demonstrate the robustness of the energy stabilized boundary closures (i.e. ESWENO). Next, we quantify the efficacy of the stabilization terms by comparing WENO and ESWENO formulations on problems characterized by sharp gradients that move across boundaries.

Three exact solutions, $u(x,t) = \sin^r(x-t), r = 1, 2, 4$ on the domain $x \in [-\pi, \pi]$, with two different values of the WENO stencil biasing parameter ε are simulated on a grid containing $N = 25$ points.[4] The results of this test are tabulated in Table 5. Conventional WENO simulations that diverge before $t = 10$ are marked with a $\mathscr{U}$, while stable simulations are denoted with an **S**. Longer simulation times do not change the outcome of the test.

Table 5 summarizes the impact of the energy stabilization terms for both the $\text{ESWENO}_{3\text{-}6\text{-}3}$ and $\text{ESWENO}_{5\text{-}6\text{-}5}$ schemes. The instabilities in the problem occur at the left boundary, where the Dirichlet boundary condition is imposed. The highest allowable value, $\varepsilon = \mathscr{O}(\delta x^2)$ [26], stabilizes the diagonal-norm WENO operator for some resolutions but is not effective when the simulation is severely under-resolved. This study demonstrates that the energy stable terms should be included near boundaries.

6.2 The Euler Equations

The two-dimensional Euler equations expressed in the Cartesian coordinates x and y are

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} + \frac{\partial \mathbf{G}}{\partial y} = 0 \tag{62}$$

[4] Equation (15) describes the role of the parameter ε in the WENO stencil biasing mechanics. The parameter determines the amplitude of oscillation that the biasing mechanics can detect. Using $\varepsilon = \mathscr{O}(\delta x^2)$ permits larger oscillations in the solution.

where

$$\mathbf{U} = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{bmatrix}, \qquad \mathbf{F} = \begin{bmatrix} \rho u \\ \rho u^2 + P \\ \rho uv \\ (\rho E + P)u \end{bmatrix}, \qquad \mathbf{G} = \begin{bmatrix} \rho v \\ \rho vu \\ \rho v^2 + P \\ (\rho E + P)v \end{bmatrix}.$$

The variables ρ, u, v, P, and E are the density, x velocity, y velocity, pressure, and total specific energy, respectively. The equation of state is

$$P = (\gamma - 1)\rho\left[E - \frac{1}{2}\left(u^2 + v^2\right)\right]$$

where γ is the ratio of specific heats.

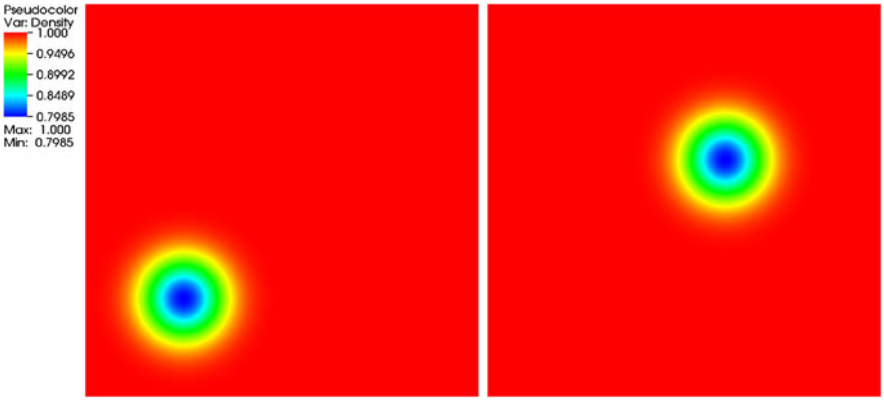
6.2.1 Two-Dimensional Inviscid Vortex

The convection of a two-dimensional Euler vortex is used to test the accuracy of the finite-domain ESWENO framework in multiple dimensions. The exact solution to the convected vortex is

$$\begin{aligned}
&f(x, y, t) = 1 - \left[\left(x - x_0 - U_\infty \cos(\alpha)t\right)^2 + \left(y - y_0 - U_\infty \sin(\alpha)t\right)^2\right], \\
&T(x, y, t) = \left[1 - \varepsilon_v^2 M_\infty^2 \frac{\gamma - 1}{8\pi^2} \exp\left(f(x, y, t)\right)\right], \qquad \rho(x, y, t) = T^{\frac{1}{\gamma - 1}}, \\
&u(x, y, t) = U_\infty \cos(\alpha) - \varepsilon_v \frac{(y - y_0 - U_\infty \sin(\alpha)t)}{2\pi} \exp\left(\frac{f(x, y, t)}{2}\right), \\
&v(x, y, t) = U_\infty \sin(\alpha) - \varepsilon_v \frac{(x - x_0 - U_\infty \cos(\alpha)t)}{2\pi} \exp\left(\frac{f(x, y, t)}{2}\right), \\
&U_\infty = M_\infty c_\infty, \ x \in (-5, 5), \ y \in (-5, 5), \ (x_0, y_0) = (-2.5, -2.5), \ t \geq 0
\end{aligned} \tag{63}$$

where the vortex strength $\varepsilon_v = 5.0$, the free stream Mach number $M_\infty = 0.5$, and $\gamma = 1.4$ are used. The vortex is advected at an angle of $\alpha = \frac{\pi}{4}$ across the grid. The density contours that are computed with the ESWENO$_{5\text{-}6\text{-}5}$ scheme at time $t = 5$ on a 128×128 grid are shown in Fig. 3.

Both boundary closures are tested for the two-dimensional vortex problem and compared to corresponding linear operators. The accuracy is affected not only by the boundary closure of the finite-difference scheme but also by the treatment of the boundary conditions. In this case, the SAT penalty method is used to specify the *exact* boundary data in a well-posed manner. Reflections from the boundaries are design-order accurate and are weakly influenced by the choice of the SAT penalty term.



(a) $t = 0$ (b) $t = 5$

Fig. 3 Density contours for two-dimensional isentropic vortex solved with ESWENO$_{5\text{-}6\text{-}5}$ scheme

Table 6 The L_2 and L_∞ errors using the ESWENO 3-6-3 scheme to simulate the isentropic vortex are tabulated

N_x	N_y	e_{L_2}	e_{L_2} rate	e_{L_∞}	e_{L_∞} rate
32	32	3.17e–02	–	1.44e–02	–
64	64	1.13e–03	4.80	7.80e–04	4.21
128	128	5.32e–05	4.42	4.69e–05	4.06

Table 7 The L_2 and L_∞ errors using the ESWENO 5-6-5 scheme to simulate the isentropic vortex are tabulated

N_x	N_y	e_{L_2}	e_{L_2} rate	e_{L_∞}	e_{L_∞} rate
32	32	2.37e–02	–	1.17e–02	–
64	64	2.17e–04	6.77	1.19e–04	6.62
128	128	2.89e–06	6.23	1.59e–06	6.22

The L_2 and L_∞ convergence rates for the ESWENO$_{3\text{-}6\text{-}3}$ and ESWENO$_{5\text{-}6\text{-}5}$ operators are tabulated in Tables 6 and 7, respectively. The results show that the linear operators achieve the design-order accuracy. The ESWENO operators converge at a rate that is higher than the design order for low resolutions but achieve design-order convergence as the resolution increases. This is the expected behavior because the nonlinear weights in the ESWENO operator asymptotically approach the linear target values as the resolution improves.

6.3 Shock Vortex Interaction

In the last Euler equation simulation, an isentropic vortex is advected through a stationary shock using the ESWENO$_{3\text{-}6\text{-}3}$ scheme. This canonical problem is commonly used to study the nonlinear physics associated with shock-vortex interactions (see references [9, 28] and the references cited therein).

The reaction of the shock to the impinging vortex depends strongly on the vortex strength ε_v (for a given M_∞). Weak vortices have minimal effect on the shock. Flow deceleration across the shock becomes more pronounced as vortex strength increases, and above a critical value of ε_v the vortex breaks down. Still stronger vortices induce significant flow nonlinearities that cause the shock to bifurcate. This strongly nonlinear case provides a challenging test problem for high resolution schemes and is used herein as our test case. Needless to say, an analytic solution does not exist for this test case.

Rather than attempt to provide further insight into the physics, we use this problem merely to demonstrate the capabilities of the ESWENO$_{3\text{-}6\text{-}3}$ scheme on a strongly nonlinear smooth/discontinuous test case. Note that the scheme used in the previous accuracy study of the inviscid vortex, is not "tuned" in any way for this test case. Thus, design order accuracy can be expected at least until the vortex impinges on the shock.

The initial and boundary conditions are imposed as follows. A stationary shock is initially located at $x = 0$. Uniform flow exists upstream of the shock given by

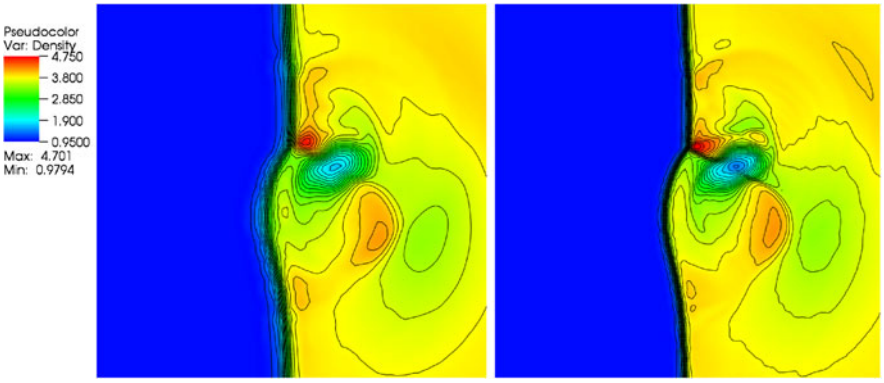
$$\begin{aligned} &\rho(x,y,0) = \rho_L = 1.0, \qquad T(x,y,0) = T_L = 1, \\ &u(x,y,0) = u_L = 1.0, \qquad v(x,y,0) = 0.0, \\ &\quad -15 \le x < 0;\ -7.5 \le y \le 7.5 \end{aligned} \tag{64}$$

where ρ, u, v, and T are the density, x-velocity, y-velocity and temperature, respectively. The free stream Mach number is $M_\infty = 3$.

The Rankine-Hugoniot relations are used to specify the uniform post-shock flow conditions

$$\begin{aligned} &\rho(x,y,0) = \rho_R = \frac{(\gamma+1)\rho_L u_L^2 M_\infty^2}{(\gamma-1)u_L^2 M_\infty^2 + 2RT_L}, \\ &u(x,y,0) = \frac{\rho_L u_L}{\rho_R}, \qquad v(x,y,0) = 0, \\ &T(x,y,0) = \frac{\gamma M_\infty^2 (2u_L^2 M_\infty^2 - RT_L(\gamma-1)/\gamma)((\gamma-1)u_L^2 M_\infty^2 + 2RT_L)}{R(\gamma+1)2u_L^2}, \\ &\quad 0 < x \le 15;\ -7.5 \le y \le 7.5 \end{aligned} \tag{65}$$

where R is the universal gas constant, and M_∞ is the free stream Mach number.



(a) 256×128 (b) 512×256

Fig. 4 Density contours on $(x, y) \in (-5, 5) \times (-5, 5)$ for the shock-vortex interaction are compared on a coarse and fine grid at $t = 75.0$

The vortex is initially located far upstream of the shock, outside the computational domain. It enters the domain after all the initial transients have left both upstream and downstream boundaries. (The discrete stationary shock solution does not satisfy the exact Rankine-Hugoniot jump conditions. Thus perturbations are formed and allowed to pass from the computational domain before the vortex arrives at the computational boundary.)

The boundary conditions for problem are (1) periodic conditions at $y = -7.5$ and $y = 7.5$ for $-15 \le x \le 15$, (2) subsonic outflow with a Dirichlet pressure boundary condition at $x = 15.0$, and (3) Dirichlet supersonic inlet conditions specified at $x = -15.0$, as determined by the following time dependent vortex inlet condition

$$
\begin{aligned}
f(x, y, t) &= 1 - \left[\big(x - x_0 - U_\infty \cos(\alpha)t\big)^2 + \big(y - y_0 - U_\infty \sin(\alpha)t\big)^2\right], \\
T(x, y, t) &= \left[1 - \varepsilon_v^2 M_\infty^2 \frac{\gamma - 1}{8\pi^2} \exp\big(f(x, y, t)\big)\right], \qquad \rho(x, y, t) = T^{\frac{1}{\gamma - 1}}, \\
u(x, y, t) &= U_\infty \cos(\alpha) - \varepsilon_v \frac{(y - y_0 - U_\infty \sin(\alpha)t)}{2\pi} \exp\left(\frac{f(x, y, t)}{2}\right), \\
v(x, y, t) &= U_\infty \sin(\alpha) - \varepsilon_v \frac{(x - x_0 - U_\infty \cos(\alpha)t)}{2\pi} \exp\left(\frac{f(x, y, t)}{2}\right), \\
U_\infty &= M_\infty c_\infty, \ (x_0, y_0) = (-70.0, 0), \ t \ge 0
\end{aligned}
\tag{66}
$$

where $\varepsilon_v = 2.0$, $M_\infty = 3.0$, $\alpha = 0$, and $\gamma = 1.4$ are used. The simulation is run on a coarser grid using 256×128 cells and a finer grid with 512×256 uniform cells. The resulting density contours at $t = 75.0$ are shown in Fig. 4.

The $\text{ESWENO}_{3\text{-}6\text{-}3}$ scheme accurately propagates the vortex up to the shock across the smooth portion of the domain. This portion of the problem is essentially equivalent to the previous inviscid vortex test case in which design order convergence was observed. As the vortex encounters the shock, the nonlinear stencil biasing mechanics of the ESWENO scheme suppress the $O(1)$ Gibbs oscillations near the shock. It is well known that the global accuracy (even WENO/ESWENO) degrades to first-order along characteristics passing through the shock. (See references [16] and [7] for further discussions on the accuracy of captured discontinuities.) Despite the loss in formal accuracy, the smooth components of the solution downstream of the shock appear to converge rapidly. Furthermore, the smooth post-shock components propagate along characteristics, and are expected to propagate with design order accuracy. A more detailed study is required to assess the efficacy of the sixth-order ESWENO, relative to lower order formulations. Studies of this nature are currently being performed, and will be documented in subsequent articles.

7 Conclusions

A general strategy was presented in 2009 by Yamaleev and Carpenter [26, 27], for constructing Energy Stable Weighted Essentially Non-Oscillatory (ESWENO) finite-difference schemes on *periodic* domains. Fisher et al. [10] provided boundary closures for the fourth-order ESWENO scheme that maintain, the WENO stencil biasing properties and satisfy the summation-by-parts (SBP) operator convention, thereby ensuring stability in an L_2 norm. Herein, boundary closures are developed for the sixth-order ESWENO scheme that maintain the WENO stencil biasing properties, while satisfying the summation-by-parts (SBP) operator convention, thereby ensuring stability in an L_2 norm. A novel set of non-uniform flux interpolation points is necessary near the boundaries to simultaneously achieve (1) accuracy, (2) the SBP convention, and (3) WENO stencil biasing mechanics. The novelty lies in the recognition that the discrete set of flux points must be consistent with (i.e., derived from) the stability norm that is used in the SBP formulation. Using the new flux points, third-order, and fifth-order boundary closures are developed that achieve stability in a diagonal and block norm, respectively. Consistency analysis is included that establishes the permissible deviations of the nonlinear weights $\bar{w}^{(r)}$ from their preferred values $\bar{d}^{(r)}$. Interior weights should satisfy the constraints $\bar{w}^{(r)} = \bar{d}^{(r)} + O(\delta x^4), \forall r$, while the near-boundary deviations of weights depend on the order of accuracy of the boundary closures.

Numerical validation is presented to assess the efficacy of the new boundary closures. The test problems included (1) the unsteady one-dimensional linear wave equation with smooth and discontinuous initial data, (2) unsteady propagation of a two-dimensional Euler vortex, and (3) unsteady interaction of a stationary shock with a traveling vortex. Design order convergence is demonstrated for both boundary closures for all smooth test cases, i.e. the third-order, diagonal-norm boundary

closures achieves fourth-order global accuracy, while the fifth-order, block-norm boundary closures achieve sixth-order global accuracy. The efficacy of the energy stabilization terms is quantified in a test case that passes a strong gradient across the near-boundary, inward-biased stencils. Conventional WENO schemes have difficulty (some diverge) for this test problem, while the ESWENO schemes do not.

Three appendices are included to provide detailed instructions on the implementation of the new ESWENO schemes. The appendices include (1) the smoothness indicators that are used in the ESWENO schemes, including near-wall stencil definitions, (2) a numerical recipe that includes the pseudo-code for the implementation of the ESWENO schemes on nonlinear hyperbolic wave equations, and (3) the diagonal-norm interpolation and weights formulae, as well as the dissipation coefficients that are needed to stabilize the baseline WENO scheme.

Appendix

8.1 Implementation of WENO

The nonlinear WENO scheme is defined as

$$\mathscr{D}_{weno}\mathbf{f} = \mathscr{P}^{-1}\Delta\bar{\mathbf{f}} = \mathscr{P}^{-1}\Delta\sum_r \bar{w}^{(r)}\bar{\mathbf{f}}^{(r)} = \mathscr{P}^{-1}\Delta\sum_r \bar{w}^{(r)}{}_M\mathscr{I}_N^{(r)}\mathbf{f}. \tag{67}$$

The fluxes are constructed by using the formula

$$\bar{f}_j = +\bar{w}_j^{L2}\bar{f}_j^{L2} + \bar{w}_j^{L1}\bar{f}_j^{L1} + \bar{w}_j^{R1}\bar{f}_j^{R1} + \bar{w}_j^{R2}\bar{f}_j^{R2}, \quad 7 \le j \le N-6. \tag{68}$$

The expressions for $\bar{f}_j$ at the six boundary closure points are slightly more complicated but conform to the same matrix conventions,

$$\bar{f}_j = \sum_{i=1}^{n_b} \bar{w}_j^i \bar{f}_j^i, \quad 1 \le j \le 6 \tag{69}$$

where n_b is the number of flux terms in each boundary flux.

The nonlinear weights $\bar{w}^{(r)}$ are defined by

$$\begin{aligned} &\bar{w}_j^{(r)} = \frac{\bar{\alpha}_j^{(r)}}{\sum_r \bar{\alpha}_j^{(r)}}; \qquad \bar{\alpha}_j^{(r)} = \bar{d}_j^{(r)}\left(1 + \frac{\bar{\tau}_j}{\varepsilon + \bar{\beta}_j^{(r)}}\right), \quad \forall r,\ 1 \le j \le N-1, \\ &\bar{w}^{(r)} = \mathrm{Diag}\big[\bar{w}_0^{(r)}, \ldots, \bar{w}_N^{(r)}\big], \quad \forall r. \end{aligned} \tag{70}$$

The coefficients $\bar{d}_j^{(r)}$ are the target weights of the candidate stencils, and $\bar{\beta}_j^{(r)}$ and $\bar{\tau}_j$ are the smoothness indicators.

The smoothness indicators $\bar{\beta}_j^{(r)}$ have the same form at all flux points, although they may not all be defined near the boundaries. Experimentation has determined that the choice of f affects the solution smoothness and dissipation properties, and that the best practice is to base the smoothness indicator on the characteristic variables rather than the characteristic fluxes. The interior smoothness indicators

$$\begin{aligned}
\bar{\beta}_j^{L2} &= (f_{j-3} - 3f_{j-2} + 2f_{j-1})^2 + (f_{j-3} - 2f_{j-2} + f_{j-1})^2; \\
\bar{\beta}_j^{L1} &= (-f_{j-1} + f_{j-0})^2 + (f_{j-2} - 2f_{j-1} + f_{j-0})^2; \\
\bar{\beta}_j^{R1} &= (-f_{j-1} + f_{j-0})^2 + (f_{j-1} - 2f_{j-0} + f_{j+1})^2; \\
\bar{\beta}_j^{R2} &= (-2f_{j-0} + 3f_{j+1} - f_{j+2})^2 + (f_{j-0} - 2f_{j+1} + f_{j+2})^2;
\end{aligned} \qquad 7 \le j \le N-6. \tag{71}$$

The indicators $\bar{\beta}^{L3}$ and $\bar{\beta}^{R3}$ (needed only near boundaries) as well as the $\bar{\beta}^r$ stencil coefficients used near the boundaries are included in Appendix 8.3.

To guarantee that the downwind stencil weight does not exceed that of the central or upwind weights, the downwind smoothness indicator is modified by using the expression

$$\bar{\beta}_j^d = \left(\frac{1}{n_s} \sum_r \left[\bar{\beta}_j^{(r)} \right]^k \right)^{1/k} \tag{72}$$

where k is an even integer, and n_s is the number of distinct interpolants contributing to the flux f_j.

An additional stencil biasing parameter, $\bar{\tau}_j$, is needed for the ESWENO scheme. Here, $\bar{\tau}_j$ is a quadratic function of the sum of fourth-order undivided differences that are available on the stencil for $\bar{f}_j$ (see Fig. 1),

$$\begin{aligned}
\bar{\tau}_j = &(-f_{j-2} + 4f_{j-1} - 6f_{j-0} + 4f_{j+1} - f_{j+2})^2 \\
&+ (-f_{j-1} + 4f_{j-0} - 6f_{j+1} + 4f_{j+2} - f_{j+3})^2, \quad 3 \le j \le N-3.
\end{aligned} \tag{73}$$

At the points $\bar{x}_1$ and $\bar{x}_2$, $\bar{\tau}_j$ is constructed from a single fourth-order undivided difference, biased toward the interior of the domain as

$$\bar{\tau}_j = (-f_1 + 4f_2 - 6f_3 + 4f_4 - f_5)^2. \tag{74}$$

The $\bar{\tau}_j$ is biased in a mirrored fashion at $\bar{x}_{N-2}$ and $\bar{x}_{N-1}$.

The parameter ε is a function of the number of points in the discretization:

$$\varepsilon = \max\left(\|f_0\|, \left\|f_0'\right\| \right)_{x \neq x_d} (\delta x)^4, \quad \delta x = \frac{1}{N} \tag{75}$$

where $\|f_0\|$ and $\|f_0\|'$ represent a norm of the flux and the gradient of the flux, respectively, as determined using initial conditions but excluding points near discontinuities.[5]

8.2 Recipe

Consider the Euler equations $U_t + F(U)_x = 0$. A recipe for calculating the gradient term $F(U)_x$ using a sixth-order ESWENO scheme is summarized below.

1. Construct the convective flux F at solution points ($\mathbf{x}$) according to the governing differential equations.
2. Construct the flux Jacobian matrices: $A = \frac{\partial F}{\partial U}$ at the flux points ($\bar{\mathbf{x}}$) by using the Roe-averaged variables that are formed from nearest neighbor solution point data. Form the eigenvector decomposition $A = S\Upsilon S^{-1}$, where S is the matrix of right eigenvectors and Υ is the diagonal matrix of eigenvalues.
3. For the flux point $\bar{x}_j$, use the eigenvector matrix $\bar{S}_j^{-1}$ to transform the solution and the flux at all points $\mathbf{x}$ into characteristic form $U_c = \bar{S}^{-1}U$; $F_c(U) = \bar{S}^{-1}F(U) = \Upsilon U_c$.
   Remark Forming the characteristic variables and fluxes is only necessary for nodal points that are located in the stencil of a given flux point. The presentation herein assumes that the transformed fluxes and variables are available at all points but can be reduced to improve performance.
4. Form the Lax-Friedrichs characteristic fluxes $\mathbf{f}_\mathbf{c}^{\pm} = \frac{1}{2}(F_c \pm \Upsilon_{max} U_c)$, where Υ_{max} is a diagonal matrix of the maximum local eigenvalues contained within each stencil.
5. Perform interpolations on each candidate stencil, $\bar{f}_j^{(r)} = ({}_M\mathscr{I}_N^{(r)}\mathbf{f}_\mathbf{c})_j$.
6. Calculate the stencil biasing parameters:
   a. Calculate $\bar{\beta}^r$ for each flux point according to (71), except at the end points.
   b. Calculate $\bar{\tau}_j$ according to (73) and (74).
      Remark The smoothness indicators are calculated by using the characteristic variables U_c in the stencil of the flux point.
7. Calculate and normalize the weights using the stencil biasing parameters and the target weights in (70) (see (15) in text).
8. Calculate Λ_i from the weights by using (88) through (91) for the diagonal-norm scheme.
9. Modify the diagonal of Λ_i to be smoothly positive according to (32).
10. Calculate the WENO flux from the weights and candidate interpolations by using (68) (see (20) in the text).
11. Calculate the energy stable flux $\bar{\boldsymbol{\psi}}$ from (33).

[5]For applications where the solution changes drastically during simulation, ε can be rescaled as necessary.

12. Reconstruct the fluxes in characteristic space, $\bar{f}_j = \bar{f}_j^+ + \bar{f}_j^-$ and $\psi_j = \psi_j^+ + \psi_j^-$, and then transform the fluxes back to physical space by using $\hat{f}_j = S(\bar{f}_j + \psi_j)$.
13. After calculating the ESWENO interpolation at all flux points, calculate the gradient using the inverse of the $\mathscr{P}$ norm, as in (34).

Remark Note that all of the presented equations are for the forward-propagating waves, $\bar{f}_j^+$. The equations for interpolation of backward-propagating waves ($\bar{f}_j^-$) are found in exactly the same manner, except that the downwind stencil in (72) is the far left instead of the far right stencil. The Λ_i terms are to be negative instead of positive; therefore, we use

$$[\hat{\lambda}_j]_i = -\frac{1}{2}\left(\sqrt{\left([\lambda_j]_i\right)^2 + \delta_i^2} + [\lambda_j]_i\right), \quad i = 0, 5, \;\; j = 0, N. \tag{76}$$

8.3 Sixth-Order Diagonal-Norm Operator: $\mathscr{D}_{3\text{-}6\text{-}3}$

8.3.1 Differentiation Matrix

The upper left matrix quadrant of the target SBP differentiation operator $\mathscr{D}_{3\text{-}6\text{-}3}$ can be written in the form

$$\mathscr{D}_{3\text{-}6\text{-}3} = \frac{1}{\delta x} \times$$

$$\begin{pmatrix}
-\frac{21600}{13649} & \frac{104009}{54596} & \frac{30443}{81894} & -\frac{33311}{27298} & \frac{16863}{27298} & -\frac{15025}{163788} & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{104009}{240260} & 0 & -\frac{311}{72078} & \frac{20229}{24026} & -\frac{24337}{48052} & \frac{36661}{360390} & 0 & 0 & 0 & 0 & 0 & 0 \\
-\frac{30443}{162660} & \frac{311}{32532} & 0 & -\frac{11155}{16266} & \frac{41287}{32532} & -\frac{21999}{54220} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{33311}{107180} & -\frac{20229}{21436} & \frac{485}{1398} & 0 & \frac{4147}{21436} & \frac{25427}{321540} & \frac{72}{5359} & 0 & 0 & 0 & 0 & 0 \\
-\frac{16863}{78770} & \frac{24337}{31508} & -\frac{41287}{47262} & -\frac{4147}{15754} & 0 & \frac{342523}{472620} & -\frac{1296}{7877} & \frac{144}{7877} & 0 & 0 & 0 & 0 \\
\frac{15025}{525612} & -\frac{36661}{262806} & \frac{21999}{87602} & -\frac{25427}{262806} & -\frac{342523}{525612} & 0 & \frac{32400}{43801} & -\frac{6480}{43801} & \frac{720}{43801} & 0 & 0 & 0 \\
0 & 0 & 0 & -\frac{1}{60} & \frac{3}{20} & -\frac{3}{4} & 0 & \frac{3}{4} & -\frac{3}{20} & \frac{1}{60} & 0 & 0 \\
0 & 0 & 0 & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0
\end{pmatrix}. \tag{77}$$

The matrix elements in the lower right quadrant are structurally equivalent relative to the outflow boundary, but with the opposite sign. The diagonal mass matrix $\mathscr{P}$ is

$$\mathscr{P}_{3\text{-}6\text{-}3} = \delta x \,\mathrm{Diag}\left(\frac{13649}{43200} \quad \frac{12013}{8640} \quad \frac{2711}{4320} \quad \frac{5359}{4320} \quad \frac{7877}{8640} \quad \frac{43801}{43200} \quad 1 \quad \cdots\right).$$

The skew-symmetric matrix $\mathscr{Q}_{3\text{-}6\text{-}3}$ follows immediately from $\mathscr{P}_{3\text{-}6\text{-}3} D_{3\text{-}6\text{-}3}$.

The distribution the flux points $\bar{\mathbf{x}}$ as described in Lemma 1 follows immediately from the norm $\mathscr{P}_{3\text{-}6\text{-}3}$ and can be written as

$$\bar{\mathbf{x}} = \left[0, \bar{x}_1, \ldots, \bar{x}_5, \frac{11\delta x}{2}, \ldots, \left(1 - \frac{11\delta x}{2}\right), (1-\bar{x}_5), \ldots, (1-\bar{x}_1), 1\right]^T \tag{78}$$

where

$$\begin{aligned} \bar{x}_1 &= \frac{13649\delta x}{43200}; \qquad \bar{x}_2 = \frac{36857\delta x}{21600}; \qquad \bar{x}_3 = \frac{4201\delta x}{1800}; \\ \bar{x}_4 &= \frac{77207\delta x}{21600}; \qquad \bar{x}_5 = \frac{193799\delta x}{43200}. \end{aligned} \tag{79}$$

Then, the vector $\Delta\bar{\mathbf{x}}$ is the reciprocal of the diagonal elements of $\mathscr{P}_d$. The product $\mathscr{P}_d^{-1}\Delta\bar{\mathbf{x}}$ is trivially the identity vector $\mathbf{1}$ because $\mathscr{P}_d$ is a diagonal matrix. Note that all integer coefficients are exact. Case should be taken to ensure that the rational coefficients retain the full working precision of the coding language (e.g. 64 bit arithmetic).

8.3.2 Interpolation Operators

The tri-diagonal interpolation matrices ${}_M\mathscr{I}_N^{(r)}$, $r = \{L4, L3, L2, L1, R1, R2, R3, R4\}$ for the ESWENO$_{3\text{-}6\text{-}3}$ scheme, presented in compressed format using the arrays I_r, $r = \{L4, L3, L2, L1, R1, R2, R3, R4\}$, are given by

$$I^{L4} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{370193}{86400} & -\frac{59699}{5400} & \frac{74599}{9600} \\ \frac{13}{3} & -\frac{67}{6} & \frac{47}{6} \\ \vdots & \vdots & \vdots \\ \frac{13}{3} & -\frac{67}{6} & \frac{47}{6} \\ \frac{75481}{17280} & -\frac{81001}{7200} & \frac{681007}{86400} \\ \frac{44641}{10800} & -\frac{230957}{21600} & \frac{2177}{288} \\ \frac{69251}{14400} & -\frac{88447}{7200} & \frac{40681}{4800} \\ \frac{5011}{1350} & -\frac{13993}{1440} & \frac{151319}{21600} \\ \frac{422857}{86400} & -\frac{67351}{5400} & \frac{82351}{9600} \\ 0 & 0 & 0 \end{pmatrix}, \qquad I^{L3} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{4813}{2400} & -\frac{120641}{21600} & \frac{24731}{5400} \\ \frac{31079}{17280} & -\frac{36599}{7200} & \frac{370193}{86400} \\ \frac{11}{6} & -\frac{31}{6} & \frac{13}{3} \\ \vdots & \vdots & \vdots \\ \frac{11}{6} & -\frac{31}{6} & \frac{13}{3} \\ \frac{53401}{28800} & -\frac{56401}{10800} & \frac{75481}{17280} \\ \frac{36889}{21600} & -\frac{3873}{800} & \frac{44641}{10800} \\ \frac{30859}{14400} & -\frac{2857}{480} & \frac{69251}{14400} \\ \frac{10211}{7200} & -\frac{89209}{21600} & \frac{5011}{1350} \\ \frac{38191}{17280} & -\frac{43951}{7200} & \frac{422857}{86400} \\ 0 & 0 & 0 \end{pmatrix}, \tag{80}$$

$$I^{L2} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{2083}{14400} & -\frac{4487}{7200} & \frac{7097}{4800} \\ \frac{931}{2160} & -\frac{31027}{21600} & \frac{4813}{2400} \\ \frac{8999}{28800} & -\frac{11999}{10800} & \frac{31079}{17280} \\ \frac{1}{3} & -\frac{7}{6} & \frac{11}{6} \\ \vdots & \vdots & \vdots \\ \frac{1}{3} & -\frac{7}{6} & \frac{11}{6} \\ \frac{29401}{86400} & -\frac{25801}{21600} & \frac{53401}{28800} \\ \frac{127}{450} & -\frac{4277}{4320} & \frac{36889}{21600} \\ \frac{763}{1600} & -\frac{11663}{7200} & \frac{30859}{14400} \\ \frac{269}{2160} & -\frac{11723}{21600} & \frac{10211}{7200} \\ \frac{15151}{28800} & -\frac{18751}{10800} & \frac{38191}{17280} \\ 0 & 0 & 0 \end{pmatrix}, \qquad I^{L1} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -\frac{3653}{21600} & \frac{13649}{21600} & \frac{967}{1800} \\ -\frac{109}{576} & \frac{2507}{2400} & \frac{2083}{14400} \\ -\frac{3097}{21600} & \frac{5129}{7200} & \frac{931}{2160} \\ -\frac{15001}{86400} & \frac{18601}{21600} & \frac{8999}{28800} \\ -\frac{1}{6} & \frac{5}{6} & \frac{1}{3} \\ \vdots & \vdots & \vdots \\ -\frac{1}{6} & \frac{5}{6} & \frac{1}{3} \\ -\frac{15001}{86400} & \frac{5}{6} & \frac{29401}{86400} \\ -\frac{3097}{21600} & \frac{18601}{21600} & \frac{127}{450} \\ -\frac{109}{576} & \frac{5129}{7200} & \frac{763}{1600} \\ -\frac{3653}{21600} & \frac{2507}{2400} & \frac{269}{2160} \\ -\frac{13649}{86400} & \frac{13649}{21600} & \frac{15151}{28800} \\ 0 & 0 & 0 \end{pmatrix}. \quad (81)$$

The interpolants I_r, $r = \{R1, R2, R3, R4\}$ are mirror images of I_r, $r = \{L1, L2, L3, L4\}$. For example, the interpolants I_r, $r = \{R1, R2\}$ are given by

$$I^{R1} = \begin{pmatrix} 0 & 0 & 0 \\ \frac{15151}{28800} & \frac{13649}{21600} & -\frac{13649}{86400} \\ \frac{269}{2160} & \frac{2507}{2400} & -\frac{3653}{21600} \\ \frac{763}{1600} & \frac{5129}{7200} & -\frac{109}{576} \\ \frac{127}{450} & \frac{18601}{21600} & -\frac{3097}{21600} \\ \frac{29401}{86400} & \frac{5}{6} & -\frac{15001}{86400} \\ \frac{1}{3} & \frac{5}{6} & -\frac{1}{6} \\ \vdots & \vdots & \vdots \\ \frac{1}{3} & \frac{5}{6} & -\frac{1}{6} \\ \frac{8999}{28800} & \frac{18601}{21600} & -\frac{15001}{86400} \\ \frac{931}{2160} & \frac{5129}{7200} & -\frac{3097}{21600} \\ \frac{2083}{14400} & \frac{2507}{2400} & -\frac{109}{576} \\ \frac{967}{1800} & \frac{13649}{21600} & -\frac{3653}{21600} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad I^{R2} = \begin{pmatrix} 0 & 0 & 0 \\ \frac{38191}{17280} & -\frac{18751}{10800} & \frac{15151}{28800} \\ \frac{10211}{7200} & -\frac{11723}{21600} & \frac{269}{2160} \\ \frac{30859}{14400} & -\frac{11663}{7200} & \frac{763}{1600} \\ \frac{36889}{21600} & -\frac{4277}{4320} & \frac{127}{450} \\ \frac{53401}{28800} & -\frac{25801}{21600} & \frac{29401}{86400} \\ \frac{11}{6} & -\frac{7}{6} & \frac{1}{3} \\ \vdots & \vdots & \vdots \\ \frac{11}{6} & -\frac{7}{6} & \frac{1}{3} \\ \frac{31079}{17280} & -\frac{11999}{10800} & \frac{8999}{28800} \\ \frac{4813}{2400} & -\frac{31027}{21600} & \frac{931}{2160} \\ \frac{7097}{4800} & -\frac{4487}{7200} & \frac{2083}{14400} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \quad (82)$$

Reconstructing ${}_M\mathscr{I}_N^{(r)}$ from the data I_r is accomplished by populating the appropriate diagonals in the interpolation matrices. The middle column of I^r coincides with the diagonal of ${}_M\mathscr{I}_N^{(r)}$, $r = \{L4, L3, L2, L1, R1, R2, R3, R4\}$ that is shifted $\{-4, -3, -2, -1, 0, 1, 2, 3\}$ relative to the main diagonal.

8.3.3 Target Weights

The target weights for the five interpolation stencils are

$$\mathbf{d}_{3\text{-}6\text{-}3} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \frac{14400}{15151} & \frac{3475559}{5786318410} \\ 0 & 0 & 0 & \frac{17609}{29224} & \frac{34899909}{19653140} & -\frac{720681767}{494416620} \\ 0 & 0 & -\frac{113713}{74988} & \frac{364030091}{204342300} & \frac{6384599}{6180300} & -\frac{73697}{247212} \\ 0 & \frac{86153}{1039608} & -\frac{140213393}{604921905} & \frac{903169}{758765} & -\frac{948239}{9439656} & \frac{15}{254} \\ -\frac{15025}{2221158} & \frac{6395825773}{172578423705} & -\frac{367011523}{8390397630} & \frac{72016320}{134993999} & \frac{190085760}{441044401} & \frac{1440}{29401} \\ 0 & 0 & \frac{1}{20} & \frac{9}{20} & \frac{9}{20} & \frac{1}{20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \frac{1}{20} & \frac{9}{20} & \frac{9}{20} & \frac{1}{20} \\ 0 & 0 & \frac{1440}{29401} & \frac{190085760}{441044401} & \frac{72016320}{134993999} & -\frac{367011523}{8390397630} \\ 0 & 0 & \frac{15}{254} & -\frac{948239}{9439656} & \frac{903169}{758765} & -\frac{140213393}{604921905} \\ 0 & 0 & -\frac{73697}{247212} & \frac{6384599}{6180300} & \frac{364030091}{204342300} & -\frac{113713}{74988} \\ 0 & \frac{58297}{735192} & -\frac{720681767}{494416620} & \frac{34899909}{19653140} & \frac{17609}{29224} & 0 \\ -\frac{15025}{2537142} & \frac{13296322573}{242239975305} & \frac{3475559}{5786318410} & \frac{14400}{15151} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}. \tag{83}$$

For brevity, the last two columns are excluded from the table. Recognizing the relationship between columns 4 and 7 (and 5 and 6), columns 6 and 7 can be recovered from the mirror images of columns 1 and 2, respectively. Note that the WENO/ESWENO$_{3\text{-}6\text{-}3}$ scheme requires the inclusion of two extra stencils near each boundary.

8.3.4 Smoothness Indicators Coefficients

The smoothness indicators $\bar{\beta}^r$ measure the smoothness of the data used by each interpolant. This is accomplished by comparing the L_2 of the undivided first and

second derivatives, evaluated at the flux point $\bar{x}_j$.

$$\bar{\beta}^r = \left[\delta x\, f_x^r(\bar{x}_j)\right]^2 + \left[\delta x^2 f_{2x}^r(\bar{x}_j)\right]^2, \quad 1 \le j \le N-1. \tag{84}$$

The leading order truncation term is identical in all cases. For example, the interior terms

$$\begin{aligned}
\bar{\beta}^{L2} &= {f_x}^2\delta x^2 + \left({f_{2x}}^2 - \frac{23}{12} f_x f_{3x}\right)\delta x^4 + (+3 f_{2x} f_{3x} - 2 f_x f_{4x})\delta x^5 + O\left(\delta x^6\right),\\
\bar{\beta}^{L1} &= {f_x}^2\delta x^2 + \left({f_{2x}}^2 + \frac{1}{12} f_x f_{3x}\right)\delta x^4 + (+1 f_{2x} f_{3x})\delta x^5 + O\left(\delta x^6\right),\\
\bar{\beta}^{R1} &= {f_x}^2\delta x^2 + \left({f_{2x}}^2 + \frac{1}{12} f_x f_{3x}\right)\delta x^4 + (-1 f_{2x} f_{3x})\delta x^5 + O\left(\delta x^6\right),\\
\bar{\beta}^{R2} &= {f_x}^2\delta x^2 + \left({f_{2x}}^2 - \frac{23}{12} f_x f_{3x}\right)\delta x^4 + (-3 f_{2x} f_{3x} + 2 f_x f_{4x})\delta x^5 + O\left(\delta x^6\right).
\end{aligned} \tag{85}$$

The coefficients needed for the first derivative term $\delta x \frac{\partial}{\partial x}$ in the smoothness indicators throughout the spatial domain can be expressed as follows.

$$L4 = \begin{pmatrix}
0 & 0 & 0\\
0 & 0 & 0\\
0 & 0 & 0\\
0 & 0 & 0\\
0 & 0 & 0\\
\frac{128999}{43200} & -\frac{150599}{21600} & \frac{172199}{43200}\\
3 & -7 & 4\\
\vdots & \vdots & \vdots\\
3 & -7 & 4\\
\frac{130201}{43200} & -\frac{151801}{21600} & \frac{173401}{43200}\\
\frac{63193}{21600} & -\frac{73993}{10800} & \frac{84793}{21600}\\
\frac{5699}{1800} & -\frac{6599}{900} & \frac{7499}{1800}\\
\frac{60343}{21600} & -\frac{71143}{10800} & \frac{81943}{21600}\\
\frac{137551}{43200} & -\frac{159151}{21600} & \frac{180751}{43200}\\
0 & 0 & 0
\end{pmatrix}, \qquad
L3 = \begin{pmatrix}
0 & 0 & 0\\
0 & 0 & 0\\
0 & 0 & 0\\
0 & 0 & 0\\
\frac{44807}{21600} & -\frac{55607}{10800} & \frac{66407}{21600}\\
\frac{85799}{43200} & -\frac{107399}{21600} & \frac{128999}{43200}\\
2 & -5 & 3\\
\vdots & \vdots & \vdots\\
2 & -5 & 3\\
\frac{87001}{43200} & -\frac{108601}{21600} & \frac{130201}{43200}\\
\frac{41593}{21600} & -\frac{52393}{10800} & \frac{63193}{21600}\\
\frac{3899}{1800} & -\frac{4799}{900} & \frac{5699}{1800}\\
\frac{38743}{21600} & -\frac{49543}{10800} & \frac{60343}{21600}\\
\frac{94351}{43200} & -\frac{115951}{21600} & \frac{137551}{43200}\\
0 & 0 & 0
\end{pmatrix}, \tag{86}$$

$$L2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{1501}{1800} & -\frac{2401}{900} & \frac{3301}{1800} \\ \frac{23207}{21600} & -\frac{34007}{10800} & \frac{44807}{21600} \\ \frac{42599}{43200} & -\frac{64199}{21600} & \frac{85799}{43200} \\ 1 & -3 & 2 \\ \vdots & \vdots & \vdots \\ 1 & -3 & 2 \\ \frac{43801}{43200} & -\frac{65401}{21600} & \frac{87001}{43200} \\ \frac{19993}{21600} & -\frac{30793}{10800} & \frac{41593}{21600} \\ \frac{2099}{1800} & -\frac{2999}{900} & \frac{3899}{1800} \\ \frac{17143}{21600} & -\frac{27943}{10800} & \frac{38743}{21600} \\ \frac{51151}{43200} & -\frac{72751}{21600} & \frac{94351}{43200} \\ 0 & 0 & 0 \end{pmatrix}, \quad L1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \frac{4457}{21600} & -\frac{15257}{10800} & \frac{26057}{21600} \\ -\frac{299}{1800} & -\frac{601}{900} & \frac{1501}{1800} \\ \frac{1607}{21600} & -\frac{12407}{10800} & \frac{23207}{21600} \\ -\frac{601}{43200} & -\frac{20999}{21600} & \frac{42599}{43200} \\ 0 & -1 & 1 \\ \vdots & \vdots & \vdots \\ 0 & -1 & 1 \\ \frac{601}{43200} & -\frac{22201}{21600} & \frac{43801}{43200} \\ -\frac{1607}{21600} & -\frac{9193}{10800} & \frac{19993}{21600} \\ \frac{299}{1800} & -\frac{1199}{900} & \frac{2099}{1800} \\ -\frac{4457}{21600} & -\frac{6343}{10800} & \frac{17143}{21600} \\ \frac{7951}{43200} & -\frac{29551}{21600} & \frac{51151}{43200} \\ 0 & 0 & 0 \end{pmatrix}. \tag{87}$$

The coefficients needed for the first derivative term $\delta x \frac{\partial}{\partial x}$ in the rightward biased stencils ($r = R1, R2, R3, R4$) are the mirror images of those used for leftward biased stencils ($r = L1, L2, L3, L4$).

The second derivative stencil coefficients are always

$$\delta x^2 f^r_{2x}(\bar{x}_j) = f^r_{r+j-1} - 2f^r_{r+j} + f^r_{r+j+1}; \quad 1 \le j \le N-1, \forall r.$$

8.3.5 Energy-Stable Terms

The stabilization matrix $\mathrm{Diag}(\Lambda_0)$ is identically zero for the $\mathrm{ESWENO}_{3\text{-}6\text{-}3}$ scheme (as it is with every consistent scheme). The nonzero terms of the diagonal matrices Λ_i, $i = 1, 5$ are

$$\mathrm{Diag}(\Lambda_1) =$$

$$\begin{pmatrix}
0 \\
(-14612\bar{w}_2^{L1} + 12498\bar{w}_3^{L2} + 173268\bar{w}_4^{L3} + 370193\bar{w}_5^{L4} + 18155\bar{w}_1^{R1} - 27298\bar{w}_1^{R2} \cdots \\
\quad - 27298\bar{w}_1^{R3} - 27298\bar{w}_1^{R4})/172800 \\
(21044\bar{w}_2^{L1} - 16350\bar{w}_3^{L1} - 41346\bar{w}_3^{L2} + 37240\bar{w}_4^{L2} - 309296\bar{w}_4^{L3} + 155395\bar{w}_5^{L3} \cdots \\
\quad - 584991\bar{w}_5^{L4} + 374400\bar{w}_6^{L4} + 13649\bar{w}_1^{R1} - 50268\bar{w}_2^{R1} + 104555\bar{w}_1^{R2} - 61028\bar{w}_2^{R2} \cdots \\
\quad - 86400\bar{w}_1^{R3} - 61028\bar{w}_2^{R3} - 86400\bar{w}_1^{R4} - 61028\bar{w}_2^{R4})/172800 \\
(45054\bar{w}_3^{L1} - 12388\bar{w}_4^{L1} + 57552\bar{w}_3^{L2} - 86868\bar{w}_4^{L2} + 26997\bar{w}_5^{L2} + 86400\bar{w}_4^{L3} \cdots \\
\quad - 283793\bar{w}_5^{L3} + 158400\bar{w}_6^{L3} + 86400\bar{w}_5^{L4} - 590400\bar{w}_6^{L4} + 14612\bar{w}_2^{R1} + 12354\bar{w}_3^{R1} \cdots \\
\quad - 45453\bar{w}_1^{R2} + 36132\bar{w}_2^{R2} - 28848\bar{w}_3^{R2} + 336457\bar{w}_1^{R3} - 86400\bar{w}_2^{R3} - 28848\bar{w}_3^{R3} \cdots \\
\quad - 86400\bar{w}_1^{R4} - 86400\bar{w}_2^{R4} - 28848\bar{w}_3^{R4})/172800 \\
(468\bar{w}_4^{L1} - 15001\bar{w}_5^{L1} + 36772\bar{w}_4^{L2} - 68995\bar{w}_5^{L2} + 28800\bar{w}_6^{L2} + 36772\bar{w}_4^{L3} \cdots \\
\quad + 86400\bar{w}_5^{L3} - 288000\bar{w}_6^{L3} + 86400\bar{w}_5^{L4} + 86400\bar{w}_6^{L4} + 16350\bar{w}_3^{R1} - 25244\bar{w}_4^{R1} \cdots \\
\quad - 10760\bar{w}_2^{R2} + 98754\bar{w}_3^{R2} - 49628\bar{w}_4^{R2} - 190955\bar{w}_1^{R3} + 234304\bar{w}_2^{R3} - 86400\bar{w}_3^{R3} \cdots \\
\quad + 654759\bar{w}_1^{R4} - 86400\bar{w}_2^{R4} - 86400\bar{w}_3^{R4})/172800 \\
(17405\bar{w}_5^{L1} - 14400\bar{w}_6^{L1} + 44402\bar{w}_5^{L2} - 72000\bar{w}_6^{L2} + 28800\bar{w}_7^{L2} + 44402\bar{w}_5^{L3} \cdots \\
\quad + 86400\bar{w}_6^{L3} + 44402\bar{w}_5^{L4} + 86400\bar{w}_6^{L4} + 12388\bar{w}_4^{R1} - 12597\bar{w}_5^{R1} - 41202\bar{w}_3^{R2} \cdots \\
\quad + 61156\bar{w}_4^{R2} - 41998\bar{w}_5^{R2} - 122532\bar{w}_2^{R3} + 329106\bar{w}_3^{R3} - 422857\bar{w}_1^{R4} + 518876\bar{w}_2^{R4} \cdots \\
\quad - 86400\bar{w}_3^{R4})/172800 \\
(14400\bar{w}_6^{L1} - 14400\bar{w}_7^{L1} + 43200\bar{w}_6^{L2} - 72000\bar{w}_7^{L2} + 28800\bar{w}_8^{L2} + 43200\bar{w}_6^{L3} \cdots \\
\quad + 43200\bar{w}_6^{L4} + 15001\bar{w}_5^{R1} - 14400\bar{w}_6^{R1} - 24384\bar{w}_4^{R2} + 73803\bar{w}_5^{R2} - 43200\bar{w}_6^{R2} \cdots \\
\quad - 185154\bar{w}_3^{R3} - 320704\bar{w}_2^{R4} + 645858\bar{w}_3^{R4})/172800 \\
(14400\bar{w}_7^{L1} - 14400\bar{w}_8^{L1} + 43200\bar{w}_7^{L2} - 72000\bar{w}_8^{L2} + 28800\bar{w}_9^{L2} + 14400\bar{w}_6^{R1} \cdots \\
\quad - 14400\bar{w}_7^{R1} - 29401\bar{w}_5^{R2} + 72000\bar{w}_6^{R2} - 43200\bar{w}_7^{R2} - 415506\bar{w}_3^{R4})/172800 \\
(\bar{w}_8^{L1} - \bar{w}_9^{L1} + 3\bar{w}_8^{L2} - 5\bar{w}_9^{L2} + 2\bar{w}_{10}^{L2} + \bar{w}_7^{R1} \cdots \\
\quad - \bar{w}_8^{R1} - 2\bar{w}_6^{R2} + 5\bar{w}_7^{R2} - 3\bar{w}_8^{R2})/12 \\
\vdots \\
(\bar{w}_{N-8}^{L1} - \bar{w}_{N-7}^{L1} + 3\bar{w}_{N-8}^{L2} - 5\bar{w}_{N-7}^{L2} + 2\bar{w}_{N-6}^{L2} + \bar{w}_{N-9}^{R1} \cdots \\
\quad - \bar{w}_{N-8}^{R1} - 2\bar{w}_{16}^{R2} + 5\bar{w}_{N-9}^{R2} - 3\bar{w}_{N-8}^{R2})/12 \\
(14400\bar{w}_{N-7}^{L1} - 14400\bar{w}_{N-6}^{L1} + 43200\bar{w}_{N-7}^{L2} - 72000\bar{w}_{N-6}^{L2} + 29401\bar{w}_{N-5}^{L2} + 415506\bar{w}_{N-3}^{L4} \cdots \\
\quad + 14400\bar{w}_{N-8}^{R1} - 14400\bar{w}_{N-7}^{R1} - 28800\bar{w}_{N-9}^{R2} + 72000\bar{w}_{N-8}^{R2} - 43200\bar{w}_{N-7}^{R2})/172800 \\
\vdots \\
(18155\bar{w}_{N-1}^{L1} + 27298\bar{w}_{N-1}^{L2} + 27298\bar{w}_{N-1}^{L3} + 27298\bar{w}_{N-1}^{L4} + 14612\bar{w}_{N-2}^{R1} - 12498\bar{w}_{N-3}^{R2} \cdots \\
\quad - 173268\bar{w}_{N-4}^{R3} - 370193\bar{w}_{N-5}^{R4})/172800 \\
0
\end{pmatrix}, \tag{88}$$

$$\mathrm{Diag}(\Lambda_2) =$$

$$\begin{pmatrix}
0 \\
(14612\bar{w}_2^{L1} - 24996\bar{w}_3^{L2} - 519804\bar{w}_4^{L3} - 1480772\bar{w}_5^{L4} - 13649\bar{w}_1^{R1} - 13649\bar{w}_1^{R2} \cdots \\
\quad - 13649\bar{w}_1^{R3} - 13649\bar{w}_1^{R4})/172800 \\
(16350\bar{w}_3^{L1} + 16350\bar{w}_3^{L2} - 74480\bar{w}_4^{L2} + 98788\bar{w}_4^{L3} - 466185\bar{w}_5^{L3} + 274201\bar{w}_5^{L4} \cdots \\
\quad - 1497600\bar{w}_6^{L4} - 14612\bar{w}_2^{R1} + 90906\bar{w}_1^{R2} - 14612\bar{w}_2^{R2} - 100049\bar{w}_1^{R3} - 14612\bar{w}_2^{R3} \cdots \\
\quad - 100049\bar{w}_1^{R4} - 14612\bar{w}_2^{R4})/172800 \\
(12388\bar{w}_4^{L1} + 12388\bar{w}_4^{L2} - 53994\bar{w}_5^{L2} + 12388\bar{w}_4^{L3} + 101401\bar{w}_5^{L3} - 475200\bar{w}_6^{L3} \cdots \\
\quad + 101401\bar{w}_5^{L4} + 273600\bar{w}_6^{L4} - 16350\bar{w}_3^{R1} + 21520\bar{w}_2^{R2} - 16350\bar{w}_3^{R2} + 572865\bar{w}_1^{R3} \cdots \\
\quad - 101012\bar{w}_2^{R3} - 16350\bar{w}_3^{R3} - 272849\bar{w}_1^{R4} - 101012\bar{w}_2^{R4} - 16350\bar{w}_3^{R4})/172800 \\
(15001\bar{w}_5^{L1} + 15001\bar{w}_5^{L2} - 57600\bar{w}_6^{L2} + 15001\bar{w}_5^{L3} + 100800\bar{w}_6^{L3} + 15001\bar{w}_5^{L4} \cdots \\
\quad + 100800\bar{w}_6^{L4} - 12388\bar{w}_4^{R1} + 82404\bar{w}_3^{R2} - 12388\bar{w}_4^{R2} + 367596\bar{w}_2^{R3} - 102750\bar{w}_3^{R3} \cdots \\
\quad + 1691428\bar{w}_1^{R4} - 273812\bar{w}_2^{R4} - 102750\bar{w}_3^{R4})/172800 \\
(14400\bar{w}_6^{L1} + 14400\bar{w}_6^{L2} - 57600\bar{w}_7^{L2} + 14400\bar{w}_6^{L3} + 14400\bar{w}_6^{L4} - 15001\bar{w}_5^{R1} \cdots \\
\quad + 48768\bar{w}_4^{R2} - 15001\bar{w}_5^{R2} + 555462\bar{w}_3^{R3} + 1282816\bar{w}_2^{R4} - 275550\bar{w}_3^{R4})/172800 \\
(7200\bar{w}_7^{L1} + 7200\bar{w}_7^{L2} - 28800\bar{w}_8^{L2} - 7200\bar{w}_6^{R1} + 29401\bar{w}_5^{R2} - 7200\bar{w}_6^{R2} \cdots \\
\quad + 831012\bar{w}_3^{R4})/86400 \\
(\bar{w}_8^{L1} + \bar{w}_8^{L2} - 4\bar{w}_9^{L2} - \bar{w}_7^{R1} + 4\bar{w}_6^{R2} - \bar{w}_7^{R2})/12 \\
\vdots \\
(\bar{w}_{N-7}^{L1} + \bar{w}_{N-7}^{L2} - 4\bar{w}_{N-6}^{L2} - \bar{w}_{N-8}^{R1} + 4\bar{w}_{N-9}^{R2} - \bar{w}_{N-8}^{R2})/12 \\
(7200\bar{w}_{N-6}^{L1} + 7200\bar{w}_{N-6}^{L2} - 29401\bar{w}_{N-5}^{L2} - 831012\bar{w}_{N-3}^{L4} - 7200\bar{w}_{N-7}^{R1} + 28800\bar{w}_{N-8}^{R2} \cdots \\
\quad - 7200\bar{w}_{N-7}^{R2})/86400 \\
\vdots \\
(13649\bar{w}_{N-1}^{L1} + 13649\bar{w}_{N-1}^{L2} + 13649\bar{w}_{N-1}^{L3} + 13649\bar{w}_{N-1}^{L4} - 14612\bar{w}_{N-2}^{R1} + 24996\bar{w}_{N-3}^{R2} \cdots \\
\quad + 519804\bar{w}_{N-4}^{R3} + 1480772\bar{w}_{N-5}^{R4})/172800 \\
0
\end{pmatrix}, \tag{89}$$

$$\mathrm{Diag}(\Lambda_3) =$$

$$\begin{pmatrix}
0\\
0\\
(12498\bar{w}_3^{L2} + 519804\bar{w}_4^{L3} + 2221158\bar{w}_5^{L4} - 45453\bar{w}_1^{R2} \cdots \\
\quad - 236408\bar{w}_1^{R3} - 659265\bar{w}_1^{R4})/172800\\
(37240\bar{w}_4^{L2} + 210508\bar{w}_4^{L3} + 466185\bar{w}_5^{L3} + 1206571\bar{w}_5^{L4} + 2246400\bar{w}_6^{L4} - 10760\bar{w}_2^{R2} \cdots \\
\quad - 572865\bar{w}_1^{R3} - 133292\bar{w}_2^{R3} - 1418579\bar{w}_1^{R4} - 453996\bar{w}_2^{R4})/172800\\
(26997\bar{w}_5^{L2} + 182392\bar{w}_5^{L3} + 475200\bar{w}_6^{L3} + 552585\bar{w}_5^{L4} + 1224000\bar{w}_6^{L4} - 41202\bar{w}_3^{R2} \cdots \\
\quad - 367596\bar{w}_2^{R3} - 226356\bar{w}_3^{R3} - 2537142\bar{w}_1^{R4} - 1009004\bar{w}_2^{R4} - 641862\bar{w}_3^{R4})/172800\\
(4800\bar{w}_6^{L2} + 31200\bar{w}_6^{L3} + 93600\bar{w}_6^{L4} - 4064\bar{w}_4^{R2} - 92577\bar{w}_3^{R3} - 320704\bar{w}_2^{R4} \cdots \\
\quad - 231079\bar{w}_3^{R4})/28800\\
(28800\bar{w}_7^{L2} - 29401\bar{w}_5^{R2} - 2493036\bar{w}_3^{R4})/172800\\
(\bar{w}_8^{L2} - \bar{w}_6^{R2})/6\\
\vdots\\
(\bar{w}_{N-6}^{L2} - \bar{w}_{N-8}^{R2})/6\\
(29401\bar{w}_{N-5}^{L2} + 2493036\bar{w}_{N-3}^{L4} - 28800\bar{w}_{N-7}^{R2})/172800\\
\vdots\\
(45453\bar{w}_{N-1}^{L2} + 236408\bar{w}_{N-1}^{L3} + 659265\bar{w}_{N-1}^{L4} - 12498\bar{w}_{N-3}^{R2} \cdots \\
\quad - 519804\bar{w}_{N-4}^{R3} - 2221158\bar{w}_{N-5}^{R4})/172800\\
0\\
0
\end{pmatrix}, \tag{90}$$

$$\mathrm{Diag}(\Lambda_4) =$$

$$\begin{pmatrix}
0\\
0\\
(173268\bar{w}_4^{L3} - 1480772\bar{w}_5^{L4} + 190955\bar{w}_1^{R3} + 1036669\bar{w}_1^{R4})/172800\\
(155395\bar{w}_5^{L3} - 895781\bar{w}_5^{L4} - 1497600\bar{w}_6^{L4} + 122532\bar{w}_2^{R3} \cdots \\
\quad + 1691428\bar{w}_1^{R4} + 763940\bar{w}_2^{R4})/172800\\
(79200\bar{w}_6^{L3} - 453600\bar{w}_6^{L4} + 92577\bar{w}_3^{R3} + 641408\bar{w}_2^{R4} + 508083\bar{w}_3^{R4})/86400\\
(69251\bar{w}_3^{R4})/7200\\
0\\
\vdots\\
0\\
-(69251\bar{w}_{N-3}^{L4})/7200\\
(92577\bar{w}_{N-3}^{L3} - 508083\bar{w}_{N-3}^{L4} - 641408\bar{w}_{N-2}^{L4} + 79200\bar{w}_{N-6}^{R3} + 453600\bar{w}_{N-6}^{R4})/86400\\
(122532\bar{w}_{N-2}^{L3} - 763940\bar{w}_{N-2}^{L4} - 1691428\bar{w}_{N-1}^{L4} + 155395\bar{w}_{N-5}^{R3} \cdots \\
\quad + 1497600\bar{w}_{N-6}^{R4} + 895781\bar{w}_{N-5}^{R4})/172800\\
(190955\bar{w}_{N-1}^{L3} - 1036669\bar{w}_{N-1}^{L4} + 173268\bar{w}_{N-4}^{R3} + 1480772\bar{w}_{N-5}^{R4})/172800\\
0\\
0
\end{pmatrix}, \tag{91}$$

$$\mathrm{Diag}(\Lambda_5) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ (370193\bar{w}_5^{L4} - 422857\bar{w}_1^{R4})/172800 \\ (5850\bar{w}_6^{L4} - 5011\bar{w}_2^{R4})/2700 \\ -(69251\bar{w}_3^{R4})/28800 \\ 0 \\ \vdots \\ 0 \\ (69251\bar{w}_{N-3}^{L4})/28800 \\ (5011\bar{w}_{N-2}^{L4} - 5850\bar{w}_{N-6}^{R4})/2700 \\ (422857\bar{w}_{N-1}^{L4} - 370193\bar{w}_{N-5}^{R4})/172800 \\ 0 \\ 0 \\ 0 \end{pmatrix}. \tag{92}$$

References

1. R. Borges, M. Carmona, B. Costa, and W. S. Don. An improved weighted essentially non-oscillatory scheme for hyperbolic conservation laws. *Journal of Computational Physics*, 227(6):3191–3211, 2008.
2. M. H. Carpenter, D. Gottlieb, and S. Abarbanel. Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Methodology and application to high-order compact schemes. *Journal of Computational Physics*, 111(2):220–236, 1994.
3. M. H. Carpenter and C. A. Kennedy. Fourth-order $2n$-storage Runge-Kutta schemes. Technical Report TM 109112, NASA, 1994.
4. M. H. Carpenter, J. Nordström, and D. Gottlieb. A stable and conservative interface treatment of arbitrary spatial accuracy. *Journal of Computational Physics*, 148(2):341–365, 1999.
5. M. H. Carpenter, J. Nordström, and D. Gottlieb. Revisiting and extending interface penalties for multi-domain summation-by-parts operators. Technical Report TM 214892, NASA, 2007.
6. M. H. Carpenter, J. Nordström, and D. Gottlieb. Revisiting and extending interface penalties for multi-domain summation-by-parts operators. *Journal of Scientific Computing*, 45(1–3):118–150, 2010.
7. J. Casper and M. H. Carpenter. Computational considerations for the simulation of shock-induced sound. *SIAM Journal on Scientific Computing*, 19(3):813–828, 1998.
8. B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor. *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations*. Springer, Berlin, 1998.
9. G. Erlebacher, M. Hussaini, and C.-W. Shu. Interaction of a shock with a longitudinal vortex. *Journal of Fluid Mechanics*, 337:129–153, 1997.
10. T. C. Fisher, M. H. Carpenter, N. K. Yamaleev, and S. H. Frankel. Boundary closures for fourth-order energy stable weighted essentially non-oscillatory finite-difference schemes. *Journal of Computational Physics*, 230(10):3727–3752, 2011.
11. B. Gustaffson. *High Order Finite Difference Methods for Time Dependent PDE*. Springer, Berlin, 2008.
12. B. Gustafsson. The convergence rate for difference approximations to mixed initial boundary value problems. *Mathematics of Computation*, 29:396–406, 1975.

13. G. Jiang and C.-W. Shu. Efficient implementation of weighted ENO schemes. *Journal of Computational Physics*, 126(1):202–228, 1996.
14. A. Kitson, R. I. McLachlan, and N. Robidoux. Skew-adjoint finite difference methods on nonuniform grids. *New Zealand Journal of Mathematics*, 32(2):139–159, 2003.
15. H.-O. Kreiss and G. Scherer. Finite element and finite difference methods for hyperbolic partial differential equations. In *Mathematical Aspects of Finite Elements in Partial Differential Equations*, pages 195–212. Academic Press, San Diego, 1974.
16. P. Lax and M. Mock. The computation of discontinuous solutions of linear hyperbolic equations. *Communications in Pure and Applied Mathematics*, 31:423–430, 1978.
17. M. Martin, E. Taylor, M. Wu, and V. Weris. A bandwidth-optimized WENO scheme for the effective direct numerical simulation of compressible turbulence. *Journal of Computational Physics*, 220(1):270–289, 2006.
18. K. Mattsson. Summation by parts operators for finite difference approximations of second-derivatives with variable coefficients. *Journal of Scientific Computing*, 29(1):1–33, 2011.
19. R. I. McLachlan and N. Robidoux. EQUADIFF 99. In *Antisymmetry, Pseudospectral Methods, and Conservative PDEs*, pages 994–999. World Scientific, Singapore, 2000.
20. J. Nordström. Conservative finite difference formulations, variable coefficients, energy estimates and artificial dissipation. *Journal of Scientific Computing*, 29(3):375–404, 2006.
21. J. Nordström and M. H. Carpenter. Boundary and interface conditions for high-order finite-difference methods applied to the Euler and Navier-Stokes equations. *Journal of Computational Physics*, 148(2):621–645, 1999.
22. J. Nordström, J. Gong, E. van der Weide, and M. Svärd. A stable and conservative high order multi-block method for the compressible Navier-stokes equation. *Journal of Computational Physics*, 228(24):9020–9035, 2009.
23. C.-W. Shu. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations: Lecture Notes in Mathematics*, volume 1697, pages 325–432. Springer, Berlin, 1998.
24. B. Strand. Summation by parts for finite difference approximations for d/dx. *Journal of Computational Physics*, 110(1):47–67, 1994.
25. M. Svärd and J. Nordström. On the order of accuracy for difference approximations of initial-boundary value problems. *Journal of Computational Physics*, 218(1):333–352, 2006.
26. N. K. Yamaleev and M. H. Carpenter. A systematic methodology for constructing high-order energy stable WENO schemes. *Journal of Computational Physics*, 228(11):4248–4272, 2009.
27. N. K. Yamaleev and M. H. Carpenter. Third-order energy stable WENO scheme. *Journal of Computational Physics*, 228(8):3025–3047, 2009.
28. S. Zhang, S. Jiang, Y. Zhang, and C. Shu. The mechanism of sound generation in the interaction between a shock wave and two counter-rotating vortices. *Physics of Fluids*, 21:076101, 2009.

A Multiscale Method Coupling Network and Continuum Models in Porous Media II—Single- and Two-Phase Flows

Jay Chu, Björn Engquist, Maša Prodanović, and Richard Tsai

Abstract We present a numerical multiscale method for coupling mass conservation laws at the continuum scale with a discrete, pore scale network model for two-phase flow in porous media. Our previously developed single-phase flow algorithm is extended to two-phase flows, for the situations in which the saturation profile go through a sharp transition from fully saturated to almost unsaturated states. Our method evaluates the continuum equation by simulations using small representative networks centering at different physical locations, and thereby computes the effective dynamics of the two phase flow at the continuum scale. On the other hand, the initial and boundary data for the network simulations are determined by the variables used in the continuum model. We present numerical results for single-phase flows with nonlinear flux-pressure dependence, as well as two-phase flows.

J. Chu (✉) · B. Engquist · R. Tsai
Department of Mathematics, The University of Texas at Austin, Austin, TX 78712, USA
e-mail: ccchu@math.utexas.edu

B. Engquist
e-mail: engquist@math.utexas.edu

R. Tsai
e-mail: ytsai@math.utexas.edu

B. Engquist · R. Tsai
ICES, The University of Texas at Austin, Austin, TX 78712, USA

M. Prodanović
Department of Petroleum and Geosystems Engineering, The University of Texas at Austin, Austin, TX 78712, USA
e-mail: masha@ices.utexas.edu

M. Prodanović
CPGE, The University of Texas at Austin, Austin, TX 78712, USA

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5_7,

1 Introduction

Modeling and computing transport in the subsurface is a difficult problem that requires good understanding of the relations among different processes at various different length and time scales, and their effective properties. Already at the pore scale, ranging from a few micrometers to millimeters, direct flow simulation in a detailed medium geometry assuming Stokes flow is extremely costly. Network modeling [6, 31, 35] provides an alternative for approximating flows at the pore scale with reduced complexity; in a network model the complicated geometry related to the medium is mapped onto a representative network of idealized pores, throats, and cracks. The fluid displacements within a physical region are then modeled as discrete events that take place in the corresponding pore-throat network. At larger length scales, one typically constructs continuum PDE models involving Darcy's law. The coefficients in these models are typically derived from grid blocks that contain sufficiently many pores so that the average fluid quantities within evolve smoothly with time. However, the specific microscopic structure of the pore space frequently plays a critical role in determining macroscopic flow properties, and often cannot be ignored. Continuum models accounting for two scales—the so-called dual porosity models [1, 2, 27]—have been constructed, and some efforts have also been made to build hybrid models [4, 39]. Balhoff et al. [4] focused on a scenario in which a pore network domain is connected to a continuum Darcy model for simulating single-phase fluid flow. In [4], the network and the continuum domain were physically disjoint except for a shared interface where information from the domains is exchanged. Another example is given in [3], and also [5], where a mortar method is adapted to include pore scale models.

In this paper, the multiscale numerical model of [14] is extended so that the saturation of two non-mixing fluids in porous media can be computed efficiently over a much larger time and length scales, using effective properties of the underlying discrete network model. We also consider a single-phase flow in networks containing edges connecting two physically remote pores. Such network configurations are adequate for describing media near a wellbore. See Fig. 6. The algorithm has the form of the heterogeneous multiscale method (HMM) [17], and couples a network model on the microscale with continuum scale over the same physical domain. The HMM [17] is a general framework for designing multiscale methods. HMM starts with an incomplete macroscopic model for macro variables on a chosen coarse grid that covers the full physical domain. The missing quantities and data in the macroscopic model are obtained by solving an accurate microscale model locally over small domains. The continuum two-phase models in this paper are conservation laws written formally in two PDEs, one for the pressure and the other for the saturation of the fluids. The macroscopic PDE for the pressure is discretized over a grid using a finite volume method. The fluxes of the macroscopic variables through the cell that is outlined by the dashed rectangle are computed by network simulations over the four small domains. See Fig. 1. The evolution for the fluid saturation at the macroscopic scale is simplified to a front propagation problem. The saturation is assumed to be essentially constant with a sharp transition along an evolving curve. The normal

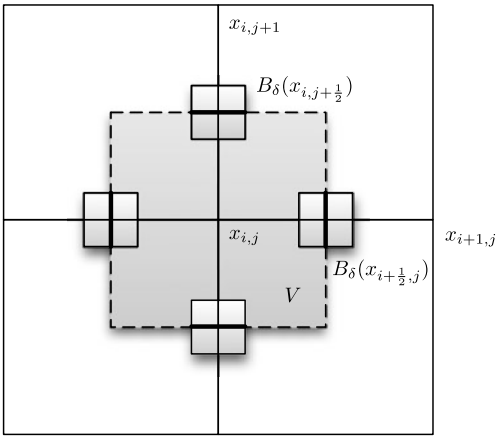


Fig. 1 Schematic of 2D coupled model. The *shaded cell with dashed boundary* is a macro grid cell and the *smaller shaded boxes* on the four sides are local network domains

velocity of this curve is evaluated by two-phase flow network simulations on local domains placed over the curve. The network simulations over the small domains require boundary conditions that are determined by the values of the macroscopic variables. See Fig. 4. Typically, Dirichlet boundary conditions are imposed on the boundaries of each small domain. In such cases, one may simply interpolate the macroscopic variables to obtain the boundary data for the small network domains.

Our algorithm shares certain similarities with other upscaling approaches for both single phase flow [10, 12, 19, 32], and two phase flow [11, 16, 18]. We refer the reader to our previous paper [14] for an extensive comparison of conceptual approaches. Our algorithm couples a given pore scale network in 2D or 3D, structured or un-structured, to an effective conservation law on continuum scale that can be posed in 1D, 2D or 3D. The pore scale network properties may depend on effective quantities from the larger scale.

2 Steady State Single-Phase Flow

In this section, we present our multiscale coupling algorithm for the pressure in the system. For simplicity of presentation, we will describe our algorithm under the assumption that the effective flow through the underlying two or three dimensional networks is essentially one dimensional and can be approximated by the effective continuum equations posed in one dimension.

2.1 Network Models

Generally speaking, predictions of macroscopic two-phase flow in a porous medium can be achieved by averaging of Navier-Stokes equations on the pore level assuming

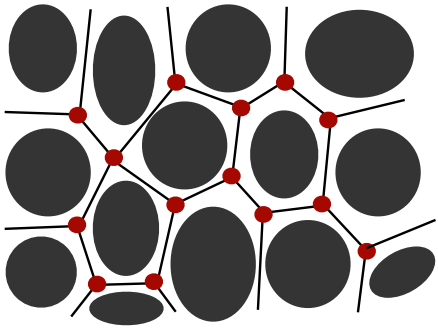


Fig. 2 An example of the relation between a block of grains and a network model. In the network model, the grains are neglected, the pores are represented by balls (*nodes*) and the throats are represented by cylindrical tubes (*segments*)

appropriate boundary conditions. This is how Darcy's equation can be derived [38]. However, obtaining a closed system of averaged equations requires the introduction of constitutive relationships between the different parameters, such as capillary pressure-saturation and relative permeability-saturation. These relationships can be approximated from pore scale simulations using either the corresponding network models or more complicated models, such as the lattice-Boltzmann method, smoothed particle hydrodynamics, and level set method, that work in exact porous medium geometry. Network flow modeling, pioneered by Fatt [22–24], retains the interconnectivity or the pathways in the original porous medium, as well as a set of microscopic properties such as sizes and geometries of the pores and tight cross-sections. The cross-sectional geometrical details are typically simplified so that the overall complexity of the model is reduced. Capillary forces and the resulting effect are built directly into the network models. Network simulations are used to study pore scale controls on permeability [8], and the relations between saturation and macroscopic parameters such as relative permeability [29, 36, 37], capillary pressure and interfacial area [30], phase distributions, as well as the influence of wettability on different relationships [15, 35]. More details on the network flow models can be found in the reviews articles by Celia et al. [9] and Blunt et al. [6, 7].

In network models, pores are simply represented as nodes and throats as links. In the simplest form, throats are approximated by cylindrical tubes and pores by spheres. The nodes and tubes are usually depicted by vertices and edges respectively. Thus a network model has a topology of a graph. However, as each pore has a physical location, we shall refer a network that models a medium in a d dimensional domain as a d dimensional network. See Fig. 2 for an illustration of a two dimensional network. We shall also consider networks that have "nonlocal" edges—this is an idealized situation of a partially cemented natural fracture [28] or a man-made structure (wellbore or a drain pipe) that connects two non-neighboring porous regions.

For convenience, we number all nodes in the domain and collect them in the set I. Furthermore, we shall denote by $I^{(0)}$ the index set containing all the indices of the nodes lying in the interior of the network domain. Let I_i denote the set consisting of all node indices j that connect to the node i by a throat. Further, p_i denotes the microscopic pressure inside pore i and c_{ij} denotes the conductance of the throat

which connects pore i and the pore j for each $j \in I_i$. The pressure flux from pore i to pore j is simply $c_{ij}(p_i - p_j)$. The law of mass conservation leads to

$$\sum_{j \in I_i} c_{ij}(p_i - p_j) = g_i, \tag{1}$$

where g_i is the sink or source in the pore i. In general, the conductance c_{ij} may be a nonlinear function,

$$c_{ij} := c(p_i, p_j), \tag{2}$$

depending on the nearby pressures p_i and p_j. However, for particular cases such as the Newtonian fluid in a cylindrical throat where gravity can be ignored, the conductance is given by a constant

$$c_{ij} = \frac{\pi r^4}{8l\mu},$$

where r is the radius and l the length of the throat, and μ is the viscosity of the fluid. System (1) should be coupled with suitable boundary conditions on the boundary nodes. The boundary conditions are typically Dirichlet, periodic or Neumann conditions.

In this section, we assume a three dimensional network, and we impose Dirichlet boundary conditions on two opposite faces (the left and the right faces) of the cubic volume and periodic boundary condition (or no flow condition) on the remaining parts of the boundary. Periodic boundary conditions can be used in regular lattice networks, as well as in irregular networks from periodic model sphere packings.

Let I_{x_0} be the index set consisting the indices of the nodes that are connected by throats that cut through the plane $x = x_0$. Then the flux through this plane is computed by summing up the fluxes through all the throats that cross this plane:

$$f = \sum_{i \in I_{x_0}} \sum_{j \in I_i} c_{ij}(p_i - p_j). \tag{3}$$

2.2 Macroscopic Model

Consider a network model over the physical domain $[x_L, x_R] \times [y_1, y_2] \times [z_1, z_2]$, with the Dirichlet condition on the boundaries at x_L and x_R, and periodic boundary condition (or no flow Neumann) on the other four faces. We assume that in macroscopic scale, the average fluid velocity $\mathbf{v}$ depends on pressure P, pressure gradient ∇P and the background geological data. The dependence of geological data is described by the location variable $\mathbf{x}$. Mass conservation implies

$$\nabla \cdot \mathbf{v}(\mathbf{x}, P, \nabla P) = G(\mathbf{x}), \tag{4}$$

where G is a source or sink term. Let $B_\delta(x)$ be the subdomain $[x-\delta/2, x+\delta/2] \times [y_1, y_2] \times [z_1, z_2]$ and $\Sigma(x;\delta)$ be the boundary surface of $B_\delta(x)$. By integrating (4) over $B_\delta(x)$ and applying the boundary conditions, we have

$$\int_{B_\delta} G\, dv = \int_{B_\delta} \nabla \cdot \mathbf{v}\, dv = \oint_{\Sigma} \mathbf{v} \cdot \mathbf{n}\, ds = F_{\Sigma_R} - F_{\Sigma_L}, \tag{5}$$

where F_{Σ_R} and F_{Σ_L} are the fluxes through boundaries at $x+\delta/2$ and $x-\delta/2$ respectively. Dividing δ on the both sides of (5) and taking the limit as δ to 0 lead to

$$\frac{d}{dx}F = \lim_{\delta \to 0} \frac{1}{\delta} \int_{B_\delta(x)} G dv =: Q(x), \quad x \in (x_L, x_R). \tag{6}$$

Hence we obtain a one dimensional macroscopic equation over $[x_L, x_R]$ with the macroscopic pressure P being the unknown which can be viewed as an average pressure of small scale pressure p on the cross section $\Sigma(x;0)$. We assume the flux F is a function of pressure P, pressure gradient P_x and location x. We shall evaluate F from simulations using the network models.

Let N be the number of partitions of $[x_L, x_R]$ and $\Delta x = (x_R - x_L)/N$, $x_l = x_L + l\Delta x$ for $l = 0, 1, \ldots, N$. Let P_l be the approximation of $P(x_l)$ and $F_{l-\frac{1}{2}}$ be the approximation of the flux F at $x_{l-\frac{1}{2}} = (x_l + x_{l-1})/2 = x_L + (l - \frac{1}{2})\Delta x$. The main goal of our multiscale method is to find $P_0, P_1, P_2, \ldots, P_{N-1}, P_N$ such that $P_0 = P_L$, $P_N = P_R$ and

$$F_{l+\frac{1}{2}} - F_{l-\frac{1}{2}} = \int_{x_{l-\frac{1}{2}}}^{x_{l+\frac{1}{2}}} Q dx =: Q_i \Delta x \quad \text{for } l = 1, 2, \ldots, N-1. \tag{7}$$

In the next subsection, we describe how to evaluate the macroscopic fluxes and how to map the macroscopic fluxes to the values of the macroscopic pressures.

2.3 The Basic Coupling Algorithm

The macroscopic flux $F_{l-\frac{1}{2}}$ is determined by the network model as follows. For each grid node x_l, we choose a representative subdomain $B_\delta(x_{l-\frac{1}{2}})$. We call the corresponding portion of our network over this subdomain the local network centered at $x_{l-\frac{1}{2}}$. The Dirichlet boundary conditions for the subdomain $B_\delta(x_{l-\frac{1}{2}})$ at $x_{l-\frac{1}{2}} \pm \delta/2$ are defined as the values of the macroscopic pressure at the corresponding locations. At the discretization level, they are approximated by linear interpolation of P_l and P_{l-1} on $[x_{l-1}, x_l]$ to define an approximation of the pressure P at $x_{l-\frac{1}{2}} \pm \delta/2$. Thus, the flux through the local network is a function depending on two macroscopic pressure values and the center of the subdomain

$$\hat{f}_{l-\frac{1}{2}} = f(x_{l-\frac{1}{2}}, P_{l-1}, P_l),$$

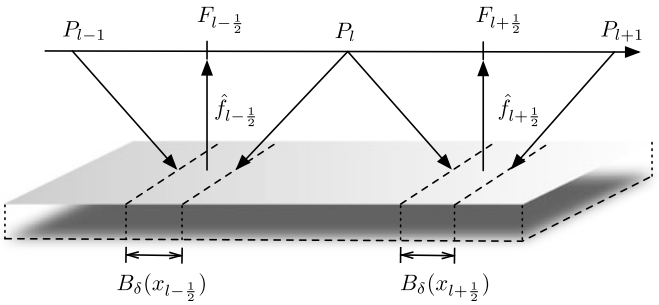


Fig. 3 Continuous (macro) scale is discretized using points x_l, $l = 0, \ldots, N$. Macro flux $F_{l-\frac{1}{2}}$ is updated using micro scale simulation (network model) on a representative region within the segment $[x_{l-1}, x_l]$ (local network domain). At the same time, the boundary conditions (in this sketch, pressure boundary conditions) required for the micro-scale model come from the macro-scale information (pressure) at end points $[x_{l-1}, x_l]$

where f is the function defined by (3) in Sect. 2.1. More precisely, the Dirichlet boundary conditions at $x_{l-\frac{1}{2}} \pm \frac{\delta}{2}$ are $P_{l-\frac{1}{2},L}$ and $P_{l-\frac{1}{2},R}$ defined by $P_{l-\frac{1}{2}} = (P_{l-1} + P_l)/2$, and

$$P_{l-\frac{1}{2},L} = P_{l-\frac{1}{2}} - D^+ P_{l-1}(\delta/2), \qquad P_{l-\frac{1}{2},R} = P_{l-\frac{1}{2}} + D^+ P_{l-1}(\delta/2),$$

where $D^+ P_{l-1} = (P_l - P_{l-1})/\Delta x$ is the standard divided centered differencing on P_{l-1}.

The macroscopic flux $F_{l-\frac{1}{2}}$ is defined as the flux, denoted by $\hat{f}_{l-\frac{1}{2}}$, through the corresponding local network:

$$F_{l-\frac{1}{2}}(P_{l-1}, P_l) = \hat{f}_{l-\frac{1}{2}}.$$

The source term $\int_{x_{l-\frac{1}{2}}}^{x_{l+\frac{1}{2}}} Q\, dx = \int_B S dv$ is obtained by summing all source terms s_i in each pore inside subdomain B. In particular, $Q_i \Delta x = \int_{x_{l-\frac{1}{2}}}^{x_{l+\frac{1}{2}}} G dx = 0$ if we assume $g_i = 0$ in the network model.

Figure 3 shows a schematic diagram of the proposed coupling. The dashed boxes shown there correspond to the representative local networks, which can be two or three dimensions. Under this setting, *the flux F can be obtained for any given pressures P_{l-1} and P_l, but the explicit expression is unknown, when the underlying network model is nonlinear*. The formal algebraic equations (7) for the macroscopic pressure P_l may be nontrivial to solve as the relation between $F_{l-\frac{1}{2}}$, P_{l-1} and P_l are not available explicitly. In particular, the Newton's method is not applicable and thus an alternate root finding scheme is required. We propose a quasi-Newton-like scheme in the next section.

2.3.1 Recovering the Pressure from Macroscopic Flux Values

We now describe our proposed method for recovering the macroscopic pressure values. In the following discussion, we assume that there is no source term in the system. As one can see from the above discussion, the difficulty to be overcome here is that no convenient analytical relation between the macroscopic flux F and the pressure P is available (or rather assumed). Our strategy is to resort to Taylor expansions, using the fact that the flux should be zero when there is no pressure gradient; i.e.

$$F(x, P, P_x) = f(x, P, P_x) = 0, \quad \text{whenever } P_x = 0,$$

and thus

$$F(x, P, P_x) = F_{P_x}(x, P, \xi) P_x, \tag{8}$$

where F_{P_x} refers to the partial derivative of F with respect to the third variable and ξ is an intermediate value, which depends on P and x, between 0 and P_x.

At the discrete level, we want to solve the following equations for P_l:

$$F\big(x_{l+\frac{1}{2}}, P_{l+\frac{1}{2}}, D^+ P_l\big) = \hat{f}_{l+\frac{1}{2}}(P_l, P_{l+1}), \tag{9}$$

$$D^- F\big(x_{l+\frac{1}{2}}, P_{l+\frac{1}{2}}, D^+ P_l\big) = Q_l \Delta x, \quad l = 1, 2, \ldots, N-1, \tag{10}$$

with the boundary condition $P_0 = P_L$, $P_N = P_R$. See Fig. 3 for a diagram. Therefore, we use

$$\begin{aligned} F\big(x_{l+\frac{1}{2}}, P_{l+\frac{1}{2}}, D^+ P_l\big) &= F_{P_x}(x_{l+\frac{1}{2}}, P_{l+\frac{1}{2}}, \xi) \\ &\approx \hat{f}_{l+\frac{1}{2}}(P_l, P_{l+1}) / D^+ P_l =: -K(P_l, P_{l+1}). \end{aligned} \tag{11}$$

We propose to solve the above coupled equations by iterations:

$$-D^-\big(K\big(P_l^{(n)}, P_{l+1}^{(n)}\big) D^+ P_l^{(n+1)}\big) = Q_l \Delta x. \tag{12}$$

This iterative scheme can be explicitly written as

$$\mathbf{P}^{(n+1)} = \left(\frac{1}{\Delta x^2}\mathbf{K}^{(n)}\right)^{-1} \mathbf{h}^{(n)} = \mathbf{P}^{(n)} - \left(\frac{1}{\Delta x^2}\mathbf{K}^{(n)}\right)^{-1} \mathbf{G}\big(\mathbf{P}^{(n)}\big), \tag{13}$$

where

$$K_{l-\frac{1}{2}}^{(n)} = F_{l-\frac{1}{2}}^{(n)} \cdot \frac{\delta}{(P_{l-\frac{1}{2},L}^{(n)} - P_{l-\frac{1}{2},R}^{(n)})} = -\frac{F_{l-\frac{1}{2}}^{(n)}}{D^+ P_{l-1}^{(n)}}, \tag{14}$$

$$\mathbf{K}^{(n)} = \begin{pmatrix} K^{(n)}_{\frac{1}{2}} + K^{(n)}_{\frac{3}{2}} & -K^{(n)}_{\frac{3}{2}} & 0 & \cdots & 0 \\ -K^{(n)}_{\frac{3}{2}} & K^{(n)}_{\frac{3}{2}} + K^{(n)}_{\frac{5}{2}} & -K^{(n)}_{\frac{5}{2}} & \ddots & \vdots \\ 0 & -K^{(n)}_{\frac{5}{2}} & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & -K^{(n)}_{N-\frac{3}{2}} \\ 0 & \cdots & 0 & -K^{(n)}_{N-\frac{3}{2}} & K^{(n)}_{N-\frac{3}{2}} + K^{(n)}_{N-\frac{1}{2}} \end{pmatrix}, \tag{15}$$

$$\mathbf{h}^{(n)} = \left[Q_1 + K^{(n)}_{1-\frac{1}{2}} P_L/\Delta x^2, Q_2, \ldots, Q_{N-2}, Q_{N-1} + K^{(n)}_{N-\frac{1}{2}} P_R/\Delta x^2\right]^T, \tag{16}$$

and

$$\mathbf{G}(\mathbf{P}) = \left[D^+ F_{1-\frac{1}{2}} - Q_1, D^+ F_{2-\frac{1}{2}} - Q_2, \ldots, D^+ F_{N-1-\frac{1}{2}} - Q_{N-1}\right]^T. \tag{17}$$

We refer the readers to [14] for more detailed discussion about this iterative scheme and its convergence. The treatment for the cases involving non-zero source terms is also presented in [14].

3 Two-Phase Flows

In this section, we generalize the algorithm described in the previous section to coupling a continuum model for two phase flow with dynamic two-phase flow network models. The dynamic network model follows the work of Joekar-Niasar, Hassanizadeh and Dahle [26]. For completeness, we summarize their algorithm in the following section.

3.1 Two-Phase Dynamic Pore-Network Modeling

In the model of [26], each pore is filled with one or two fluids (wetting and non-wetting), and each fluid has its own pressure, denoted by p^w and p^n respectively. The local capillary pressure p^c_i for pore body i is defined as the difference of non-wetting and wetting pressures. The capillary pressure is assumed to be determined by local saturation s^w in the pore body i only. That is,

$$p^c_i = p^n_i - p^w_i = p^c_i\left(s^w\right). \tag{18}$$

A flux f^α_{ij} for phase α fluid in throat ij is simply given by

$$f^\alpha_{ij} = c^\alpha_{ij}\left(p^\alpha_i - p^\alpha_j\right), \quad \alpha = w, n, \tag{19}$$

where c_{ij}^{α} is the conductance of the throat for phase α fluid. Since the flow is incompressible, the mass conservation law implies the total mass is conserved:

$$\sum_{j\in I_i} f_{ij}^{w} + f_{ij}^{n} = \sum_{j\in I_i} c_{ij}^{w}\left(p_i^{w} - p_j^{w}\right) + c_{ij}^{n}\left(p_i^{n} - p_j^{n}\right) = g_i, \tag{20}$$

where g_i describes either the sink or the source in pore i. For simplicity, we assume $g_i = 0$ for all pores. To reduce the numbers of unknowns in (18) and (20), the pressure equation (20) is reformulated in terms of total pressure $\bar{p}_i = s_i^{w} p_i^{w} + s_i^{n} p_i^{n}$:

$$\begin{aligned}&\sum_{j\in I_i}\left(c_{ij}^{w} + c_{ij}^{n}\right)\left(\bar{p}_i - \bar{p}_j\right)\\ &\quad = -\sum_{j\in I_i}\left[\left(c_{ij}^{n} s_i^{w} - c_{ij}^{w}\left(1 - s_i^{w}\right)\right) + \left(c_{ij}^{w}\left(1 - s_j^{w}\right) - c_{ij}^{n} s_j^{w}\right) p_j^{c}\right].\end{aligned} \tag{21}$$

Notice that for given conductance c^{n} and c^{w}, only boundary condition $\bar{p}$ and saturation s are involved to solve (21) since p^{c} depends on s also. Once $\bar{p}$ is solved, $p^{w} = \bar{p} - s^{n} p^{c}$, $p^{n} = \bar{p} + s^{w} p^{c}$ and fluxes f_{ij}^{n} and f_{ij}^{w} in each throat can be calculated.

For saturation, a volume balance for each fluid gives

$$V_i \frac{\Delta s_i^{\alpha}}{\delta t} = -\sum_{j\in I_i} f_{ij}^{\alpha}, \tag{22}$$

where V_i is the volume of the pore i. Equations (18), (20) and (22) form a complete set of governing equations for two-phase dynamic pore-network modeling. Typically the boundary conditions of pressure and saturation are Dirichlet on one pair of opposite faces (or sides for 2D), and periodic on the rest of faces (or sides). In our later simulation, the local network model is under this setting and the Dirichlet boundary conditions are determined by the nearby coarse scale pressure and saturation.

In the following, we summarize how to evolve the saturation, as proposed in [26], using equations (18), (20), and (22). We start with the given initial saturation data and boundary conditions are given for both pressure and saturation.

Step 1. Compute the local capillary pressure. The local capillary pressure p_i^{c} is a function depending on the wetting phase saturation s_i^{w} and the interfacial tension σ^{nw}. Detail of the function is derived from the shape of the pore body. An example is presented in Experiment 3 discussed in Sect. 4.

Step 2. Determine if a throat is invaded by the non-wetting phase. A throat is invaded when the capillary pressure in a neighboring pore body is larger than the entry capillary pressure (a critical value) of the throat α_{ij}. In this case, we assign the capillary pressure in the throat p_{ij}^{c} to be equal to the capillary pressure of the upstream pore body. That is, if $\alpha_{ij} < p_i^{c}$ (or $\alpha_{ij} < p_i^{c}$), then the throat ij is invaded and $p_{ij}^{c} = p_i^{c}$ (or $p_{ij}^{c} = p_j^{c}$).

Step 3. Calculate the conductances of throats. There are two cases during the simulation:

- Case 1 of Step 3: The throat is not invaded by the non-wetting phase. Then the conductances are obtained by

$$c_{ij}^w = \frac{\pi}{8\mu^w l_{ij}} \left(\sqrt{\frac{4}{\pi}} r_{ij} \right)^4 \quad \text{and} \quad c_{ij}^n = 0,$$

where r_{ij} and l_{ij} are the inscribed radius and length of the throat ij respectively, and μ^w is the viscosity of the wetting phase.

- Case 2 of Step 3: The throat is invaded by the non-wetting phase. Then the conductances of each phase are given by

$$c_{ij}^w = \frac{4-\pi}{\beta \mu^w l_{ij}} \left(r_{ij}^c \right)^4 \quad \text{and} \quad c_{ij}^n = \frac{\pi}{8\mu^n l_{ij}} \left(r_{ij}^{eff} \right)^4,$$

where μ^n is the viscosity of the non-wetting phase and

$$r_{ij}^c = \frac{\sigma^{nw}}{p_{ij}^c}, \quad \text{and} \quad r_{ij}^{eff} = \frac{1}{2} \left(\sqrt{\frac{r_{ij}^2 - (4-\pi)(r_{ij}^c)^2}{\pi}} + r_{ij} \right).$$

Here β is a resistance factor that depends on the geometry of the throats (see [40]).

Step 4. Solve pressure equations (21) with the given Dirichlet boundary conditions to get fluxes f_{ij}^n and f_{ij}^w in each throat.

Step 5. Update saturation s_i^w. Once flux f_{ij}^w is obtained, saturation of next time step can be updated by discretizing equation (22) explicitly, for example, with the Euler scheme. However, the explicit saturation update scheme may be unstable [34]. Therefore, Joekar-Niasar et al. proposed a semi-implicit scheme to overcome this difficulty [26]. Their scheme uses total flux $f_{ij}^{tot} = f_{ij}^w + f_{ij}^n$ and total conductance $c_{ij}^{tot} = c_{ij}^w + c_{ij}^n$ as unknowns. The non-wetting flux f_{ij}^n is related to the total flux by a formula analogous to the fractional-flow equation:

$$f_{ij}^n = \frac{c_{ij}^n}{c_{ij}^{tot}} f_{ij}^{tot} + \frac{c_{ij}^n c_{ij}^w}{c_{ij}^{tot}} \left(p_i^c - p_j^c \right). \tag{23}$$

The key in their scheme is to approximate $p_i^c - p_j^c$ by the derivatives with respect to s^w:

$$p_i^c - p_j^c = \frac{\partial p_{ij}^c}{\partial s_{ij}^w} \left(s_i^w - s_j^w \right). \tag{24}$$

Similar to the evaluation of p_{ij}^c in Step 2, $\partial p_{ij}^c / \partial s_{ij}^w$ is calculated from the upstream pore body. By substituting (24) into (23), a semi-implicit discretization of (22) is given by

$$V_i \frac{(s_i^w)^{k+1} - (s_i^w)^k)}{\delta t} - \sum_{j \in I_i} \left(\frac{c_{ij}^n}{c_{ij}^{tot}} f_{ij}^{tot} + \frac{c_{ij}^n c_{ij}^w}{c_{ij}^{tot}} \frac{\partial p_{ij}^c}{\partial s_{ij}^w} \left(\left(s_i^w \right)^{k+1} - \left(s_j^w \right)^{k+1} \right) \right) = 0. \tag{25}$$

After updating saturation, go back to Step 1 and repeat the process until the saturation is unchanged.

3.2 Two-Phase Flow Continuum Model

The two-phase flow setting under consideration involves a wetting and a non-wetting fluids. We use the saturation (volume fraction) S^w to denote the ratio between the volume of wetting fluid and the total volume of pore space. The saturation of non-wetting phase S^n is defined analogously. Obviously the saturations satisfy $S^w + S^n = 1$. Due to the curvature and surface tension of the interface of the two phases, the pressure in the non-wetting fluid P^n is higher than that in wetting fluid P^w; the pressure difference is determined by the capillary pressure $P^c = P^n - P^w$. The balance of volume for phase saturation results in the following evolution equations:

$$\phi \frac{\partial S^\alpha}{\partial t} + \nabla \cdot \mathbf{v}^\alpha = 0, \quad \alpha = w, n, \tag{26}$$

where ϕ is the porosity, which is the ratio of the volume of all the pores to the total volume of the medium, and $\mathbf{v}^\alpha$ is the average velocity for phase α.

This model can be simplified by the introduction of the so-called global pressure P and the total fluid velocity $\mathbf{v} = \mathbf{v}^w + \mathbf{v}^n$; see [13]. In the simplified model, the mass conservation equation is written as

$$\nabla \cdot \mathbf{v}\big(\mathbf{x}, P, \nabla P, S^w\big) = G(\mathbf{x}), \tag{27}$$

where G is a source or sink term. The saturation equations (26) are reduced to a single one involving the evolution of S^w:

$$\phi \frac{\partial S^w}{\partial t} + \nabla \cdot \big(f\big(S^w\big)\mathbf{v}\big) = 0, \tag{28}$$

where f is the fractional flow of wetting fluid. In classical models, the velocity is assumed to satisfy Darcy's law: $\mathbf{v} = -\kappa(\mathbf{x}, S^w)\nabla P$. The positive definite tensor function $\kappa(\mathbf{x}, S^w)$ may depend on saturation S^w. A common form of $\kappa(\mathbf{x}, S^w)$ is $\kappa(\mathbf{x}, S^w) = \lambda(S^w)\kappa$, where κ is the absolute permeability and $\lambda(S^w)$ is mobility. By Darcy's law, equation (27) becomes an elliptic partial differential equation:

$$-\nabla \cdot \big(\kappa\big(\mathbf{x}, S^w\big)\nabla P\big) = G(\mathbf{x}). \tag{29}$$

Equations (29) and (28) are called the pressure and saturation equations respectively. *In our algorithm described in this paper, the coarse pressure is simply described by global pressure P and we do not covert P to single-phase pressure P^w or P^n. The fluxes are directly estimated from local network simulation without using fractional flow information.* Therefore the details about formulas of global pressure P and the fractional flow f are omitted here and we refer interested readers to [13] for more information.

The capillary force is typically either neglected in the equations for pressure and saturation, or is embedded as a factor in the mobility λ and the fractional flow function f. In this paper, we assume that the macroscopic variables satisfy equations (27) and (28), and that their dependence on the capillary force is evaluated through simulations on the chosen local network models.

3.3 Multiscale Coupling of Two-Phase Flow Models

In this section, we explain how to couple the network modeling described in the previous section with the following continuum conservation laws:

$$\nabla \cdot \mathbf{v}\big(\mathbf{x}, P, \nabla P, S^w\big) = 0, \tag{30}$$

$$\phi \frac{\partial S^w}{\partial t} + \nabla \cdot \mathbf{v}^w = 0. \tag{31}$$

Here S^w is the average saturation for wetting phase, P is the average pressure and ϕ is the porosity, and we assume the source term $G = 0$ for simplicity. Recall that $\mathbf{v}^w$ is the velocity for wetting phase and total velocity $\mathbf{v} = \mathbf{v}^w + \mathbf{v}^n$.

We adapt the implicit pressure and explicit saturation (IMPES) approach: First we fix saturation S^w and solve (30) to get updated pressure P. Below we describe details of each step separately in two dimensional problems, but the method can be easily generalized to three-dimensional problems.

3.3.1 Solution of the Pressure Equation

We evaluate the flux for α phase fluid by $F^\alpha = \int_\Sigma \mathbf{v}^\alpha \cdot \mathbf{n}\, ds$ through suitable surfaces Σ for different profiles of P, and total flux is given by $F = F^w + F^n$. On the other hand, the macro-quantities (macroscopic pressure and saturation) determine the boundary conditions for the local network simulations and value of saturation that are used to evaluate the total flux. As in the single-phase case, the coupled system is solved by iterations, starting with the initial macroscopic data for the saturation, and some initial guess for the macroscopic pressure. Local network simulations are performed to evaluate the macroscopic fluxes. The macroscopic saturation and pressure are then updated.

We use a finite volume discretization to solve the PDEs (30) on a rectangular domain. Divide the domain into $N_1 \times N_2$ coarse blocks. On each coarse grid, we assign average pressure $P_{i,j}$ and saturation $S^w_{i,j}$. Let $\mathbf{x}_o$ be the center of a block, and V be the corresponding control volume. See Fig. 1 for an illustration. Hence, (30) implies that

$$\oint_{\partial V} \mathbf{v} \cdot \mathbf{n}\, ds = 0. \tag{32}$$

Let F_N, F_S, F_W, and F_E denote the total fluxes through the four edges of V. Equation (32) implies

$$F_N + F_S + F_W + F_E = 0. \tag{33}$$

The total flux F across each side of V is evaluated by $\hat{f}$ coming from local network simulations on a $\delta \times \delta$ size sampling domain B_δ with boundary condition from given macroscopic pressure P and saturation in each pore from downscaling macro saturation S^w. We describe in detail the case of computing F_E below. The fluxes through the other edges of V can be easily computed analogously. We define

$$F_E(\mathbf{x}_{i+\frac{1}{2},j}) = \mathbf{v}_{i+\frac{1}{2},j} \cdot \mathbf{n}_x \ \Delta y = \hat{f}^{(x)}(P_{i,j}, P_{i+1,j})\Delta y/\delta,$$

where $\hat{f}^{(x)}$ is the total flux through the local network over $B_\delta(\mathbf{x}_{i+\frac{1}{2},j})$ in x-direction. The dependence of $P_{i,j}$ and $P_{i+1,j}$ comes from the interaction of boundary conditions. More precisely, $\hat{f}^{(x)}$ is evaluated from a local network simulation according to Steps 1, 2, 3, 4 described in Sect. 3.1. The saturation s_i used in the local network is obtained from interpolating coarse saturation S^w. A particular downscaling method of coarse saturation S is presented in Sect. 3.3.2. Combining with boundary condition from interpolation of coarse pressure, total flux $f = f^w + f^n$ can be calculated by solving $\bar{p}$ in Step 4. Under this setting, the flux is a function of macroscopic pressure, and we look for macroscopic pressure $P_{i,j}$ such that the corresponding flux satisfies (33). For simplicity, we assume the upscaled conductance is isotropic. The boundary conditions for local network simulation is setup as Dirichlet boundary condition in x-direction and periodic boundary conditions in y-direction. The Dirichlet boundary condition is computed by linear interpolation using $P_{i,j}$ and $P_{i+1,j}$. For anisotropic upscaled conductance, we can apply more complicated boundary condition that is discussed in [14].

As in the one dimensional case, an explicit algebraic formula for the macroscopic flux F as a function of pressure and pressure gradient is not readily available. Moreover, due to capillary pressure inside the network, $\hat{f}^{(x)}$ can be nonzero even if the boundary conditions are the same on both sides. However, from Taylor expansion, we have

$$\begin{aligned}\hat{f}^{(x)}(P_{i,j}, P_{i+1,j}) &= \hat{f}^{(x)}\big(P_{i,j}, P_{i,j} + \Delta^+ P_{i,j}\big) \\ &= \hat{f}^{(x)}(P_{i,j}, P_{i,j}) + \partial_2 \hat{f}^{(x)}(P_{i,j}, \xi)\Delta^+ P_{i,j},\end{aligned}$$

where $\Delta^+ P_{i,j} = P_{i+1,j} - P_{i,j}$, ∂_2 is the partial derivative operator with respect to second variable and ξ is an intermediate value of $P_{i,j}$ and $P_{i+1,j}$. In the macroscopic scale, $\hat{f}^{(x)}(P_{i,j}, P_{i,j})$ is usually small and can be neglected. Hence

$$\hat{f}^{(x)}(P_{i,j}, P_{i+1,j}) \simeq \partial_2 \hat{f}^{(x)}(P_{i,j}, \xi)\Delta^+ P_{i,j}.$$

On the other hand, the macro flux F_E defined in Sect. 3.3.1 can be written as

$$F_E := \hat{f}^{(x)}(P_{i,j}, P_{i+1,j})\frac{\Delta y}{\delta} = \big(K_{i+\frac{1}{2},j} D_+^x P_{i,j}\big)\Delta y, \tag{34}$$

where $K_{i+\frac{1}{2},j}$ is given by

$$K_{i+\frac{1}{2},j} = \hat{f}^{(x)}(P_{i,j}, P_{i+1,j})/\big(\delta\ D_+^x P_{i,j}\big), \tag{35}$$

and $D_+^x P_{i,j}$ is the forward difference. A simple calculation can show that $K_{i+\frac{1}{2},j} \simeq \partial_2 \hat{f}^{(x)}(P_{i,j}, \xi)\Delta x/\delta$.

Now we are ready to describe our macro-micro iterations. For a given macroscopic pressure $P_{i,j}^{(n)}$, we compute the coefficients $K_{i\pm\frac{1}{2},j}^{(n)}$ and $K_{i,j\pm\frac{1}{2}}^{(n)}$ as in (35). The updated macroscopic pressure $P_{i,j}^{(n+1)}$ is obtained by solving the sparse linear system

$$F_N^{(n)} + F_S^{(n)} + F_W^{(n)} + F_E^{(n)} = 0,$$

where

$$F_E^{(n)} = \big(K_{i+\frac{1}{2},j}^{(n)} D_+^x P_{i,j}^{(n+1)}\big)\Delta y,$$

$$F_N^{(n)} = \big(K_{i,j+\frac{1}{2}}^{(n)} D_+^y P_{i,j}^{(n+1)}\big)\Delta x.$$

Under certain assumptions, the iterations converge to a solution of (33). See [14] for more detail.

3.3.2 Evolution of Fluid Saturation

We consider a simplified but commonly occurring setting of multiphase flow in porous media for drainage process. In a drainage process, one injects a non-wetting fluid into a bulk volume that is initially filled with a wetting fluid. The initial saturation is equal to 1 for all pores except ones on the injecting boundary. The boundary condition is simply $s = 1$ on the injecting side and $s = 0$ on the opposite side. In the evolution of saturation, a sharp transition layer is usually formed through which the saturation goes from 0 to 1. We refer this transition layer as a front.

In this setting, the macro saturation is simply characterized by an evolving curve $\Gamma(t)$ that represents the interface of a sharp transition from $S = 1$ to $S = S_0 \simeq 0$. For a given front $\Gamma(t)$, we first downscale the macro saturation to pore scale saturation by the following procedure: the saturation at a pore is set to 1 before $\Gamma(t)$ arrives there, and to the minimum saturation value after $\Gamma(t)$ passes through it. Physically due to the shape of pore bodies, it is impossible to completely displace the wetting fluid by non-wetting fluid in each pore. Thus each pore body has a minimum saturation $s_{i,min}^w$. This minimum saturation depends on the geometry of pore bodies. An example of defining the minimum saturation is given in Experiment 3 described in Sect. 4. Similarly, a throat is invaded or not is determined by its location. If both end pores of the throat are behind $\Gamma(t)$, then the throat is marked as invaded. Otherwise, the throat is not invaded. The downscaled information is used in the local network to calculate macroscopic pressures P as discussed in Sect. 3.3.1.

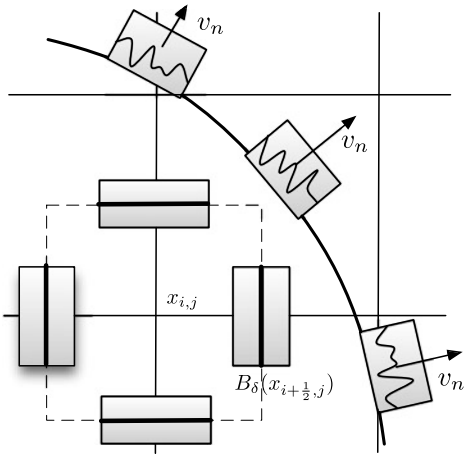


Fig. 4 Microscale network simulations of the two-phase model are performed in the *shaded boxes*. The *boxes* B_δ at the edges of the macro cells are used in the pressure calculation and the boxes along the front determine the macroscale front velocity

The evolution of $\Gamma(t)$ is approximated by a front tracking strategy that involves advancing a set of marker points placed on $\Gamma(t)$. At each time t_n, the normal speed of $\Gamma(t)$ at the marker point is evaluated from a two-phase network simulation performed on a local network whose size is $\delta \times \delta$. The orientation of the local network coincides with the normal direction of $\Gamma(t)$ at the marker point. For each marker point, the local network domain is defined with one of its faces tangent to $\Gamma(t)$ and the marker point at the face's center. Thus, we have a multiscale front propagation problem, see Fig. 4 for an illustration. Similar approaches have been used to study combustion [21] as well as epitaxial crystal growth [20].

The detailed setup for each local network simulation is described in the following. At t_n, the initial value for the saturation is set to be 1 over the interior of the local network domain. The saturation value on the face that tangents $\Gamma(t_n)$ is set to be the minimum saturation value, and 1 on the opposite face. Periodic boundary condition is imposed on the other faces of the network that are normal to $\Gamma(t_n)$. The Dirichlet boundary conditions for the microscopic pressure are obtained from interpolating the macroscopic pressure P. The local network simulation solving (21) and (22) is performed for a duration $\tau = k\delta t$ so that significant change in the saturation profile in the local network can be observed. The normal speed of $\Gamma(t_n)$ at the marker point is then calculated by

$$\mathbf{v_n} := \frac{|\sum_i V_i (s_i^{new} - s_i^{old})|\,\delta}{\tau \sum_i V_i},$$

where V_i, s_i^{old} and s_i^{new} are the volume, the initial saturation and saturation at τ of pore i respectively. In Fig. 5, we present three snapshots of saturation in a local simulation used in Experiment 3. We evolve $\Gamma(t)$ with a time step size Δt bigger than τ by moving the marker points in the normal directions with distance equal to $\mathbf{v_n}\Delta t$.

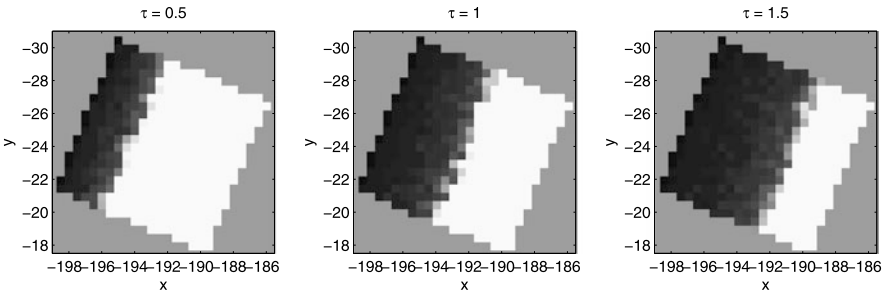


Fig. 5 Snapshot of saturation of local network simulation by using different time duration $\tau = 0.5, 1, 1.5$. Change of saturation in the local network is used to calculate speed of the front at marker points. The *gray region* is not in the local network simulation

4 Simulations

In this section, we present simulation results for three different model problems. The purpose of these examples is to showcase a proof of concept of the proposed multiscale algorithm and its applications to what conventional methods cannot compute. We first compare the results computed from full network simulations and the proposed multiscale simulations for the steady state single-phase flow with a nonlinear network flux. This example is designed to demonstrate the convergence of our multiscale coupling algorithm for nonlinear pressure equation. In the second experiment, we presented a very particular setting which is not uncommon in porous media containing either bore holes or fractures that are cemented naturally except at the tips. In such settings, the underlying network contain a highly conducting throat that connects only two widely separated physical locations. In the third example, we represent our result of multiphase simulation using the proposed multiscale front tracking algorithm. We note that with the dynamic network model, it is virtually impossible to perform direct simulation over the macroscopic spatial and temporal domains used in the last example.

Experiment 1 (Quadratic flux for high velocity flows) The flux f_{ij} in the network model is given by

$$\frac{f_{ij}}{c_{ij}} + \beta f_{ij}^2 = -(p_i - p_j).$$

The formula is derived from the Forchheimer equation:

$$-\frac{\mathrm{d}p}{\mathrm{d}x} = \frac{\mu}{\mathbf{K}} \cdot v + \rho\beta v^2,$$

where p is the pressure, v is the flux velocity, $\mathbf{K}$ is the permeability and μ is the viscosity, ρ is the fluid density and β is the non-Darcy coefficient of the porous medium. The Forchheimer equation is the standard equation for describing high-

Table 1 The average errors of Experiment 1 from 1000 realizations in pressure and flux

	e_p			e_f		
	$\delta = 10$	$\delta = 15$	$\delta = 20$	$\delta = 10$	$\delta = 15$	$\delta = 20$
$N = 5$	0.0173	0.0139	0.0114	0.0366	0.0240	0.0193
$N = 10$	0.0127	0.0102	0.0084	0.0322	0.0213	0.0169
$N = 20$	0.0094	0.0071	0.0058	0.0297	0.0193	0.0141
$N = 30$	0.0074	0.0054	0.0041	0.0305	0.0201	0.0150

velocity flow in petroleum engineering [25, 33]. In our simulation, by solving the quadratic equation, we used the following formula:

$$f_{ij} = \frac{-1 + \sqrt{1 - 4\beta c_{ij}^2 (p_i - p_j)}}{2\beta c_{ij}} \simeq -\big(c_{ij} + \beta c_{ij}^3 |p_i - p_j|\big)(p_i - p_j).$$

The conductance in this case is $(c_{ij} + \beta c_{ij}^2 |p_i - p_j|)$ and depends on nearby pressures p_i and p_j. The parameter β is chosen to be 10^{12} on purpose in order to amplify nonlinear effects in our simulations.

The testing full network model has 1001×21 nodes arranged in a $[0, 1] \times [0, 0.02]$ rectangle domain. Each node is connected by 6 nearby nodes and the length of throats is 0.001 unit in horizontal and vertical direction, and $\sqrt{2}/1000$ unit in diagonal direction. The radii of the throats are randomly generated from the uniform distribution $[(1 - \lambda)r_0, (1 + \lambda)r_0]$ and the conductance c is determined by

$$c_{ij} = \frac{\pi r^4}{8l\mu}.$$

We choose $r_0 = 0.01$, $\lambda = 0.5$, and $\mu = 1$. The resulting conductances range from 10^{-18} to 10^{-7}. We apply Dirichlet boundary condition in x direction: $p = 100$ on the left hand side and $p = 0$ on the right hand side, and periodic boundary condition in y direction. In the simulations using the proposed multiscale algorithms, we divide the domain into N blocks, each of the dimension $\delta \times 0.02$, so that the center of each block corresponds to the node $x_{\delta + \frac{l}{2}}$ described in Sect. 2.3. At the microscopic level, we experimented with a few local networks with different sizes.

We fix $\delta = 0.01$, 0.015, 0.02 and set $N = 5$, 10, 20, 30 to test the convergence of the proposed algorithm. We compare the flux F_D and the pressure P_D computed from direct full simulation on 1001×21 nodes with the flux F_H and the pressure P_H computed by the proposed multiscale algorithm. The pressure P_D is the average value of fine scale pressure on each y-direction section. The relative errors of flux e_F and of pressure e_P are defined by

$$e_F = \frac{|F_H - F_D|}{|F_D|} \quad \text{and} \quad e_P = \frac{\|P_H - P_D\|_\infty}{\|P_D\|_\infty},$$

where $\|\cdot\|_\infty$ is the supremum norm of vectors.

The average error from 1000 realizations is given in Table 1.

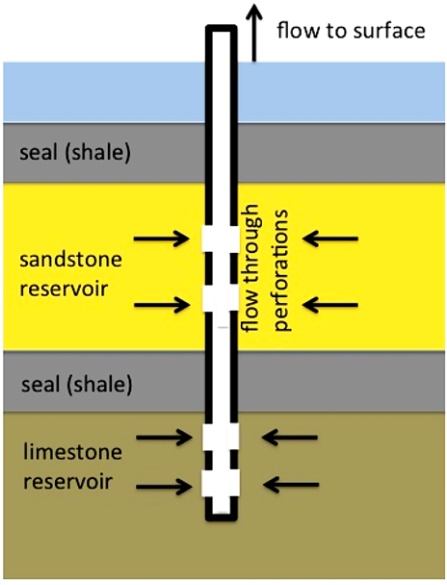


Fig. 6 Schematic of a wellbore drilled for exploration of hydrocarbons. The oil flows from the reservoirs through perforations on the wellbore casing: the perforations are the only connection to the neighboring porous formation

Experiment 2 (Nonlocal connections in the network model) Figure 6 depicts a realistic scenario in which edges in a network connecting two physically remote pores can be found. In this experiment, we report simulation results in a simplified setting. We first set up a regular network as in Experiment 1, except that here we use the linear flux $f_{ij} = -c_{ij}(p_i - p_j)$. The boundary condition is $p = 10$ on both left hand side and $p = 0$ on the right sides and periodic in y direction.

We then create a nonlocal throat connecting pore A at $(0.1, 0.01)$ and pore B at $(0.9, 0.01)$ with conductance c_{nl}. When c_{nl} is relatively larger than other conductances in the network, the pressure values of pore A and pore B are close even their physical positions are far away. The standard coupling method discussed earlier can not capture such behavior. However, the method is modified as following.

In addition to the original macroscopic pressure P_i, we introduce two pressures Q_1 and Q_2 on macro-scale in order to approximate micro pressure of pores A and B. We divide the network into N coarse blocks. Suppose pore A is in the first block and pore B is in the last block. The flux in the nonlocal throat is denoted by $G_0 = -(Q_2 - Q_1)c_{nl}$ and the fluxes on the left boundary and right boundary of the first block is denoted by F_0 and $F_{\frac{1}{2}}$ respectively. Then F_0 and $F_{\frac{1}{2}}$ are functions of P_0, P_1 and Q_1, and we have $F_0 = G_0 + F_{\frac{1}{2}}$. Similarly, we denote $F_{N-\frac{1}{2}}$ and F_N for fluxes on the left boundary and right boundary of the last block. They are functions of P_{N-1}, P_N and Q_2, and we have $F_0 = G_0 + F_{\frac{1}{2}}$. Flux $F_{l-\frac{1}{2}}$ in other blocks is defined as the same as before. Then we look for the macroscopic pressure P_l and Q_1, Q_2 to satisfy mass conservation of flux: $F_0 = G_0 + F_{\frac{1}{2}}$, $F_{l-\frac{1}{2}} = F_{l+\frac{1}{2}}$ for $l = 1, \ldots, N-1$ and $F_{N-\frac{1}{2}} + G_0 = F_N$. See Fig. 7 for illustration.

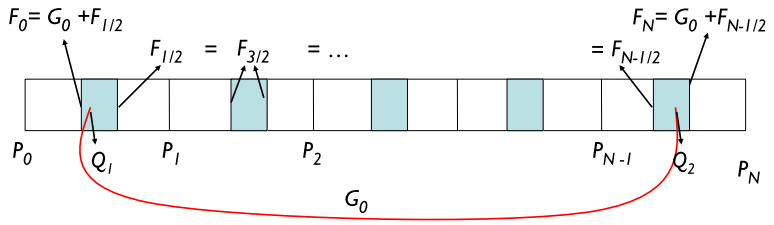


Fig. 7 Discretization of the network model with a nonlocal throat. P_i and Q_1, Q_2 denote the macroscopic pressure and F_0, F_N, $F_{i+\frac{1}{2}}$, G_0 denote the macro flux

Table 2 The errors of Experiment 2 in pressure and flux

	$\delta = 0.04$	$\delta = 0.08$	$\delta = 0.12$	$\delta = 0.16$	$\delta = 0.2$
e_p	0.0937	0.0508	0.0228	0.0110	0.000014
e_f	0.2463	0.1180	0.0608	0.0138	0.000006

In the simulations, $c_{nl} = 10^{-5}$ and $N = 5$ coarse blocks are used. The sampling size δ is chosen to be 0.04, 0.08, 0.12, 0.16, 0.2. The error of pressure and flux is given in Table 2.

We observe that the pressure error is reasonably small for all choices of δ. The flux error is large when sampling size is not wide enough, but the error decays when enlarging the sampling size. If we sample all pores in the network ($\delta = 0.2$), we obtain very accurate approximations for pressure and flux of full network simulation. This example demonstrates that our method can be applied to an unstructured network which is very different from discretizations of partial differential equations.

Experiment 3 (Two dimensional problem with two-phase flows) In the following simulation, we consider a two-dimensional network with 1001×1001 pores whose physical domain is $[-250, 250]^2$. The network structure is a regular lattice and each pore connects to four adjacent pores. The spacing between layers of the network in x- and y-directions is 0.5.

While this is a two-dimensional problem, we nevertheless assume 3D shapes for pores and throats so that wetting layers can be accommodated. Pore bodies are of cubic shape and throats have square cross-sections.

The radius of inscribed sphere in pore body i is denoted by R_i, and R_i's are generated from a truncated log-normal distribution with no spatial correlation. The density function of R_i takes nonzero value only when $R_{min} \le R_i \le R_{max}$ and is given by

$$f(R_i) = c \exp\left[-\frac{1}{2}\left(\frac{\ln(R_i/R_m)}{\sigma_{nd}}\right)^2\right], \quad R_{min} \le R_i \le R_{max},$$

where R_{min} is the lower range of truncation, R_{max} is the upper range of truncation, R_m is the mean of inscribed sphere radii, and σ_{nd} is the standard deviation. The

constant c is chosen such that $\int_{R_{min}}^{R_{max}} f(R)\,dR = 1$. The inscribed radius r_{ij} and length l_{ij} of the throat ij are then determined based on the values of R_i and R_j as described in [26]. We create a heterogeneity (region with smaller radii of pores and throats) in our domain as follows. We set $R_m = 0.1$, $R_{min} = 0.05$, $R_{max} = 0.25$ and $\sigma_{nd} = 0.1$ in the region $x < y^2/10 + 25$, and $R_m = 0.02$, $R_{min} = 0.01$, $R_{max} = 0.05$ and $\sigma_{nd} = 0.1$ in the region $x > y^2/10 + 25$. The resulting pore radii as well as throat radii are smaller inside the parabolic region $\Omega = \{(x, y)|x > y^2/10 + 25\}$ than ones outside the region Ω.

Initially the network is filled with wetting fluid. At time $t = 0$ and onwards, the network is assumed to be connected to a non-wetting fluid reservoir on the left hand side and a wetting fluid reservoir on the right hand side. Let P_g denote the non-wetting fluid reservoir pressure. Throughout the simulation on the left hand side, the boundary condition is then $p_i^n = P_g$, $p_i^w = 0$ and $s^w = 0$, on the right hand side boundary we have $p_i^n = p_i^w = 0$, and $s^w = 1$. Periodic boundary conditions are imposed at the rest of boundaries.

From the assumed specific shape of pore bodies and throats, the local capillary pressure p_i^c can be computed from the wetting phase saturation s_i^w using the function

$$p_i^c\left(s_i^w\right) = \frac{2\sigma^{nw}}{R_i(1 - \exp(-6.83 s_i^w))}, \tag{36}$$

where σ^{nw} is the interfacial tension. The entry capillary pressure is defined by

$$\alpha_{ij} = \frac{\sigma^{nw}}{r_{ij}}\left(\frac{1 - \pi/4}{1 - \sqrt{\pi/4}}\right).$$

Physically, the capillary pressure p_i^c cannot be larger than the capillary pressure applied on the RHS boundary (P_g). From the $p_i^c - s_i^w$ relationship (36), we then impose the minimum saturation $s_{i,min}^w$ as follows:

$$s_{i,min}^w = -\frac{1}{6.83}\ln\left(1 - \frac{2\sigma^{nw}}{R_i P_g},\right). \tag{37}$$

The parameters we use in this numerical experiment are $P_g = 30$, $\sigma = 0.0725$, $\mu_n = \mu_w = 0.001$.

With our current MATLAB/Octave implementation, direct numerical simulation of two-phase flows in the entire network over a time interval of interest could not be completed in any reasonable time. We therefore present a multiscale result obtained using the method described in Sects. 3.3.1 and 3.3.2. We discretize the domain $[-250, 250]^2$ into 20×20 coarse blocks and use a fixed time step size $\Delta t = 3$. We refer to the grid lines by x_i and y_j. The boundary conditions for macroscopic pressure P are consistent with the network model: $P = P_g = 30$ in the left side and $P = 0$ in the right hand side, and periodic in the rest of boundary.

The front $\Gamma(t)$ is assumed to be well described by a function of y. The initial front $\Gamma(t)$ is simply the grid line $x = -250$ and the evolution of $\Gamma(t)$ is tracked by the marker points $(x_j(t), y_j)$. At each time t_n, we solve for the pressure field P_{ij} at the grid node (x_i, y_j). The marker points is then moved according to the local

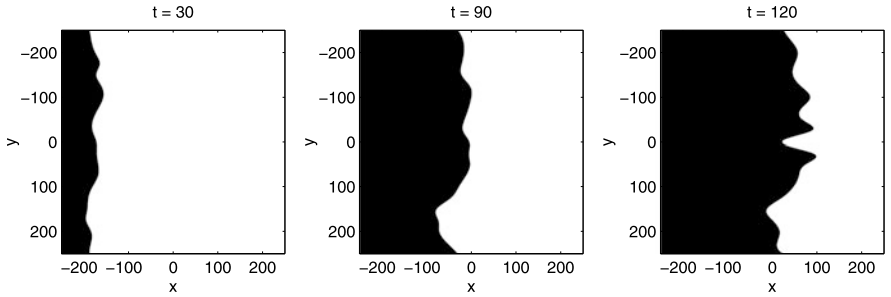


Fig. 8 Snapshot of saturation at $t = 30, 90, 120$. The *black region* is where pore bodies are filled with non-wetting phase fluid ($s_i^w = s_{i,min} \sim 0$). The *white region* is where pore bodies are not invaded ($s_i^w = 1$)

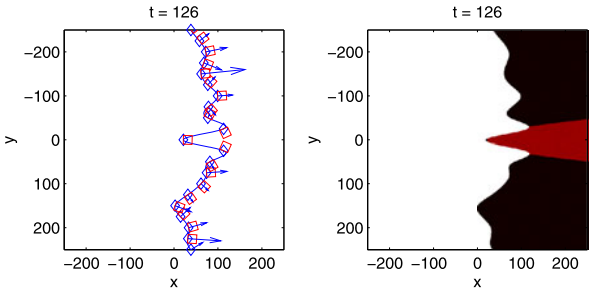


Fig. 9 *Left sub-figure*: Illustrations of front curve with sampling domains when $t = 126$. The *blue arrows* indicate the direction of velocity of marker point obtained from local network simulation. *Right sub-figure*: Capillary pressure p^c at the pores when $t = 126$. The value in the *black region* is 30, the value in the *red region* is in between 7 and 8, and the value in the *white region* is in between 1 and 2

two-phase network simulations. We choose the local network size to be 10×10 and the local simulation duration τ to be 1.5 to estimate the normal velocity of front at the marker points. See the left sbu-figure in Fig. 9 for an illustration. We use cubic spline interpolation on updated marker points to obtain full $\Gamma(t)$ at t_{n+1}, and reassign marker points as $(x_j(t_{n+1}), y_j)$, i.e. the intersection of $\Gamma(t_{n+1})$ and grid line y_j. In Fig. 8, we show snapshots of saturation at $n\Delta t$, $n = 10, 30, 40$. We observe that the front $\Gamma(t)$ is roughly a straight line before it reaches the area $x > y^2/10 + 25$. When it reaches the region Ω, the front can not move into the area because the speed of the marker points on the intersection of $\Gamma(t)$ and the boundary of Ω is very small (due to the parabolic region with smaller pores/throats). See the left sub-figure in Fig. 9. The capillary pressure of each pore at $t = 126$ is shown in the right sub-figure in Fig. 8. By the definition of minimum saturation and our downscaling method, the capillary pressure of the pores behind the front is $P_g = 30$. Because the throats' radii inside the region Ω are smaller, the capillary force inside the region Ω is larger. When the entry capillary pressure α_{ij} is lager than P_g, the non-wetting fluid can never invade the pores.

This phenomenon cannot be easily captured by conventional continuum approaches that neglect capillary pressure. Furthermore, the approach we present can be conceptually extended to cases where the pore scale simulation parameters dynamically responded to changes in macroscopic parameters (see fracture example in [14]).

5 Conclusions

The algorithm presented in this article is built upon the HMM method introduced in [14], where single-phase flows are considered. In this paper, we presented some additional numerical convergence studies for the proposed method. The existing single-phase algorithm is here extended to handle the special scenarios where nonlocal edges in the network develops during the drilling—this is a scenario for which conventional PDE based methods may become inadequate. The new multiscale two-phase front tracking algorithm is able to advance the sharp transition layer of the fluids saturation using a dynamic two-phase network model. We remark here that there is no difficulty in building a level set method that allows for different portions of the front to merge or break up. The purpose of the present paper is to serve as a proof of concept in evaluation of macroscopic front speed using more accurate two-phase network models.

Acknowledgements Chu and Tsai are partially supported by NSF DMS-0714612, and NSF DMS-0914840.

References

1. T. Arbogast. *Gravitational forces in dual-porosity systems: 1. Model derivation by homogenization*. Transp. Porous Media, **13**(2) (1993), 179–203.
2. T. Arbogast. *Gravitational forces in dual-porosity systems: 2. Computational validation of the homogenized model*. Transp. Porous Media, **13**(2) (1993), 205–220.
3. T. Arbogast, L.C. Cowsar, M.F. Wheeler, and I. Yotov. *Mixed finite element methods on nonmatching multiblock grids*. SIAM J. Numer. Anal., **37**(4) (2000), 1295–1315.
4. M.T. Balhoff, K.E. Thompson, and M. Hjortso. *Coupling pore-scale networks to continuum-scale models of porous media*. Comput. Geosci., **33**(3) (2007), 393–410.
5. M. Balhoff, S. Thomas, and M. Wheeler. *Mortar coupling and upscaling of pore-scale models*. *Comput. Geosci.*, **12** (2008), 15–27. doi:10.1007/s10596-007-9058-6.
6. M.J. Blunt. *Flow in porous media—pore-network models and multiphase flow*. Curr. Opin. Colloid Interface Sci., **6** (2001), 197–207.
7. M.J. Blunt, M.D. Jackson, M. Piri, and P.H. Valvatne. *Detailed physics, predictive capabilities and macroscopic consequences for pore-network models of multiphase flow*. Adv. Water Resour., **25** (2002), 1069–1089.
8. S.L. Bryant, P.R. King, and D.W. Mellor. *Network model evaluation of permeability and spatial correlation in a real random sphere packing*. Transp. Porous Media, **11** (1993), 53–70.
9. M.A. Celia, P.C. Reeves, and L.A. Ferrand. *Recent advances in pore-scale models for multiphase flow in porous media*. Rev. Geophys. Suppl., **33** (1995), 1049–1057.

10. Y. Chen and L.J. Durlofsky. *An adaptive local-global upscaling for general flow scenarios in heterogeneous formations*. Transp. Porous Media, **62** (2006), 157–185.
11. Y. Chen and L.J. Durlofsky. *Efficient incorporation of global effects in upscaled models of two-phase flow and transport in heterogeneous formations*. SIAM MMS, **5** (2006), 445–475.
12. Y. Chen, L.J. Durlofsky, M. Gerritsen, and X.H. Wen. *A coupled local-global upscaling approach for simulating flow in highly heterogeneous formations*. Adv. Water Resour., **26** (2003), 1041–1060.
13. Z. Chen, G. Huan, and Y. Ma. *Computational methods for multiphase flows in porous media*. Computational Science & Engineering. SIAM, Philadelphia, 2006.
14. J. Chu, B. Engquist, M. Prodanović, and R. Tsai. *A multiscale method coupling network and continuum models in porous media I—steady state single phase flow*. To appear in SIAM Multiscale Model. Simul. doi:10.1137/110836201.
15. A.B. Dixit, J.S. Buckley, S.R. McDougall, and K.S. Sorbie. *Empirical measures of wettability in porous media and the relationship between them derived from pore-scale modelling*. Transp. Porous Media, **40** (2000), 27–54.
16. L.J. Durlofsky, Y. Efendiev, and V. Ginting. *An adaptive local-global multiscale finite volume element method for two-phase flow simulations*. Adv. Water Resour., **30** (2007), 576–588.
17. W. E and B. Engquist. *The heterogeneous multi-scale methods*. Commun. Math. Sci., **1**(1) (2003), 87–133.
18. Y. Efendiev, V. Ginting, T.Y. Hou, and R. Ewing. *Accurate multiscale finite element methods for two-phase flow simulations*. J. Comput. Phys., **220**(1) (2006), 155–174.
19. Y. Efendiev and T.Y. Hou. *Multiscale finite element methods: theory and applications*. Springer, New York, 2009.
20. B. Engquist, R. Caflisch, and Y. Sun. *A multiscale method for epitaxial growth*. SIAM MMS, **9**(1) (2011), 335–354.
21. B. Engquist and Y. Sun. *Heterogeneous multiscale methods for interface tracking of combustion fronts*. SIAM MMS, **5**(2) (2006), 532–563.
22. I. Fatt. *The network model of porous media I. Capillary characteristics*. Pet. Trans. AIME, **207** (1956), 144–159.
23. I. Fatt. *The network model of porous media II. Dynamic properties of a single size tube network*. Pet. Trans. AIME, **207** (1956), 160–163.
24. I. Fatt. *The network model of porous media III. Dynamic properties of networks with tube radius distribution*. Pet. Trans. AIME, **207** (1956), 164–181.
25. P. Forchheimer. *Hydrolik*. Teubner, Leipzig, 1914.
26. V. Joekar-Niasar, S.M. Hassanizadeh and H.K. Dahle. *Non-equilibrium effects in capillarity and interfacial area in two-phase flow: dynamic pore-network modelling*. J. Fluid Mech., **655** (2010), 38–71.
27. M. Karimi-Fard, B. Gong, and L. J. Durlofsky. *Generation of coarse-scale continuum flow models from detailed fracture characterization*. Water Resour. Res., **42**(10) (2006). doi:10.1029/2006WR005015.
28. J.E. Olson, S.E. Laubach, and R.H. Lander. *Natural fracture characterization in tight gas sandstones: Integrating mechanics and diagenesis*. Am. Assoc. Pet. Geol. Bull., **93**(11) (2009), 1535–1549.
29. P.E. Oren and S. Bakke. *Reconstruction of Berea sandstone and pore-scale modelling of wettability effects*. J. Pet. Sci. Eng., **39** (2003), 177–199.
30. V. Joekar-Niasar, M. Prodanović, D. Wildenschild, and S.M. Hassanizadeh. *Network model investigation of interfacial area, capillary pressure and saturation relationships in granular porous media*. Water Resour. Res., **46**(6) (2010), WR008585.
31. P.C. Reeves and M.A. Celia. *A functional relationship between capillary pressure, saturation, and interfacial area as revealed by a pore scale network model*. Water Resour. Res., **32** (1996), 2345–2358.
32. F.D. Rossa, C. D'Angelo, and A. Quarteroni. *A distributed model of traffic flows on extended regions*. Netw. Heterog. Media, **5**(3) (2010), 525–544.

33. F. Thauvin and K.K. Mohanty. *Network modeling of non-darcy flow through porous media*. Transp. Porous Media, **31** (1998), 19–37.
34. K.E. Thompson. *Pore-scale modelling of fluid transport in disordered fibrous materials*. AIChE J., **48** (2002), 1369–1389.
35. P.H. Valvatne and M.J. Blunt. *Predictive pore-scale modeling of two-phase flow in mixed wet media*. Water Resour. Res., **40** (2004). doi:10.1029/2003WR002627.
36. M.I.J. van Dijke, K.S. Sorbie, and S.R. McDougall. *Saturation-dependencies of three-phase relative permeabilities in mixed-wet and fractionally wet systems*. Adv. Water Resour., **24** (2001), 365–384.
37. X. Wang and K.K. Mohanty. *Pore-network model of flow in gas-condensate reservoirs*. SPE J., **5** (2000), 426–434.
38. S. Whitaker. *Flow in porous media i: a theoretical derivation of Darcy's law*. Transp. Porous Media, **1** (1986), 3–25.
39. S. Youssef, M. Han, D. Bauer, E. Rosenberg, S. Bekri, M. Fleury, and O. Vizika. *High resolution μCT combined to numerical models to assess electrical properties of bimodal carbonates*. Abu Dhabi, UAE, 29 October–2 November, 2008.
40. D. Zhou, M.J. Blunt, and F.M. Orr. *Hydrocarbon drainage along corners of noncircular capillaries*. J. Colloid Interface Sci., **187** (1997), 11–21.

Statistical Geometry and Topology of the Human Placenta

Rak-Kyeong Seong, Pascal Getreuer, Yingying Li, Theresa Girardi, Carolyn M. Salafia, and Dimitri D. Vvedensky

Abstract We present a method of characterising tree networks based on a structural triangulation of those networks. Each component triangle is assigned a generation number which reflects the distance of that component from the origin of the network. By interpreting the generation number as an energy level, we can associate a partition function with a tree network which, in terms of the usual statistical thermodynamic interpretation, enables the determination of the internal energy and entropy of the triangulation. These thermodynamic functions depend on a parameter analogous to an inverse temperature that assigns weights to different parts of the network based on the generation numbers of the triangular elements. The systematic variation of these weights permits an examination of the development of the network, from the initial stages at low temperatures, where lower generation numbers have the greatest weight, to the complete network at high temperature, where all generation numbers have similar weights. After working through several examples to illustrate our methodology, we analyze the arterial and venous vasculature of the chorionic plate of 13 human placentas. We attempt to examine the

R.-K. Seong · D.D. Vvedensky (✉)
The Blackett Laboratory, Imperial College London, London SW7 2BW, UK
e-mail: d.vvedensky@imperial.ac.uk

R.-K. Seong
e-mail: rak-kyeong.seong@imperial.ac.uk

P. Getreuer
Mathematics Department, Yale University, P.O. Box 208283, New Haven, CT 06520-8283, USA
e-mail: pascal.getreuer@yale.edu

Y. Li · T. Girardi · C.M. Salafia
Placental Analytics LLC, 93 Colonial Avenue, Larchmont, NY 10538, USA

Y. Li
e-mail: yingyingli1985@gmail.com

T. Girardi
e-mail: terri.girardi@gmail.com

C.M. Salafia
e-mail: carolyn.salafia@gmail.com

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5_8,

extent to which the entropy function is correlated to the infant birthweight with the sample set. A correlation is postulated as a key factor in determining lifelong health.

1 Introduction

A network is a system of edges that connect vertices or nodes [1, 4, 9, 22]. Networks describe many important functions and processes, including information on the internet, power on a grid, people and freight on road, rail and air networks, and nutrients within biological networks. Graph theory has undergone various incarnations, beginning with Euler's 1736 ground-breaking solution of the Königsberg bridge problem [12], and later in applications to electrical networks [17] and molecules [5]. In the 20th century, the seminal work of Erdos and Renyi [10, 11] and Gilbert [13] on random graphs expanded the discussion to graph component distributions and phase transitions.

Our interest here is in the structure of a particular biological network, the vasculature on the chorionic plate of a human placenta (Fig. 1). The vascular system of a placenta has a branching structure within the chorion that resembles a tree. The origin of the tree is the umbilical cord insertion in the placenta. The extraction of vasculature data is only now becoming available in sufficient quantities to enable statistically significant analysis. The vasculature can be represented as edges and vertices for a graph-theoretical analysis.

There are two levels of analysis that have been applied to such networks. The most basic level is the description of the expansion of such a system with an accompanying cost function. A benchmark for vascular systems was proposed by Murray [20] in 1926 as a compromise between the frictional and metabolic costs, with the latter expressed as a cost function. The formulation of a minimum energy hypothesis led to a scaling law, $Q \propto d^3$, between the volumetric flow rate Q and the diameter d of a blood vessel segment. This scaling law is universal for all trees whose internal flows are laminar. Uylings [30] argued that the exponent in Murray's scaling law can be reduced from 3 to near 2.3 if the fluid flow is turbulent. Numerous studies have found support for Murray's scaling law [23, 28], but with significant scatter. A recent report [18] suggests that animals and plants have reached similar solutions for efficient fluid transport.

An alternative approach in the absence of a cost function is one based solely on the geometrical properties of the network. This includes the fractal dimension and other scaling properties [7, 19, 32], which are determined by the entire network, as well as more local quantities such as the degree distribution of branching points. Fractal analysis in particular has been pervasive in the study of vascular networks and, indeed, of other biological entities, in part because of scale invariance. A noteworthy early effort [3] with scanned x-ray angiograms that produced a binary image of the edges of the blood vessels focussed largely on the fractal dimension of the vascular tree.

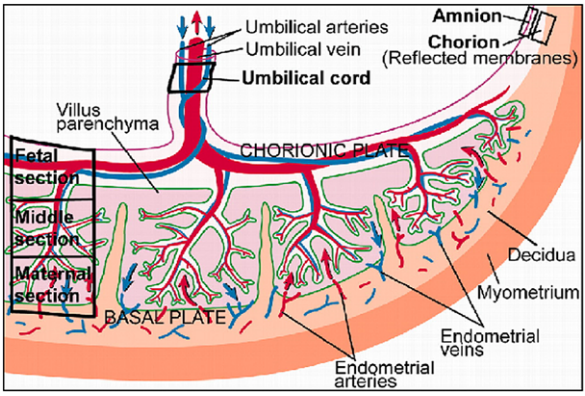


Fig. 1 Schematic view of the placenta showing the chorionic plate, the basal plate, the umbilical cord, and the branching vascular system of the placenta. The work described here analyzes the vasculature on the chorionic plate. Reproduced with permission from R. Sood, J.L. Zehnder, M.L. Druzin, and P.O. Brown, Proc. Natl. Acad. Sci. USA, 103(14), 5478 (2006). Copyright (2006) National Academy of Sciences, USA

The methods described above typically focus only on the final state of the network and ignore its developmental history. Most biological networks have organising principles based on the balance between cost and benefits of extending the network during growth [6, 8, 16]. These principles, in addition to genetic factors, determine the branching structure and final shape of the network. An example is the growth of tree leaves [24, 25, 33]. All tree leaves serve the same purpose of maximizing sunlight exposure while minimizing the costs. By using rules that maximize overall efficiency, the growth of tree leaves can be modelled in terms of parameters that determine the shape of the leaf, giving the variety seen in nature. A similar situation may be expected for the vasculature on the chorionic plate of a human placenta.

In this paper, we introduce a new method for analyzing tree networks by developing geometrical and topological measures. In the absence of loops, a graph has no noteworthy topology in the traditional sense. As a result, our method aims to address this shortfall by introducing new measures that capture mathematically and physically interesting invariants [21] at a given state of the tree's development. Integer generations that are assigned to network vertices which measure the discrete distance to the origin are used to define a partition function from statistical mechanics. The partition function is by definition a function of temperature whose variation is useful to us for the characterization of the network at various stages of its development. This feature leads to the second motivation of this paper: the analysis of the vasculature of the human placenta. The vasculature represents a tree network for which our method provides a way to compute network invariants. Given the limitation that the network is only available for analysis at the end of its development, we attempt with our method to reconstruct the developmental history of the network and compare it with biological parameters, such as the infant birthweight.

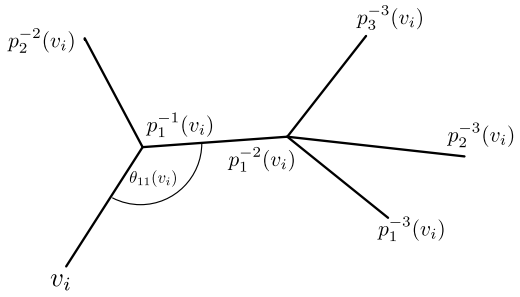


Fig. 2 Labelling of vertices in a 2D tree network

2 Network Triangulation and Partition Functions

2.1 Tree Networks and Biological Systems

A tree network $\mathscr{M} = \{v_i; g, p\}$ is a set of vertices v_i with maps g and p,

$$g: \quad v_i \to g \in \mathbb{Z}^+, \tag{1}$$

$$p: \quad v_i \to v_j, \quad i \neq j. \tag{2}$$

The network has an origin v_0, which we denote as generation $g(v_0) = 1$. All nodes connected by one edge to the origin are generation $g = 2$ nodes. The parent $p(v) = u$ of a generation $g(v)$ vertex v has generation number $g(u) = g(v) - 1$. Inversely, the child $v = p^{-1}(u)$ of a generation $g(u)$ vertex u has the generation number $g(v) = g(u) + 1$. Since we are considering tree networks, no vertex has two or more parents. With $m \in \mathbb{Z}^+$, the union of descendants $p^{-m}(v)$ and ancestors $p^m(v)$ of all vertices v of the network make up the entire set of vertices (Fig. 2).

The network can be extended to include weights w such that $\mathscr{M} = \{v_i; p, g, w\}$. Every vertex in the network can be associated to a weight,

$$w: \quad v_i \to w(v_i) \in \mathbb{R}^+. \tag{3}$$

The weight $w(v_i)$ can be interpreted as the flow rate at v_i in a transport network, or the degree of the branching point corresponding to the number of children for v_i. In order to study the growth mechanism of the network, it may be more appropriate to interpret $w(v_i)$ as a flow rate rather than a geometrical quantity. However, due to the difficulty of measuring the flow rates at every v_i, we assign a different quantity to $w(v_i)$.

The aim of most biological networks is to transport nutrients to a sustained tissue segment. Such tissue segments are located around the vasculature and in the case of human placenta form the chorionic plate. One way to interpret the system is that the further away the tissue segment is from the vasculature origin, the more energy has been consumed to build it at its location during growth. Accordingly, we relate the size of a tissue segment in the neighborhood of v_i to the number of microscopic

physical states at an energy level proportional to $g(v_i)$. We identify the size of the tissue segment at v_i as the weight $w(v_i)$.

We acknowledge possible other interpretations of $g(v_i)$ and $w(v_i)$. The strict proportionality between energy levels and $g(v_i)$, as well as between the tissue segment size and $w(v_i)$, is an inherent assumption of our method. Other interpretations are the subject of future investigations.

2.2 Triangulation of a Tree Network

The biological system will be approximated as a two-dimensional (2D) system. This is a valid approximation since the vasculature is concentrated on the chorionic plate of the placenta. Lower-generation vessels travel parallel to the surface of the chorionic plate outward from the cord insertion. While arteries frequently cross over veins, it is rare for an artery to cross another artery or a vein to cross another vein, so the separate artery and vein networks are usually planar. Vessels of later generations dive down into the villus parenchyma, and are not visible from the surface. This work focuses on the visible portion of the vasculature.

The network is first mapped to $\mathbb{R}^2$ under

$$\vec{x}: \quad v_i \to \boldsymbol{x}(v_i) = (x_i, y_i) \in \mathbb{R}^2. \tag{4}$$

Let all children of a vertex v_i be labelled in the anti-clockwise direction by $\alpha = 1, \ldots, n(v_i)$, where $n(v_i)$ is the number of children of vertex v_i, as shown in Fig. 2. We denote by $\theta_{\alpha\beta}(v_i)$ the anti-clockwise angle between $(v_i, p_\alpha^{-1}(v_i))$ and $(p_\alpha^{-1}(v_i), p_\beta^{-2}(v_i))$. We are interested in the case when

$$0 < \left(\theta_{\alpha\beta}(v_i) - \theta_{\alpha(\beta+1)}(v_i)\right) < \pi, \tag{5}$$

with

$$\left\{p_\alpha^{-1}(v_i), p_\beta^{-2}(v_i), p_{\beta+1}^{-2}(v_i)\right\} \tag{6}$$

forming corners of a triangle. Accordingly, our triangulation $\Delta(\mathscr{M}) = \{\Delta_\alpha(v_i)\}$ of the network structure consists of triangles $\Delta_\alpha(v_i)$ that take the form

$$\Delta_\alpha(v_i) = \left\{v_i, p_\alpha^{-1}(v_i), p_{\alpha+1}^{-1}\right\}, \tag{7}$$

where α is the anti-clockwise labelling of children of v_i, as described above. This labelling convention ensures that no two triangles that are associated with the same vertex v_i overlap. As a result, each vertex v_i is associated with a fan of the form

$$\Delta(v_i) = \left\{\Delta_\alpha(v_i) | \alpha = 1, \ldots, n(v_i)\right\}. \tag{8}$$

The motivation for introducing a network triangulation is that each triangle can be interpreted as a tissue segment which is sustained by the vascular system. A triangle is associated with the generation of the parent vertex v_i,

$$g\big(\Delta_\alpha(v_i)\big) = g(v_i), \tag{9}$$

and has an area in $\mathbb{R}^2$ which is denoted by $A(\Delta_\alpha(v_i))$. Accordingly, the weight of a vertex v_i is chosen to be

$$w(v_i) = \sum_{\alpha=1}^{n(v_i)} A\big(\Delta_\alpha(v_i)\big). \tag{10}$$

2.3 Interpretation of Variables

In order to relate the network triangulation to statistical mechanical quantities, such as the partition function, we introduce an interpretation of network variables which is summarized in this section.

We suppose, for simplicity, that the network variables and statistical variables are proportional. The area $A(\Delta_\alpha(v_i))$ is chosen to be proportional to the size of the sustained tissue segment in the vicinity of vertex v_i. During growth, the network is required to transport material proportional to $A(\Delta_\alpha(v_i))$ from the origin v_0 to the vicinity of v_i. The higher the level $g(v_i)$ is, the more work has done to transport material proportional to $A(\Delta_\alpha(v_i))$. Accordingly, $g(v_i)$ is interpreted as an energy level of the system $E(g) = gE_0$.

We associate $A(\Delta_\alpha(v_i))$ with the size of a set of microscopic states which are at energy level $E(g)$ and relate to vertex v_i. We define this to be the weight $w(v_i)$. The total number of states at $g(v_i)$ is then

$$A(g) = \sum_{g(v_i)=g} w(v_i) = \sum_{g(v_i)=g} \sum_{\alpha=1}^{n(v_i)} A\big(\Delta_\alpha(v_i)\big). \tag{11}$$

2.4 Partition Function for Tree Networks, Symmetries, and Scale-Invariance

The discussion in the section above motivates a description of the network in the context of statistical mechanics. Given the canonical ensemble of states in the vicinity of vertices v_i and enclosed by triangles $\Delta_\alpha(v_i)$, we define a 'partition function' of states as

$$Z(\mathscr{M}) = \sum_g A(g) t^g = \sum_{\Delta_\alpha(v_i)} A\big(\Delta_\alpha(v_i)\big) t^{g(\Delta_\alpha(v_i))}, \tag{12}$$

where $t = e^{-\beta E_0}$ and β is a parameter that controls the thermodynamic weight given to successive generations. In other words, t is the fugacity that counts the generation level associated with a triangle $\Delta_\alpha(v_i)$ with area $A(\Delta_\alpha(v_i))$. The ranges of β and t are $0 \leq \beta < \infty$ and $0 < t \leq 1$. In a traditional statistical mechanical formulation, β is associated with the inverse of the absolute temperature. Since the role of the factor t^g is to vary the weight of different generations of the network, we will use β as a control parameter that enables us to examine the development of the network from the initial ($\beta \to \infty$) to the final complete network ($\beta = 0$). We first examine the partition function in these limits.

In the limit $\beta \to 0$,

$$\lim_{\beta \to 0} Z(\mathscr{M}) = \sum_g A(g) = A(\mathscr{M}), \tag{13}$$

the partition function becomes the total area of network triangles. This is the "high-temperature" limit. In the limit $\beta \to \infty$ (the "low-temperature" limit), the partition function vanishes, as expected. Given that the limit $\beta \to 0$ captures all triangles and, hence, all microscopic states, we associate this limit with the final state of the network growth process at which the placenta vasculature is captured for analysis. The opposite limit $\beta \to \infty$ is associated with the beginning of growth close to the origin.

The partition function is invariant under certain symmetry transformations of the network in $\mathbb{R}^2$. An example is a global $\mathrm{GL}(2, \mathbb{R})$ transformation which maps the coordinates of vertices as follows,

$$\mathsf{M}: \quad \boldsymbol{x}(v_i) = (x_i, y_i) \to \boldsymbol{x}'(v_i) = \mathsf{M} \cdot \boldsymbol{x} = (ax_i + by_i, cx_i + dy_i), \tag{14}$$

where

$$\mathsf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2, \mathbb{R}), \tag{15}$$

is an invertible matrix. Under this transformation, the partition function $Z(\mathscr{M})$ gains an overall factor $|\det(M)|$ which can be set to 1. This would then be a symmetry of the partition function.

In general, in order to guarantee a scale-invariant network analysis, all spatial coordinates are normalized so that the maximum diameter of the placenta perimeter is 1.

2.5 Examples

Figure 3 shows two examples of tree networks with vertex generations and network triangulations. The partition functions corresponding to the triangulations in panel (c) and (f) are

$$Z(\mathscr{M}_c) = 16t + 16t^2 + 8t^3, \tag{16}$$

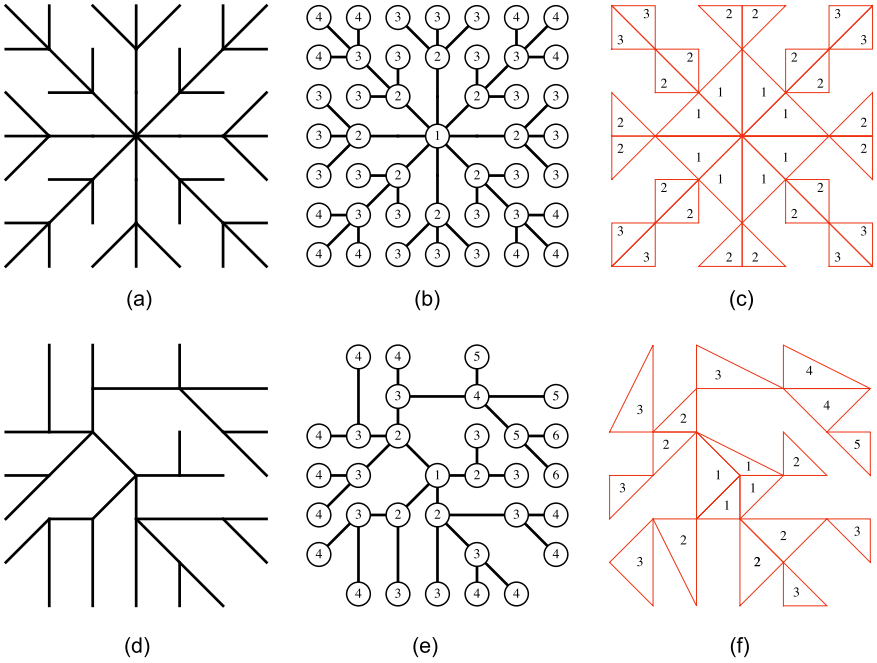


Fig. 3 Two tree networks (**a**, **d**) shown with the generation number of their vertices indicated with *circles* centered at each vertex in (**b**, **e**) and the corresponding triangulations in (**c**, **f**)

$$Z(\mathscr{M}_f) = 5t + 9t^2 + 9t^3 + 4t^4 + t^5. \tag{17}$$

In the limit $\beta \to 0$ ($t \to 1$), the total triangulated areas are

$$A(\mathscr{M}_c) = 40, \tag{18}$$

$$A(\mathscr{M}_f) = 28. \tag{19}$$

Note that the more symmetric the tree network, the larger the total triangulated area.

3 Statistical Topology of Triangulated Networks

In Sect. 2.4, we have identified a tree network and its triangulation with a statistical mechanical system through the formulation of a partition function of states. The partition function allows us to calculate 'thermodynamic' variables such as the 'internal energy' U and 'entropy' S of the canonical ensemble as a function of the weighting parameter β. These quantities are calculated in the usual way from the partition function and we will see that they have analogous interpretations as their real thermodynamic counterparts.

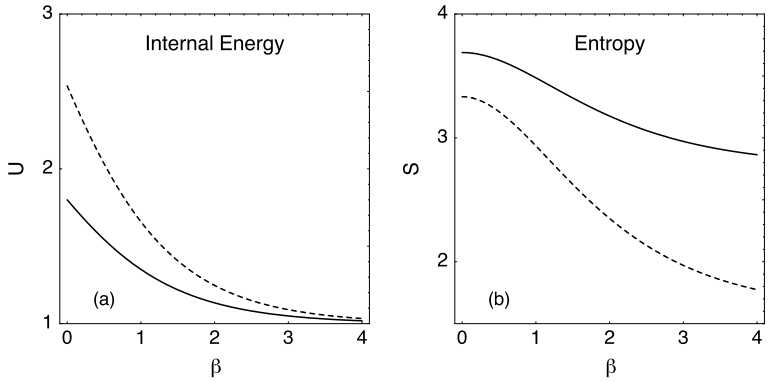


Fig. 4 (**a**) Internal energy and (**b**) entropy for the networks in Fig. 3(a) (*solid line*) and Fig. 3(d) (*dashed line*) determined from their triangulations

3.1 Internal Energy

The internal energy U of the canonical ensemble of a tree network is obtained using the standard relation

$$U = -\frac{\partial \ln Z}{\partial \beta}. \tag{20}$$

The internal energies of the two networks in Fig. 3 are shown in Fig. 4. Figure 4(a) shows that $U_d > U_a$ for all β. In the first few generations (large β), the internal energies of the two networks are similar, but diverge as higher generations are included (decreasing β). A physical rationale for this behavior is suggested by using the definition of the partition function (12) in the definition of the internal energy to obtain

$$U = E_0 \frac{\sum_{\Delta_\alpha(v_i)} A(\Delta_\alpha(v_i)) g(v_i) t^{g(v_i)}}{\sum_{\Delta_\alpha(v_i)} A(\Delta_\alpha(v_i)) t^{g(v_i)}} \equiv \left\langle E_0 g(v_i) \right\rangle, \tag{21}$$

which is the thermodynamic average of the work required to add E_0 across all generations. As $\beta \to \infty$ ($t \to 0$), i.e. in the initial stages of development,

$$\lim_{\beta \to \infty} U = E_0, \tag{22}$$

since only the first generation contributes. For each network, we work in energy units of E_0, so all internal energy profiles approach unity as $\beta \to \infty$. The other limit, where $\beta \to 0$ ($t \to 1$) is, according to (21),

$$\lim_{\beta \to 0} U = E_0 \frac{\sum_{\Delta_\alpha(v_i)} A(\Delta_\alpha(v_i)) g(v_i)}{\sum_{\Delta_\alpha(v_i)} A(\Delta_\alpha(v_i))}, \tag{23}$$

which is the average work required to add E_0 across all generations, with each generation weighted equally. Hence, the internal energy profiles of all networks

coincide initially ($\beta \to \infty$), but differences in later generations are manifested as deviations in these profiles as $\beta \to 0$. Thus, the physical interpretation of the behavior of the internal energy in Fig. 4(a) is that the structure of the asymmetric network in Fig. 3(d) displays a higher internal energy because more states in the canonical ensemble belonging later generations ($\beta \to 0$) of the system than in the symmetric network in Fig. 3(a).

3.2 Entropy

The entropy of the canonical ensemble associated with the tree network is given by the standard equation

$$S = k_B(\ln Z + \beta U), \tag{24}$$

where Boltzmann's constant k_B has been set to unity. As Fig. 4(b) shows, $S_a > S_d$ for all β. To understand this result, we interpret entropy as a measure of energy dispersal in the network, rather than as a measure of structural 'disorder.' Energy in the symmetric network is more evenly distributed between different generations than in the asymmetric network, and hence has the higher entropy. Thus, our interpretation is that entropy measures the extent to which the triangle areas in the network are distributed in different generations.

As $\beta \to \infty$ ($t \to 0$), entropy approaches the logarithm of the triangle areas associated with the first generation,

$$\lim_{\beta\to\infty} = \ln\big[A(g=1)\big]. \tag{25}$$

The other limit is $\beta \to 0$ ($t \to 1$), where entropy approaches the logarithm of the area of the entire triangulation,

$$\lim_{\beta\to 0} = \ln\left[\sum_g A(g)\right]. \tag{26}$$

As discussed above, both of these limits can be directly obtained from the partition function. The interesting situation is when the entropy curves of different networks cross. In a later section, we will provide an interpretation of this situation in the context of the vasculature on the chorionic plate of the human placenta.

4 Analysis of Chorionic Plate Vasculature

The placenta is the sole source of oxygen and nutrients for the fetus. Placental transport function is one of the main factors in the health and development of both the fetus and the placenta itself. The fetal origins of adult health hypothesis [2] suggests

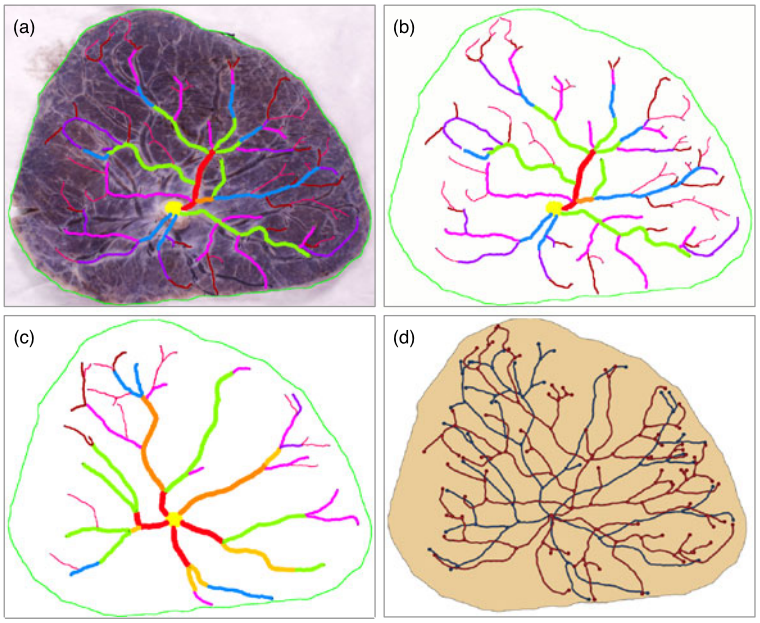


Fig. 5 (**a**) A digital photograph of the chorionic plate of the placenta with hand tracing which highlights the visible surface vasculature (arteries are traced here). (**b**) and (**c**) show the hand traced arterial and venous vasculature respectively prior to computer aided processing. (**d**) Computer generated graph showing both the arterial (*red*) and venous vasculature

that placental development is even linked to adult health. The placenta grows and develops alongside the fetus (Fig. 1). The surface of the placenta attached to the endometrium (lining) of the uterine wall is called the basal plate and the surface nearest to the fetus the chorionic plate. Between the two is a complex arborized vascular network through which oxygen, nutrient and waste exchange takes place. The vascular system of the placenta consists of two arteries and one vein that form a treelike network in the chorion with its origin placed at the umbilical cord insertion. Beyond the chorionic plate, the veins and arteries dive into the placenta with further branching. In our analysis we focus on the vascular structure on the chorionic plate, which we regard as a 2D tree network. The analysis distinguishes between the arterial and venous vasculature.

4.1 Tracing the Vasculature

Extracting the network information from the placenta is a time-consuming multi-step process. Figure 5 illustrates the steps involved in extracting the placental vasculature from a 2D digital photograph of the chorionic plate of the placenta. These steps are:

(a) The placenta is first photographed and then hand traced, highlighting the vasculature of interest, as well as the cord insertion point.
(b, c) The tracing and the cord insertion point are filtered out from the original image.
(d) The network graph is identified from the tracing using a computer program which identifies the vertices and edges of the graph. This enables the vasculature to be represented as a graph embedded in $\mathbb{R}^2$, with the origin v_0 identified with the umbilical cord insertion point on the chorionic plate. Every branching point and endpoint of the vascular system is identified with a vertex point of the network, v_i.

To carry out a scale invariant comparison between different placentas, we introduce a descriptor of the placenta vasculature. All coordinates are normalized such that the maximum diameter of the perimeter of the placenta is 1. We distinguish between the arterial and venous vasculature on the chorionic plate of the placenta. Figure 5(d) illustrates the identified vasculature, where arterial and venous vasculature is highlighted in red and blue, respectively.

In general, it is a challenge to efficiently distinguish between branching points and vascular overlap and intersection points by using a computer aided graph extraction scheme. This is because a 2D image does not convey unambiguously the fundamentally three-dimensional (3D) structure of the vasculature on the chorionic plate of the placenta. These limitations can be minimized when the analysis is extended to 3D. In fact, we expect to be able to extend the analysis method presented in this work to any higher dimension. In particular, the extension to three dimensions involves a decomposition of the tree not into triangles but tetrahedra whose volumes contribute to the associated partition function. This is an extension which we will adopt in future investigations.

4.2 Thermodynamic Analysis

We consider the arterial vasculature on the chorionic plate in Fig. 6 which has in total $g_{\max} = 12$ generations. The corresponding partition function is given by

$$Z(\mathscr{M}) = 0.0095t + 0.0107t^2 + 0.0118t^3 + 0.0122t^4 + 0.0292t^5 + 0.0393t^6 + 0.0181t^7 + 0.0232t^8 + 0.0080t^9 + 0.0092t^{10} + 0.0164t^{11} + 0.0013t^{12}. \tag{27}$$

As discussed above, the fugacity $t = e^{-\beta}$ depends on the parameter β, which allows us to weight the contributions of successive generations to the partition function. For large values of β, only the early generations contribute significantly to the partition function and, therefore, to the internal energy and the entropy. This is indicated in Fig. 6(c), where the part of the network that contributes appreciably to the partition function is indicated in red, while the remaining network is indicated in gray. As

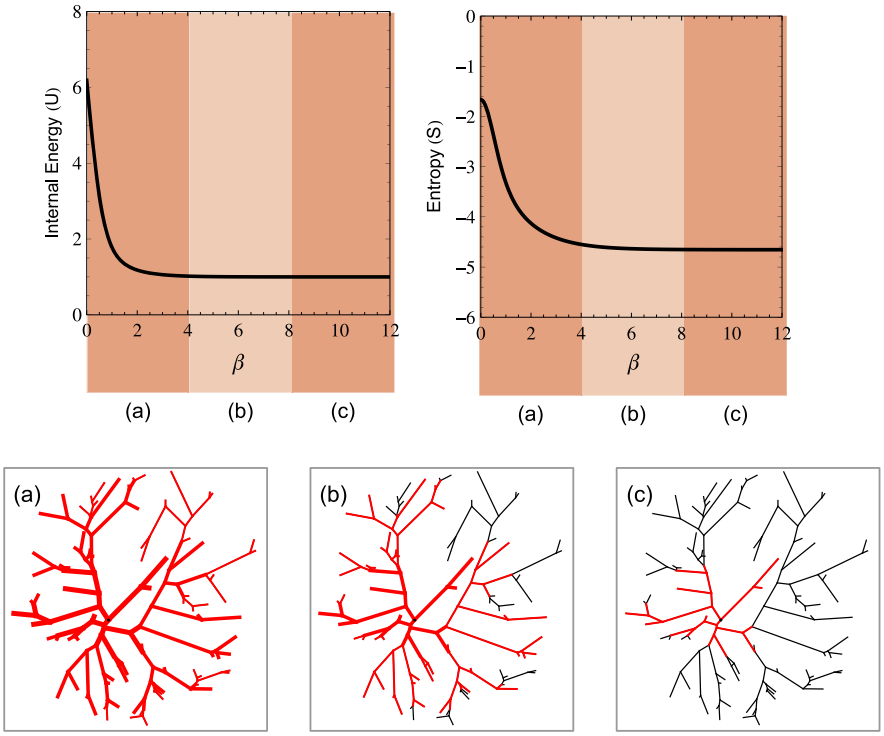


Fig. 6 An example of an arterial vascular tree on the chorionic plate of a human placenta (*lower panels*), together with the internal energy and entropy obtained from the partition function in (27) (*upper panels*). The *bottom panels* show the vascular tree along with a representation of the effect of the weighting parameter β. For large β, only the early generations contribute significantly to the partition function (**c**). As β decreases, later generations make successively larger contributions to the partition function (**a**, **b**) and eventually crossover to the limit where the entropy is the logarithm of the area of the entire triangulation

β decreases, the contributions of successive generations to the partition function increases. Figure 6(b) shows an intermediate stage, while Fig. 6(a) shows the final stage when essentially the entire network contributes to the partition function. In the limit when $\beta \to 0$, i.e. $t \to 1$, all parts of the network contribute equally to the partition function, as (27) readily indicates.

This example illustrates several important aspects of our analysis. As (25) shows, the large-β limit of the entropy is the logarithm of the triangulated area in the first generation. This provides an indication of how the vasculature grows initially, with larger values corresponding to relative symmetric growth compared to smaller values. The limiting entropy in the small-β limit (26) is the logarithm of the area of the entire triangulation which, as will be seen below, does not exhibit the same extent of variation as in the large-β limit. However, a characteristic feature of vasculature development is the crossover between these limits, which is determined by the number of generations. The greater the area in later generations the more delayed

the crossover. This will, of course, be reflected in other network statistics, but the profile of the entropy across all values of β provides an instant assessment of this key aspect of the vasculature.

4.3 Analysis According to Birth Weight

As we mentioned in the introduction, birthweight has been linked to infant health and, indeed, to lifelong health [2]. But weight is a cumulative summary of many developmental factors, so an important question is whether there are other, more refined characteristics of the human placenta that can provide an indication of health prior to birth. In this regard, the vasculature forms a natural starting point because of its role in the transport of nutrients to the fetus from the mother.

As a prelude to this discussion, Fig. 7 shows the arterial vasculatures (in black) and triangulations (in red) of six placentas with the indicated birth weights, together with their entropy profiles. Our data comes from a subset of previous studies [26, 27]. In brief, women were recruited upon presentation at 11–14 weeks gestation for aneuploidy screening and informed consent was obtained. During the nuchal translucency exam, the trans-abdominal probe (GE Voluson E8) was used to obtain a 3D volume sweep of the placenta. The volumes were obtained using power Doppler (quality-max, PRF-0.6) with the sweep angle opened to ensure inclusion of the entire placenta. Volumes were stored for offline analysis post-partum. After delivery, the trimmed placental weight was obtained to the nearest gram. A digital photograph of the chorionic plate surface was then obtained according to a previously described protocol and the perimeter, chorionic surface vasculature and cord insertion sites traced or marked using a graphics tablet [Fig. 5(a)].

The birth weights range from the higher end of what is regarded as low birth weight to what is regarded as high birth weight. The differences between the arterial vasculatures of these placentas are most apparent from their entropy profiles. With the limit $\beta \to \infty$ corresponding to low generation numbers, we can group the placentas according to their entropies in this limit into the sets {(a), (b)}, {(b), (c)} and {(e), (f)}. Given the interpretation that a higher entropy relates to a more evenly distributed microscopic physical states, one observes that placentas (c) and (d) exhibit the highest distribution at lower generations.

Figure 7 shows that the entropy profiles can discriminate between various aspects of state distributions. At this point, however, we still have a purely modified topological description of the vasculature. We have made no attempt to relate the generational distribution of triangulated areas in terms of the transport characteristics of the network. A first step in this direction would be to incorporate vascular segment diameters or thicknesses to the partition function proposed above. This can be done by incorporating the diameter of the vascular segment into the weights $w(v_i)$ on vertices v_i. Modifications in this direction may be able to extend our analysis to results from blood perfusion studies [29, 31]. Studies along these lines will be the subject of future work. For the moment, we will examine the extent to which this purely *geometrical* approach provides an understanding of birthweight variations.

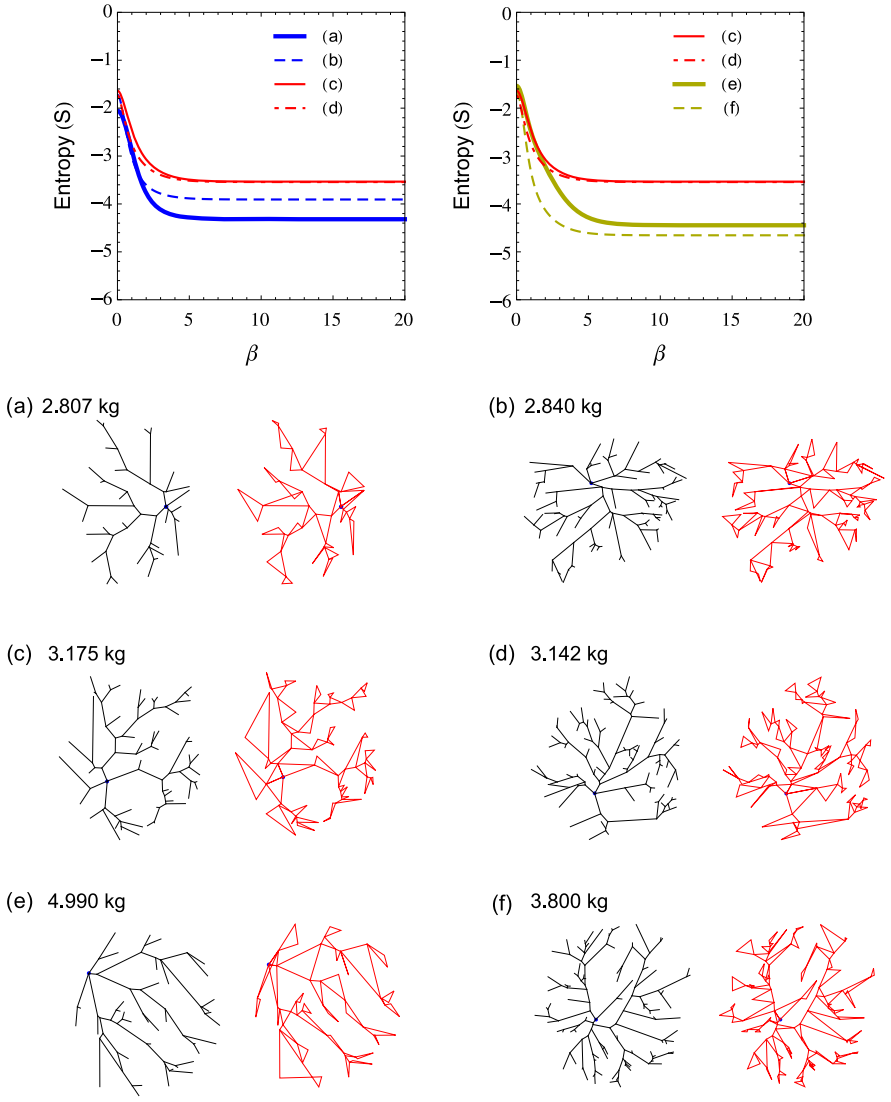


Fig. 7 Six tree networks corresponding to the vasculature of human placentas, obtained according to the procedure in Sect. 4.1, with the indicated birth weights (*lower panels*) together with their entropy profiles (*upper panels*). For each placenta, the network is shown in *gray* and the corresponding triangulation shown adjacent in *red*

Bearing in mind the limitations of a small sample size, as well as the potential for measurement errors of the placental vasculature, Fig. 8 shows several features of the entropy profiles of 13 placentas divided into categories of low (2–3 kg), normal (3–4 kg) and high (4–5 kg) infant birth weights. In common with the placentas in Fig. 7, the entropies of the entire vasculatures [Figs. 8(a), (b)] show a much smaller

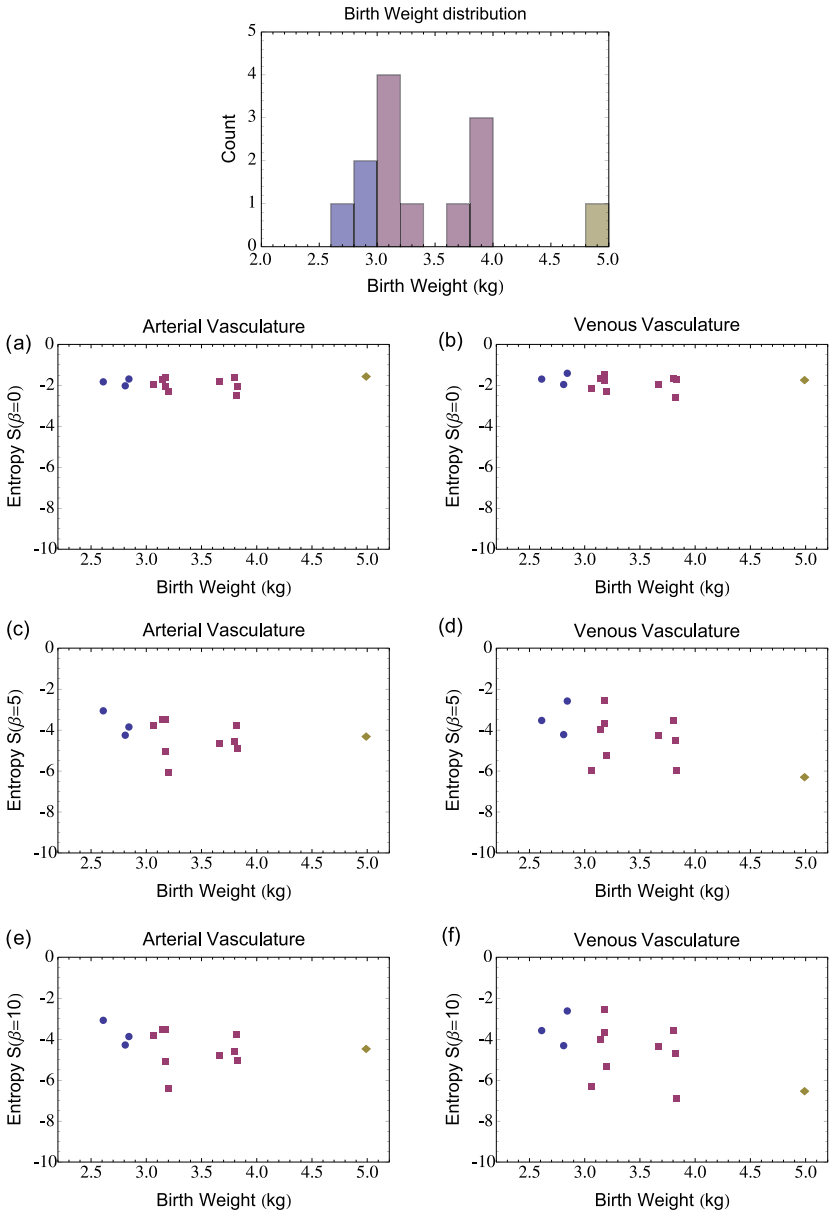


Fig. 8 The distribution of birth weights for the placentas in our sample, with 3 corresponding to the low birthweight category (2–3 kg), shown in *blue*, 9 to the normal birthweight category (3–4 kg), shown in *red*, and 1 in the high birthweight category (4–5 kg), shown in *yellow*. Panels (**a**)–(**f**) show the characteristics of the entropy profiles determined for the arterial and venous vasculature of the placentas in the sample set using the same color coding: (**a**)–(**b**) the final entropy, taken as $S(0)$, (**c**)–(**d**) the interim entropy, $S(5)$, and (**e**)–(**f**) the initial entropy taken as $S(10)$. The panels are divided into arterial vasculature plots (**a**, **c**, **e**) and venous vasculature plots (**b**, **d**, **f**)

variation than at any other point in their development. This trend is seen for the other entropy measures in Fig. 8 and shows quite clearly that arguments based on the mean alone do not provide a complete picture of the entropy trends of placental vasculature.

The small sample set limits the observation of differences between arterial [Figs. 8(a), (c), (e)] and venous [Figs. 8(b), (d), (f)] vasculatures. Given the small sample size for high birthweight placentas, it is difficult to claim that with increasing weight, the venous vasculature exhibits a greater variation in entropy than the arterial vasculature. Figures 9(a) and (b) show the total entropy changes of the arterial and venous vasculatures of the 13 placentas. Figures 9(c)–(f) show two measures of the point at which the entropy crosses over between the initial and final values. In Figs. 8(c) and (d), the crossover point is defined as the value of β at which the entropy takes the mean value over the range $0 \leq \beta \leq 20$. In Figs. 8(e) and (f), the crossover point is defined as the value of β at which the gradient of the entropy is for the first time from $\beta \to \infty$ less that -0.00025 with a resolution of $\Delta\beta = 0.1$. In all cases, there are too few high birth-weight placentas to draw any statistically significant conclusions.

Another significant parameter in the characterization of placentas is the gestational age at birth. Figure 10 shows the initial, interim and final entropies against the gestational age of the placenta. A trend observed from the small sample set of 13 placentas is a difference between the behavior of the arterial and venous vasculature when comparing the spread of entropy values at average gestation ages and initial entropy values. The spread appears to be larger for the venous network than the arterial network.

5 Conclusion

We have made use of concepts from geometry, topology and graph theory in order to propose an alternative method of analysing and comparing tree networks. The method utilizes thermodynamic functions known from statistical mechanics as simple measures of the weighted topology of a tree network. The motivation has been two-fold. The method takes a highly complex structure and returns simple measures based on topology and geometry. When expressed in terms of thermodynamic functions, these measures allow us to reconstruct the growth history of a network by weighting network segments according to their discrete distance to the origin.

The main features of the method presented here are as follows. It makes use of a triangulation of the tree network which enables us to formulate a partition function. The triangles have been ordered according to the generation of the related vertices which is used to construct the partition function and to calculate profiles of thermodynamic functions. Particular vertex generations are weighted more than others in order to evaluate network properties at specific periods of development. An especially interesting thermodynamic function for the analysis of a tree network is the entropy. A higher entropy at a certain network weighting is interpreted as a higher

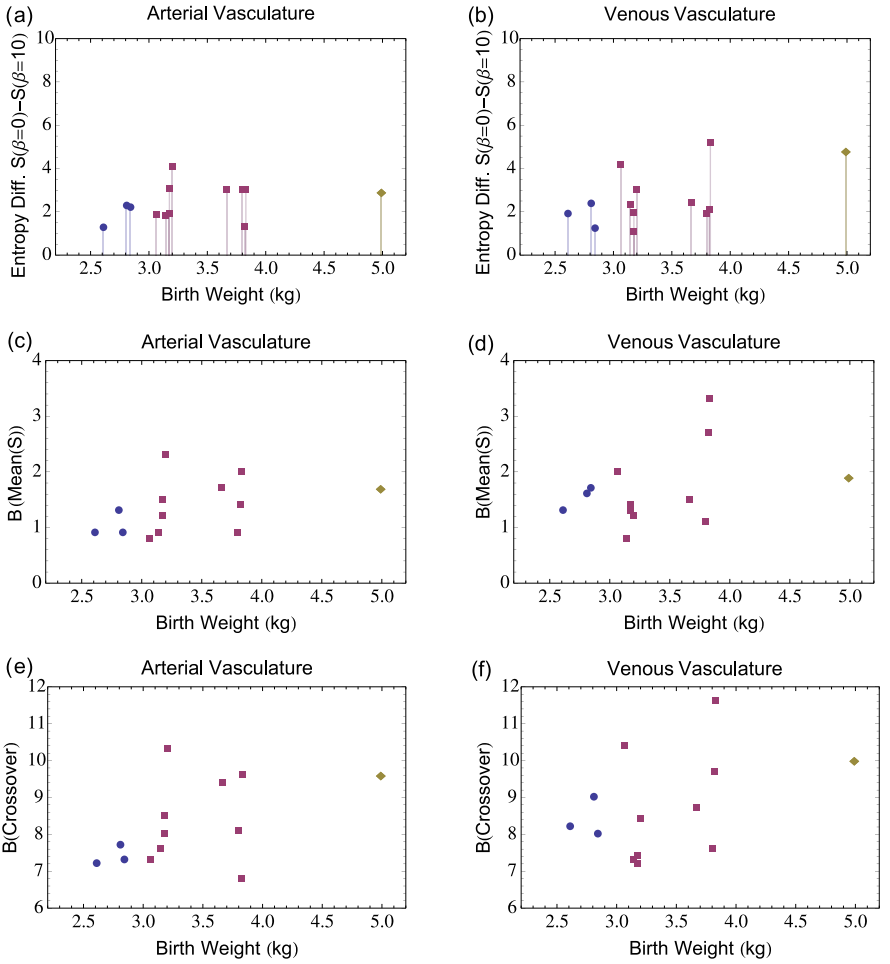


Fig. 9 Panels (**a**) and (**b**) show respectively the variation of entropy between early ($\beta = 10$) and late ($\beta = 0$) stages of the arterial and venous vasculature. Panels (**c**) and (**d**) show the value of β at which the entropy reaches the mean entropy value in the range $0 \leq \beta \leq 20$ for arterial and venous vasculature respectively. Panels (**e**) and (**f**) show respectively the β value at which the gradient of the entropy curve for the first time from $\beta \to \infty$ reaches the value -0.00025 with resolution $\Delta\beta = 0.1$ for the arterial and venous vasculature. The color coding is the same as in Fig. 8

distribution of states and their energies in the associated canonical ensemble of the thermodynamic model. As an example, a symmetric tree network is related to an overall high entropy over its entire lifespan. By comparison, an asymmetric tree network is related to an overall lower entropy.

Our motivation for this work has been to apply the proposed simple measures of tree networks to the vasculature of a human placenta with the aim of relating our proposed measures to clinical parameters such as the infant birthweight, which is a

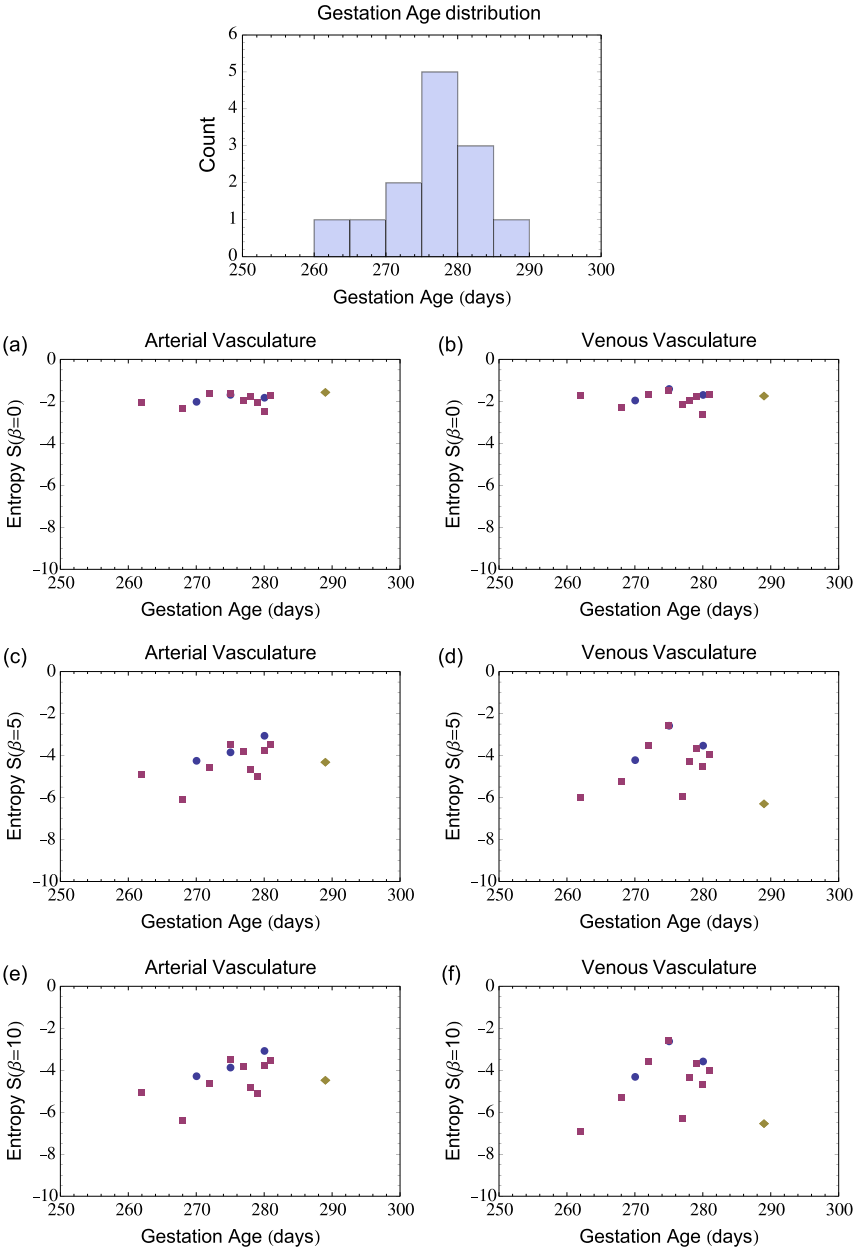


Fig. 10 Entropy as a function of gestational age for 13 placentas. The histogram shows the distribution of gestation ages for the sample set. Panels (**a**) and (**b**) show the distribution the final entropy, again taken as $S(0)$. Panels (**b**) and (**c**) show the interim distribution with entropy $S(5)$. Finally, Panels (**e**) and (**f**) show the distribution at initial entropy taken as $S(10)$. The color coding is the same as in Fig. 8 for birthweight categories. The panels are divided into arterial and venous networks of the sample placentas

primary determinant of the health of the fetus and the infant after birth. The results and interpretations for this comparison of measures have thus far been qualitative, due in part to the small sample size of available placenta vasculatures. In addition, the properties of the vasculature itself, such as the increasing diameter of vasculature segments closer to the cord insertion point as well as the 3D nature of the vasculature [3], which we have projected onto two dimensions, lead to the conclusion that the data we have used for our analysis have significant limitations. Subject to available data, future investigations could extend our method to three dimensions, where the triangulation of the network is replaced by a segmentation into tetrahedra. Furthermore, one of the central features of our analysis—the partition function—can be extended to incorporate additional information about the vasculature, such as the diameter of vasculature segments, as noted above.

The encouraging qualitative results regarding the spread of entropy and other measurable values with our method are weakened by the lack of data available for low and high infant birthweight placentas. However, the qualitative properties of entropy profiles as a tool of analysing vascular trees of placentas is observed and we are encouraged to follow this path of analysis further with the aid of a significantly larger sample set of placentas. Extensive testing of the methods we have presented above is expected to aid in developing growth models for the chorionic vasculature of the human placenta.

Taking a broader perspective, the approach developed in this paper is a 'mesoscopic' view of the placenta in which the focus is on the graph-theoretic aspects of the vasculature. This should be compared with the even more coarse-grained 'macroscopic' description of the placenta based on the distributions of shapes and measures of morphology [15], and the 'microscopic' calculations of diffusional oxygen transport in terminal villi [14]. These approaches and the characteristics they quantify are, of course, interrelated, which is just another way of saying that modelling the placenta is an inherently multiscale problem. Indeed, establishing a connection between models of the placenta and clinical outcomes will likely require a more faithful view of the vasculature as a transport network [29, 31]. Applied in other contexts [33], such an approach has proven quite fruitful in modeling biological systems.

Acknowledgements R.-K.S. was supported in part by Placental Analytics LLC. C.M.S. was partially supported by a NARSAD Young Investigator Award and by K23 MidCareer Development Award NIMH K23MH06785. This material is based upon work supported by the National Science Foundation under award DMS-1004694. This work was also supported by NO1-HD-5-3411 "NCS Formative Research Project 18–Placental Study" to Placental Analytics, LLC.

References

1. Albert, R. and Barabási, A.-L. *Statistical mechanics of complex networks*. Rev. Mod. Phys. **74** (2002), 47–97.
2. Barker, D. J. P. *Fetal origins of coronary heart disease*. BMJ **311** (1995), 171–174.

3. Bergman D. L. and Ullberg, U. *Scaling properties of the placenta's arterial tree*. J. Theor. Biol. **193** (1998), 731–738.
4. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. and Hwang, D. U. *Complex networks: structure and dynamics*. Phys. Rep. **424** (2006), 175–308.
5. Cayley, A. *On the theory of analytical forms called trees*. Philos. Mag. **13** (1857), 172–176.
6. Corson, F. *Fluctuations and redundancy in optimal transport networks*. Phys. Rev. Lett. **104** (2010), 048703.
7. Cross, S. C. *Fractals in pathology*. J. Pathol. **182** (1997), 1–8.
8. Dodds, P. S. *Optimal form of branching supply and collection networks*. Phys. Rev. Lett. **104** (2010), 048702.
9. Dorogovtsev, S. N., Goltsev, A. and Mendes, J. F. F. *Critical phenomena in complex networks*. Rev. Mod. Phys. **80** (2008), 1275–1335.
10. Erdös, P. and Rényi, A. *On random graphs. I*. Publ. Math. **6** (1959), 290–297.
11. Erdös, P. and Rényi, A. *The evolution of random graphs*. Magyar Tud. Akad. Mat. Kutató Int. Közl. **5** (1960), 17–61.
12. Euler L. *Solutio problematis ad geometriam situs pertinentis*. Comment. Acad. Sci. Imp. Petrop. **8** (1736), 128–140. [English translation: Newman, J. Leohnard Euler and the Königsberg bridges. Sci. Am. **189**(1) (1953), 66–70]
13. Gilbert, E. N. *Random graphs*. Ann. Math. Stat. **30** (1959), 1141–1144.
14. Gill, J. S., Salafia, C. M., Grebenkov, D. and Vvedensky, D. D. *Modeling oxygen transport in human placental terminal villi*. J. Theor. Biol. **291** (2011), 33–41.
15. Gill, J. S., Woods, M. P., Salafia, C. M. and Vvedensky, D. D. *Probability distributions for measures of placental shape and morphology* (unpublished). arXiv:1109.2057.
16. Katifori, E., Szöllősi, G. J. and Magnasco, M. O. *Damage and fluctuations induce loops in optimal transport networks*. Phys. Rev. Lett. **104** (2010), 048704.
17. Kirchhoff, G. *Über die Auflösung der Gleichungen, auf Welche man bei der Untersuchung der Linearen Verteilung Galvanischer Ströme Geführt Wird*. Poggendorf Ann. Phys. Chem. **72** (1847), 497–508. [English translation: *On the solution of the equations obtained from the investigation of the linear distribution of Galvanic currents*. IRE Trans. Circuit Theory **5** (1958), 4–8.]
18. McCulloh, K. A., Sperry, J. S. and Alder, F. R. *Water transport in plants obeys Murray's law*. Nature **421** (2003), 939–942.
19. Lorthois, S. and Cassot, F. *Fractal analysis of vascular networks: Insights from mortphogenesis*. J. Theor. Biol. **262** (2010), 614–633.
20. Murray, C. D. *The physiological principle of minimum work: I. The vascular system and the cost of blood volume*. Proc. Natl. Acad. Sci. USA **12** (1926), 207–214.
21. Nakahara, M. *Geometry, Topology and Physics*. IOP Publishing, Bristol, 2003.
22. Newman, M. E. J. *The structure and function of complex networks*. SIAM Rev. **45** (2003), 167–256.
23. Painter, P. R., Edén, P. and Bengtsson, H. U. *Pulsatile blood flow, shear force, energy dissipation and Murray's law*. Theor. Biol. Med. Model. **3** (2006), 31.
24. Roth-Nebelsick, A., Uhl, D., Mosbrugger, V. and Kerp, H. *Evolution and function of leaf venation architecture: a review*. Ann. Bot. **87** (2001), 553–566.
25. Runions, A., Fuhrer, M., Lane, B., Federl, P., Rolland-Lagan A.-G. and Prusinkiewicz P. *Modeling and visualization of leaf venation patterns*. ACM Trans. Graph. **24** (2005), 702–711.
26. Schwartz, N., Coletta, J., Pessel, C., Feng, R., Timor-Tritsch, I. E., Parry, S. and Salafia, C. M. *Novel 3-dimensional placental measurements in early pregnancy as predictors of adverse pregnancy outcomes*. J. Ultrasound Med. **29** (2010), 1203–1212.
27. Schwartz, N., Mandel, D., Shlakhter, O., Coletta, J., Pessel, C., Timor-Tritsch I. E. and Salafia, C. M. *Placental morphology and chorionic surface vasculature at term are highly correlated with 3-dimensional sonographic measures at 11–14 weeks*. J. Ultrasound Med. **30** (2011), 1171–1178.
28. Sherman, T. F. *On connecting large vessels to small. The meaning of Murray's law*. J. Gen. Physiol. **78** (1981), 431–453.

29. Takahashi, T., Nagaoka, T., Yanagida, H., Saitoh, T., Kamiya, A., Hein, T., Kuo, L. and Yoshida, A. *A mathematical model for the distribution of hemodynamic parameters in the human retinal microvascular network*. J. Biorheol. **23** (2010), 77–86.
30. Uylings, H. B. M. *Optimization of diameters and bifurcation angles in lung and vascular tree structures*. Bull. Math. Biol. **399** (1977), 509–519.
31. Vankan, W. J., Huyghe, J. M., Janssen, J. D., Huson, A., Hacking, W. J. G. and Schreiner, W. *Finite element analysis of blood flow through biological tissue*. Int. J. Eng. Sci. **35** (1997), 375–385.
32. West, G. B., Brown, J. H. and Enquist, B. J. *The fourth dimension of life: Fractal geometry and allometric scaling of organisms*. Science **284** (1999), 1677–1679.
33. Xia, Q. *The formation of a tree leaf*. ESAIM Control Optim. Calc. Var. **13** (2007), 359–377.

Illustrating Optimal Control Applications with Discrete and Continuous Features

Suzanne Lenhart, Erin Bodine, Peng Zhong, and Hem Raj Joshi

Abstract In this paper, we present the basic idea of optimal control of models with discrete and continuous features. We first consider ordinary differential equation (ODE) models where we emphasize problems which are linear in the control and have discrete values for the optimal control. Three examples with ODEs illustrate how the bang-bang and singular controls could be handled. The first example utilizes a simple model with one ODE. The next two examples use systems of ODEs. One example comes from a mobile robot with one or more steerable drive wheels that steer together. The other example models species augmentation where two populations of the same species are modeled with a target/endangered population and a reserve population. Then we present an extension to an integrodifference model that is discrete in time and continuous in space. This optimal pest control problem is modeled by integrodifference equations and we illustrate how to construct the necessary conditions.

1 Introduction

In an optimal control problem, controls are adjusted in a system to achieve a goal. The underlying system can have a variety of types of equations, but in this survey

S. Lenhart (✉)
Department of Mathematics, University of Tennessee, Knoxville, TN 37996-1320, USA
e-mail: lenhart@math.utk.edu

E. Bodine
Department of Mathematics & Computer Science, Rhodes College, Memphis, TN 38112, USA
e-mail: bodinee@rhodes.edu

P. Zhong
Department of Ecology, Evolution and Natural Resources, Rutgers University, New Brunswick, NJ, USA
e-mail: zhongpeng85@gmail.com

H.R. Joshi
Mathematics and CS Department, Xavier University, Cincinnati, OH 45207-4441, USA
e-mail: joshi@xavier.edu

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5_9,

paper, we are first considering systems of ordinary differential equations and then we present an extension to integrodifference equations.

In control of a single differential equation, we denote $u(t)$ as the control and $x(t)$ as the state. The state function, $x(t)$, satisfies the differential equation modeling the scenario and the control u enters into the differential equation. Usually both $u(t)$ and $x(t)$ affect the goal, which is called the objective functional. We seek to find an optimal control and corresponding state to maximize (or minimize) our objective functional.

In this paper, we are first considering optimal control problems with the control u occurring linearly in the objective functional and in the differential equation. Our objective functional is

$$J(u) = \int_0^T \big(f_1(t,x) + u(t) f_2(t,x)\big)\, dt$$

with the control set $U = \{u : [0,T] \to [a,b], |, u \text{ is piecewise continuous}\}$ and the state equation

$$\begin{aligned} x'(t) &= g_1(t,x) + u(t) g_2(t,x), \\ x(0) &= x_0, \quad a \le u(t) \le b. \end{aligned}$$

Pontryagin and his colleagues developed the theory of optimal control of differential equations around 1950 in Moscow. Pontryagin had the idea of using the adjoint function $\lambda(t)$ to attach the differential equation on to the objective functional. This idea is similar to the idea of Lagrange multipliers in multivariable calculus in which the multiplier attaches a constraint to a function to be optimized.

Given the existence of such an optimal control and corresponding state function, Pontryagin differentiated the control-to-objective functional map

$$u \to J(u)$$

to derive the necessary conditions that an optimal control must satisfy. Pontryagin's Maximum Principle converts the problem of finding a control which maximizes the objective functional subject to the state differential equation and initial condition to the problem of optimizing the Hamiltonian pointwise. Here the Hamiltonian H is a linear function of the control u:

$$H\big(t, x(t), u(t), \lambda(t)\big) = f_1(t,x) + u(t) f_2(t,x) + \lambda(t)\big[g_1(t,x) + u(t) g_2(t,x)\big].$$

Notice how the adjoint function $\lambda(t)$ attaches the right hand side of the differential equation to the Hamiltonian. We briefly state the necessary conditions for this simple case [29, 37].

1.1 Necessary Conditions

If $u^*(t)$ and $x^*(t)$ are optimal, then there exists adjoint variable $\lambda(t)$ such that

$$H\big(t, x^*(t), u, \lambda(t)\big) \leq H\big(t, x^*(t), u^*(t), \lambda(t)\big),$$

at each time, for all u with values in $U = \{u : [0, T] \to [a, b] | u$ is piecewise continuous$\}$, and

$$\lambda'(t) = -\frac{\partial(H(t, x^*(t), u^*(t), \lambda(t))}{\partial x} \quad \text{and} \quad \lambda(T) = 0.$$

The final time condition on the adjoint variable is called the transversality condition. Note that we can also use bounded, Lebesgue measurable functions with lower and upper bounds for our control set.

When using that the Hamiltonian is maximized with respect to u at u^*, one considers

$$\frac{\partial H}{\partial u} = f_2(t, x) + \lambda(t) g_2(t, x),$$

which contains no information on the control due to H being linear in the control. Viewing $f_2 + \lambda g_2$ as the slope of H as a function of u, we can choose u easily if this slope is nonzero. Define $\psi(t, x, \lambda) = f_2(t, x) + \lambda(t) g_2(t, x)$, usually called the *switching function*. Then, our characterization of u^* is

$$u^*(t) = \begin{cases} a & \text{if } \psi(t, x^*, \lambda) < 0 \\ ? & \text{if } \psi(t, x^*, \lambda) = 0 \\ b & \text{if } \psi(t, x^*, \lambda) > 0. \end{cases}$$

If $\psi = 0$ cannot be sustained over an interval of time, but occurs only at finitely many points, then the control u^* is referred to as *bang-bang*. In this case, it is piecewise constant function, switching between only the upper and lower bound.

If $\psi(t) \equiv 0$ on some interval of time, we say u^* is *singular*. A characterization of u^* on this interval must be found using other information.

The times where the control switches from the upper bound to the lower bound (or vice versa) or to a singular control are called 'switching times.'

The emphasis here is to show how to obtain the necessary conditions and to use them to characterize an optimal control. In two examples, we will show how to handle the case of a singular control. The generalized Legendre-Clebsch is used for optimality of a singular control. One can see more details about Legendre-Clebsch conditions for optimality of singular controls in [6, 26, 33].

We refer to the reader to [10, 43] for showing the existence of an optimal control in more detail. For additional biological examples, see [1, 2, 20, 29, 42].

Here each example has some type of discrete feature. After a simple single equation example, our second example has a system of differential equations that comes from simple paths from the motion of a particular robot and the optimal controls

contain bang-bang and singular controls. Then augmentation of a species is modeled in a system of differential equations, and bang-bang controls occur in our numerical illustrations. Our last example involves integrodifferential equations, which are discrete in time and continuous in space. When a model is discrete in time, the order of events in the model is important. Since Pontryagin's Maximum Principle does not directly cover this system, we show how to derive the necessary conditions through differentiating the control-to-objective functional map.

2 Example: Simple Differential Equation Example

Consider an example with one state and one control [29].

$$\begin{aligned} &\max_u \int_0^2 \big(2x(t) - 3u(t)\big)\,dt \\ &\text{subject to} \quad x'(t) = x(t) + u(t), \qquad x(0) = 5, \\ &\qquad\qquad 0 \le u \le 2. \end{aligned}$$

The Hamiltonian is

$$H = 2x - 3u + \lambda(x + u).$$

Using the necessary conditions and transversality condition, the λ differential equation can be easily solved:

$$\lambda' = -\frac{\partial H}{\partial x} = -2 - \lambda, \qquad \lambda(2) = 0 \quad \Rightarrow \quad \lambda(t) = 2e^{2-t} - 2.$$

The switching function

$$\psi(t, x, \lambda) = \frac{\partial H}{\partial u} = \lambda - 3 = 2e^{2-t} - 5$$

is not constant on any subinterval. Thus our optimal control u^* is bang-bang, and using the idea of maximizing the Hamiltonian and thinking of the switching function as the slope of H with respect to the control, we obtain

$$\begin{aligned} u^*(t) = 0 \quad &\Leftrightarrow \quad \psi < 0 \quad \Leftrightarrow \quad e^{2-t} < \frac{5}{2} \quad \Leftrightarrow \quad t > 2 - \ln\left(\frac{5}{2}\right), \\ u^*(t) = 2 \quad &\Leftrightarrow \quad \psi > 0 \quad \Leftrightarrow \quad e^{2-t} > \frac{5}{2} \quad \Leftrightarrow \quad t < 2 - \ln\left(\frac{5}{2}\right). \end{aligned}$$

On $0 \le t < 2 - \ln(\frac{5}{2})$, $u = 2 \Rightarrow x' = x + 2$. Along with $x(0) = 5$, this gives $x(t) = -2 + 7e^t$. On $2 - \ln(\frac{5}{2}) < t \le 2$, $u = 0 \Rightarrow x' = x \Rightarrow x(t) = k_0 e^t$ for some constant k_0. We require continuity of the state function, and the expressions $-2 + 7e^t$

and $k_0 e^t$ must agree at $t = 2 - \ln(\frac{5}{2})$. We can find the constant: $k_0 = 7 - 5e^{-2}$. Hence, the optimal solution is

$$u^* = \begin{cases} 2 & \text{when } t < 2 - \ln(\frac{5}{2}), \\ 0 & \text{when } t > 2 - \ln(\frac{5}{2}), \end{cases} \qquad x^* = \begin{cases} 7e^t - 2 & \text{when } t \le 2 - \ln(\frac{5}{2}), \\ 7e^t - 5e^{2-t} & \text{when } t \ge 2 - \ln(\frac{5}{2}). \end{cases}$$

3 Example: Robot Paths

To illustrate a control problem that has optimal controls with singular and bang-bang features we consider a problem from Lenhart and Reister [40]. This application comes from a mobile robot with one or more steerable drive wheels that steer together. In this example, the robot is moving at its maximum speed when it receives orders to drive through a new goal configuration (position and orientation). We want to determine the time-optimal path from the current configuration to the goal in an unobstructed environment. Because the wheels of the mobile robot steer together, we can consider the vehicle to be a unicycle.

The basic equations of motion for a single wheel are:

$$x' = v \cos \phi, \tag{1}$$

$$y' = v \sin \phi \tag{2}$$

where the Cartesian coordinates (x, y) locate the point of contact between the wheel and the floor, the velocity is a positive constant v, and ϕ is the orientation angle of the plane of the wheel with respect to the x axis. We assume that the velocity of the wheel orientation is the control variable:

$$\phi' = u \tag{3}$$

where the magnitude of the orientation velocity is bounded: $|u| \le a$. The orientation velocity u is the control variable, motivated by the use of this for the HERMIES-III vehicle [39].

The optimization problem is to find a path for the control variable u that will move the system from the initial configuration to the final configuration and minimize the transition time. The objective functional is to minimize the final time T, such that

$$\big(x(0), y(0), \phi(0)\big) = (0, 0, 0)$$

and

$$\big(x(T), y(T), \phi(T)\big) = (x_f, y_f, \phi_f).$$

The objective functional can be written as:

$$\min_{u,T} \int_0^T 1 \, dt.$$

The subscripts u, T indicate that the control and the final time are to be chosen.

Now we need to show how to form the necessary conditions when there are more than one state variable. There is one adjoint function corresponding to each state variable. The right hand side of each differential equation is attached to the Hamiltonian by the corresponding adjoint function. We have three state equations and three corresponding adjoint functions, so that the Hamiltonian becomes,

$$H = 1 + \lambda_1(v\cos\phi) + \lambda_2(v\sin\phi) + \lambda_3(u).$$

Our system of adjoint variables are:

$$\lambda_1' = \frac{\partial H}{\partial x} = 0, \tag{4}$$

$$\lambda_2' = \frac{\partial H}{\partial y} = 0, \tag{5}$$

$$\lambda_3' = \frac{\partial H}{\partial \phi} = \lambda_1 v\sin\phi - \lambda_2 v\cos\phi. \tag{6}$$

Note that the adjoint variables do not have any boundary conditions since the state variables each have two boundary conditions. Since the final time is to be determined, we have an extra condition that the Hamiltonian is 0 at the final time.

The existence for an optimal control can be shown with standard results [10], so we concentrate on explaining the necessary conditions and the structure of the optimal controls.

Pontryagin's Maximum Principle gives that the optimal set of control variables maximizes the Hamiltonian H. Because the Hamiltonian is linear in the control variable u, the optimal solution is bang-bang or singular or a combination. Note that the coefficient of u in the Hamiltonian is λ_3, so the switching function is λ_3.

When $\lambda_3(t) > 0$, the optimal control is at its lower bound $-a$ and when $\lambda_3(t) < 0$, the optimal control is at its upper bound a. When λ_3 is zero for a nontrivial subinterval of our time interval, the optimal control is singular and may have intermediate values.

We will show that the optimal path consists of a sequence of arcs and line segments. When an optimal control is bang-bang on a subinterval, the control is at its upper or lower bound and the path is an arc of a circle. When an optimal control is singular on a subinterval, the control is zero there and the path is a line segment.

We consider first the case where the control is singular. For $\lambda_3 = 0$ on a subinterval, using the state differential equations gives,

$$\lambda_3' = \lambda_1 x' - \lambda_2 y'.$$

This equation can be integrated to yield:

$$\lambda_3 = \mu_1 y - \mu_2 x + \mu_3. \tag{7}$$

Since λ_1 and λ_2 are constants, we use the constants, μ_1 and μ_2, for their values. In this singular case, $\lambda_3(t) = 0$, (7) defines a line. On this line, orientation angle ϕ

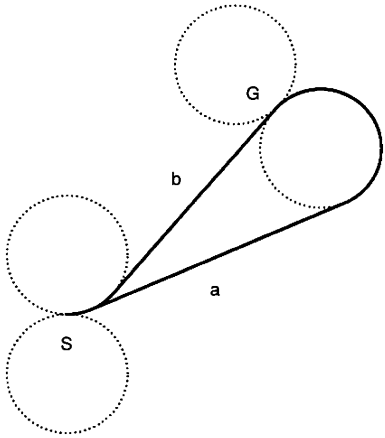


Fig. 1 Time-optimal paths when the final orientation is either 45° or 225°

is a constant and $\phi' = 0$. Thus, $u^* = 0$ on any subinterval where the singular case holds.

We consider next the case where the control is bang-bang. We will consider a subinterval where λ_3 is positive (or negative) and u is a constant, say a. We shall show that the path is an arc of a circle on such an interval. If the control is constant initially, (3) can be integrated to yield: $\phi = at$ and we can also find:

$$x = (v/a)\sin\phi, \tag{8}$$

$$y = (v/a)(1 - \cos\phi). \tag{9}$$

Thus, the path is an arc of a circle with its radius being the ratio of the wheel velocity, and the steering velocity, $R = v/a$. Its center is $(0, q) : x^2 + (y - q)^2 = R^2$, where $q = v/a$. A similar argument works for cases where such an interval does not start at $(0, 0, 0)$. All circles or arcs of circles on optimal trajectories will have the same radius.

Given a specific final position, one can compute the optimal control and corresponding state trajectory from fitting together the above equations and cases. One can show that the optimal state trajectories are paths composed of three arcs or arc-line-arc (or subpaths of those). Thus the range values of the optimal controls are discrete, $-a, 0, a$. We refer the reader to [40] for more details about the numerical calculations, how the robot actually followed such paths and the relationship between minimal time paths and minimal length paths.

In Fig. 1, the starting position (x, y) is denoted by S and the initial orientation angle is O, The final position is denoted by G, and the final orientation angle is either 45° or 225°. The path to the final orientation 45° (path b) starts with an arc (bang-bang control) and ends in a line segment (singular control).

A constant-speed vehicle cannot make tight maneuvers. Figure 2 shows the path required to turn around [the final state values are $(0, 0, 180°)$]. The path consists of three arcs, with optimal control switching between a and $-a$. Note that the path could be traveled in either direction.

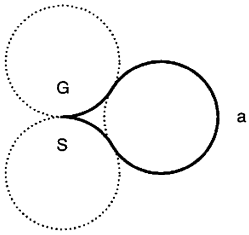


Fig. 2 Time-optimal path to rotate by 180°

While time-optimal paths at constant velocity are similar to minimal length paths, they have one significant difference. A minimal length path can reverse direction at a point (have a cusp). Our minimal time paths cannot have discontinuities in tangent vectors due to the structure of the state system. If a vehicle needs to perform tight maneuvers, it should not move at constant speed; it should slow down.

4 Example: Augmentation

Consider two populations of the same species: N, a target/endangered population, and R, a reserve population. We assume that, at the initial time, the endangered population is declining due to small population size, i.e. there is some critical population size, below which the population declines to extinction. For the reserve population to be a viable source for harvesting individuals with which to augment the target population, it must be growing at the initial time, but it is also assumed to have a lower threshold for population growth which could be crossed due to over harvesting. Therefore, each of these populations are assumed to grow according to a normalized Allee effect model [25], in which aK_N and bK_R are the critical population sizes for growth for the target and reserve populations, respectively. The control u is the rate at which individuals are moved from the reserve population to the target population. Thus, the populations are modeled by the equations

$$\frac{dN}{dt} = rN\left(1 - \frac{N}{K_N}\right)\left(\frac{N}{K_N} - a\right) + uR,$$
$$\frac{dR}{dt} = sR\left(1 - \frac{R}{K_R}\right)\left(\frac{R}{K_R} - b\right) - uR$$

where r and s are the intrinsic growth rates of N and R, respectively, K_N and K_R are the carrying capacities of N and R, respectively, aK_N and bK_R are the thresholds for population growth for N and R, respectively, and u is the rate of augmentation (translocating individuals from the reserve population to the target population). It is assumed that there is no net loss of population due to augmentation efforts. Additionally, the intrinsic growth rates r and s can be different in value. The difference could be due to different environmental conditions where the target and reserve populations are located, or could be due to difference in the underlying

genetics of the two populations. The latter would be especially true if the target and reserve populations were different subspecies of the same species.

The objective of augmentation is to maximize the target population at a given final time while minimizing the cost. This assumes there is cost associated with translocating an individual from the reserve population. The cost is assumed to be a linear function of the augmentation and the total population is to be maximized at the final time, with different relative weights applied to the reserve and target populations. It is assumed that it is not as important to maximize the reserve population as it is the target population by the final time. Thus, the optimal control formulation is

$$\max_{u\in U}\left[x(t_1)+By(t_1)-\int_{t_0}^{t_1}A_2u(t)\,dt\right] \tag{10}$$

where

$$U=\left\{u:[t_0,t_1]\to[0,1]\mid u \text{ Lebesgue measurable}\right\} \tag{11}$$

and

$$x'(t)=rx(1-x)(x-a)+puy, \qquad x(t_0)=x_0 \quad \text{where } 0<x_0<a<1, \tag{12}$$

$$y'(t)=sy(1-y)(y-b)-uy, \qquad y(t_0)=y_0 \quad \text{where } 0<b<y_0<1 \tag{13}$$

and $a, b, t_0, t_1, x_0, y_0, r, s, A_2$, and B are all non-negative constants, with $0\le B\le 1$. We have rescaled the two populations with respect to their carrying capacities ($x\equiv\frac{N}{K_N}$ and $y\equiv\frac{R}{K_R}$) and we denote $p=K_R/K_N$, i.e. the ratio of the reserve carrying capacity to the target carrying capacity. It is assume that the target population x has an initial density x_0 below its minimum threshold for growth $0<a<1$, and that the reserve population y has an initial density y_0 above its minimum threshold for growth $0<b<1$. Thus, $x_0<a$ and $y_0>b$. This optimal control problem has been examined in [5] with a quadratic the cost term in the objective functional. Here, we demonstrate the effects of a linear objective functional on the optimal control and corresponding states.

4.1 Necessary Conditions for Augmentation Example

We use Pontryagin's Maximum Principle [37] along with the generalized Legendre-Clebsch Condition [6, 26] to derive the necessary conditions that an optimal control and its corresponding states must satisfy. For specific applications of this condition, see [7, 8, 27, 28, 35]. From the boundedness of the controls and the states, we have that there exists an optimal control [5].

Since this control problem is linear in the control, there is a possibility of bang-bang and singular controls.

Theorem 1 *Given an optimal control u^* and the corresponding solutions to the state system given in (12)–(13), there exist adjoint variables λ_x and λ_y satisfying equations*

$$\frac{d\lambda_x}{dt} = r\lambda_x\big(3\big(x^*\big)^2 - 2(1+a)x^* + a\big), \tag{14}$$

$$\frac{d\lambda_y}{dt} = s\lambda_y\big(3\big(y^*\big)^2 - 2(1+b)y^* + b\big) - p\lambda_x u^* + \lambda_y u^*, \tag{15}$$

$$\lambda_x(t_1) = 1, \tag{16}$$

$$\lambda_y(t_1) = B. \tag{17}$$

Furthermore, u^* is characterized by

$$u^*(t) = \begin{cases} 0 & \text{if } \psi(t) < 0 \\ 1 & \text{if } \psi(t) > 0 \end{cases} \tag{18}$$

where

$$\psi(t) = -A_2 + \big[p\lambda_x(t) - \lambda_y(t)\big]y(t). \tag{19}$$

Let

$$\begin{aligned} F(x, y, \lambda_x, \lambda_y) = {} & pr\lambda_x'\big[3x^2 - 2(1+a)x + a\big] \\ & - 2pr^2\lambda_x xy[3x - 1 - a]\big[x^2 - (1+a)x + a\big] \\ & - s^2\lambda_y y^2\big[3y^2 - 2(1+b)y + b\big][2y - 1 - b] \\ & - sy\big[y^2 - (1+b)y + b\big]\big[2p\lambda_x' - 4\lambda_y y(3y - 1 - b) \\ & - ps\lambda_x\big(3y^2 - 2(1+b)y + b\big)\big], \end{aligned} \tag{20}$$

and

$$\begin{aligned} G(x, y, \lambda_x, \lambda_y) = {} & p^2 r\lambda_x y^2 - pr\lambda_x y\big[3x^2 - 2(1+a)x + a\big] \\ & + ps\lambda_x y\big(5y^2 - 3(1+b)y + b\big) + s\lambda_y y^2(4y - 1 - b). \end{aligned} \tag{21}$$

If $\psi(t) = 0$ on a non-empty open interval $(t_\alpha, t_\beta) \subset (t_0, t_1)$, then the singular control

$$u_s(t) = -\frac{F(x, y, \lambda_x, \lambda_y)}{G(x, y, \lambda_x, \lambda_y)} \tag{22}$$

is optimal on $[t_\alpha, t_\beta]$ if the inequalities

$$G(x, y, \lambda_x, \lambda_y) > 0 \quad \text{and} \quad 0 \le -\frac{F(x, y, \lambda_x, \lambda_y)}{G(x, y, \lambda_x, \lambda_y)} \le 1 \tag{23}$$

hold on the interval $[t_\alpha, t_\beta]$.

Proof Suppose u^* is an optimal control with corresponding states x^*, y^*. Using Pontryagin's Maximum Principle, the Hamiltonian is formed

$$H = -A_2 u + \lambda_x \big(rx(1-x)(x-a) + puy\big) + \lambda_y \big(sy(1-y)(y-b) - uy\big) \quad (24)$$

and the adjoint equations are

$$\frac{d\lambda_x}{dt} = -\frac{\partial H}{\partial x} \quad \text{and} \quad \frac{d\lambda_y}{dt} = -\frac{\partial H}{\partial y}$$

yielding (14) and (15) with transversality conditions

$$\lambda_x(t_1) = 1 \quad \text{and} \quad \lambda_y(t_1) = B.$$

Define $\psi(t)$ as the coefficient of u in the Hamiltonian

$$\psi(t) = \frac{\partial H}{\partial u} = -A_2 + \big[p\lambda_x(t) - \lambda_y(t)\big]y(t)$$

and we call $\psi(t)$ the *switching function*. In this case, if we maximize the Hamiltonian with respect to u, the characterization of the optimal control at time t is

$$u^*(t) = \begin{cases} 0 & \text{if } \psi < 0 \\ 1 & \text{if } \psi > 0. \end{cases}$$

We investigate the case where $\psi = 0$, i.e. the case of the singular control. Let u_s denote the singular control. Thus, for times in $[t_0, t_1]$ where $\psi \neq 0$ we have a bang-bang control, meaning the optimal control takes values at its lower or upper bounds. When the switching function is zero on a non-trivial interval, we have the singular case. In order to determine whether u_s, is optimal over the non-trivial interval, we check the generalized Legendre-Clebsch condition.

The generalized Legendre-Clebsch Condition [8, 26] states that if the singular control is maximizing on the interval $[t_\alpha, t_\beta]$ then

$$(-1)^q \frac{\partial}{\partial u} \frac{d^{2q}}{dt^{2q}} \psi(t) \leq 0,$$

where q is the least integer such that

$$\frac{\partial}{\partial u} \frac{d^{2q}}{dt^{2q}} \psi(t) \neq 0.$$

To find u_s, we take the first time derivative of the switching function,

$$\psi'(t) = \big[p\lambda_x' - \lambda_y'\big]y + [p\lambda_x - \lambda_y]y'. \quad (25)$$

Substituting the right hand sides of y' and λ_y' (which contain control terms) and simplifying, we obtain

$$\psi'(t) = \lambda_x' y - s\lambda_y\big[3y^3 - 2(1+b)y^2 + by\big] + p\lambda_x uy - \lambda_y uy$$

$$- ps\lambda_x\big[y^3 - (1+b)y^2 + by\big] - p\lambda_x uy + s\lambda_y\big[y^3 - (1+b)y^2 + by\big] + \lambda_y uy.$$

The terms with the control u cancel each other out, leaving no u terms in this first derivative. Simplifying,

$$\psi'(t) = p\lambda'_x y - ps\lambda_x\big[y^3 - (1+b)y^2 + by\big] - s\lambda_y\big[2y^3 - (1+b)y^2\big]. \quad (26)$$

Next, take the second time derivative of the switching function,

$$\begin{aligned}\psi''(t) &= p\lambda''_x y + p\lambda'_x y' - ps\lambda'_x\big[y^3 - (1+b)y^2 + by\big] \\ &\quad - ps\lambda_x\big[3y^2 - 2(1+b)y + b\big]y' \\ &\quad - s\lambda'_y\big[2y^3 - (1+b)y^2\big] - s\lambda_y\big[6y^2 - 2(1+b)y\big]y' \\ &= p\lambda''_x y - ps\lambda'_x\big[y^3 - (1+b)y^2 + by\big] - s\lambda'_y\big[2y^3 - (1+b)y^2\big] \\ &\quad + y'\big[p\lambda'_x - 6s\lambda_y y^2 + 2(1+b)s\lambda_y y - 3ps\lambda_x y^2 + 2p(1+b)s\lambda_x y \\ &\quad - pbs\lambda_x\big].\end{aligned}$$

Substituting in the right hand sides of y', λ'_y, and λ''_x (which contain control terms) and simplifying, we obtain

$$\begin{aligned}\psi''(t) &= pr\lambda'_x\big[3x^2 - 2(1+a)x + a\big] - 2pr^2\lambda_x xy[3x - 1 - a]\big[x^2 - (1+a)x + a\big] \\ &\quad - s^2\lambda_y y^2\big[3y^2 - 2(1+b)y + b\big][2y - 1 - b] - sy\big[y^2 - (1+b)y + b\big] \\ &\quad \times \big[2p\lambda'_x - 4\lambda_y y(3y - 1 - b) - ps\lambda_x\big(3y^2 - 2(1+b)y + b\big)\big] \\ &\quad + u\big[p^2 r\lambda_x y^2 - p\lambda'_x y + ps\lambda_x y\big(5y^2 - 3(1+b)y + b\big) \\ &\quad + s\lambda_y y^2(4y - 1 - b)\big].\end{aligned}$$

Since this second time derivative has u terms, we have

$$\frac{\partial}{\partial u}\frac{d^2}{dt^2}\psi(t) \neq 0$$

on the interval $[t_\alpha, t_\beta]$. So we take the partial derivative with respect to u of $\frac{d^2\psi}{dt^2}$

$$\begin{aligned}\frac{\partial}{\partial u}\frac{d^2\psi}{dt^2} &= p^2 r\lambda_x y^2 - p\lambda'_x y + ps\lambda_x y\big(5y^2 - 3(1+b)y + b\big) \\ &\quad + s\lambda_y y^2(4y - 1 - b).\end{aligned}$$

Substitute in the right hand side of λ'_x to get

$$\begin{aligned}\frac{\partial}{\partial u}\frac{d^2\psi}{dt^2} &= p^2 r\lambda_x y^2 - pr\lambda_x y\big[3x^2 - 2(1+a)x + a\big] \\ &\quad + ps\lambda_x y\big(5y^2 - 3(1+b)y + b\big) + s\lambda_y y^2(4y - 1 - b). \quad (27)\end{aligned}$$

The generalized Legendre-Clebsch Condition tells us that the singular control is maximizing when

$$p^2 r\lambda_x y^2 - pr\lambda_x y\big[3x^2 - 2(1+a)x + a\big] + ps\lambda_x y\big(5y^2 - 3(1+b)y + b\big) + s\lambda_y y^2(4y - 1 - b) \geq 0. \tag{28}$$

Notice the left hand side of inequality (28) is $\frac{\partial}{\partial u}\frac{d^2\psi}{dt^2}$. We can derive an equation for the singular control u_s by solving $\psi''(t) = 0$ for u. If we do this, we get

$$u_s = -\frac{F(x, y, \lambda_x, \lambda_y)}{G(x, y, \lambda_x, \lambda_y)}$$

where u_s is the singular solution, and

$$\begin{aligned} F(x, y, \lambda_x, \lambda_y) &= pr\lambda_x'\big[3x^2 - 2(1+a)x + a\big] - 2pr^2\lambda_x xy[3x - 1 - a]\big[x^2 - (1+a)x + a\big] \\ &\quad - s^2\lambda_y y^2\big[3y^2 - 2(1+b)y + b\big][2y - 1 - b] \\ &\quad - sy\big[y^2 - (1+b)y + b\big]\big[2p\lambda_x' - 4\lambda_y y(3y - 1 - b) \\ &\quad - ps\lambda_x\big(3y^2 - 2(1+b)y + b\big)\big], \end{aligned} \tag{29}$$

and

$$G(x, y, \lambda_x, \lambda_y) = p^2 r\lambda_x y^2 - pr\lambda_x y\big[3x^2 - 2(1+a)x + a\big] + ps\lambda_x y\big(5y^2 - 3(1+b)y + b\big) + s\lambda_y y^2(4y - 1 - b). \tag{30}$$

□

Notice, if $\psi(t) = 0$ on a non-empty open interval $(t_\alpha, t_\beta) \subset (t_0, t_1)$ and Inequalities (23) do *not* hold, then we cannot say whether the singular control u_s is optimal over the interval $[t_\alpha, t_\beta]$.

4.2 Numerical Illustrations for Augmentation Example

The optimal control can be numerically calculated under various parameter sets using a forward-backward sweep method [29] using 4th order Runge-Kutta to solve the state equations (12)–(13) and their corresponding adjoint equations (14)–(15). We check the Legendre-Clebsch condition (23) for any t such that $\psi(t) = 0$. To check this condition numerically, we check if $-\varepsilon < \psi(t) < \varepsilon$, for small ε.

The forward-backward sweep method makes an initial guess for the control u and then solves the state equations (12)–(13) forward in time using the Runge-Kutta

method with the initial conditions (x_0 and y_0). Then, using the state values, the adjoint equations (14)–(15) are solved backwards in time using the Runge-Kutta method with the transversality conditions (16)–(17). At this point, the optimal control is updated. For each time t, if $\psi(t) \neq 0$, then $u(t)$ is updated using (18). However, if $\psi(t) = 0$, then $u(t)$ is updated using (22) provided that inequalities (23) hold. The updated control replaces the initial control and the process is repeated until the successive iterates of control values are sufficiently close. The convergence of such an iterative method is based on the work of Hackbush [14]. Other examples using this method can be found in [11, 16, 41].

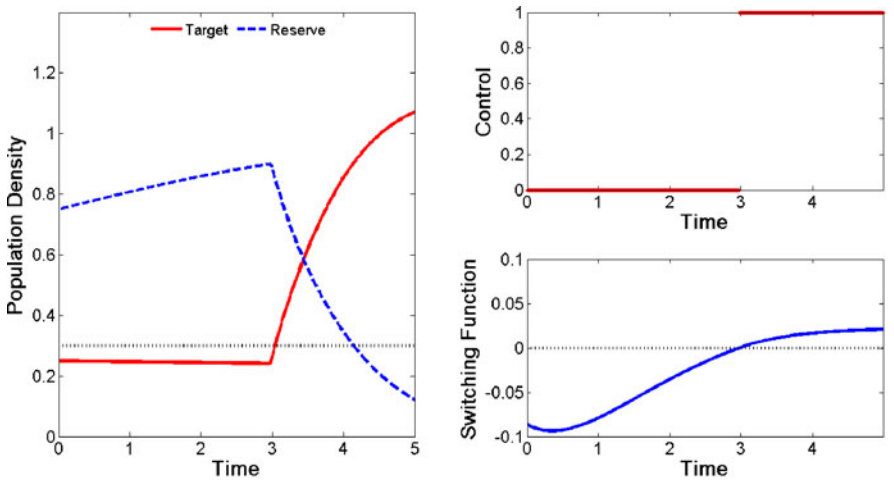
In considering various parameter scenarios, the parameter constraints $x_0 < a$ and $y_0 > b$, must be included. For the examples we consider here, we take the minimum threshold for growth for both the target and reserve populations to be 0.3 (that is 30 % of each population's carrying capacity), and $x_0 = 0.25$ and $y_0 = 0.75$. Thus, the target population is starting just below its minimum threshold for growth and the reserve population is starting well above its minimum threshold for growth. Additionally, each scenario assumes that the intrinsic growth rate of the reserve population s is greater than the intrinsic growth rate of the target population r, and in each scenario $r = 0.3$.

We first consider the impact of varying the intrinsic growth rate of the reserve population, s. In these scenarios we take $a = 0.3$, $b = 0.3$, $r = 0.3$, $p = 1$, $x_0 = 0.25$, $y_0 = 0.75$, and $A_2 = 0.001$.

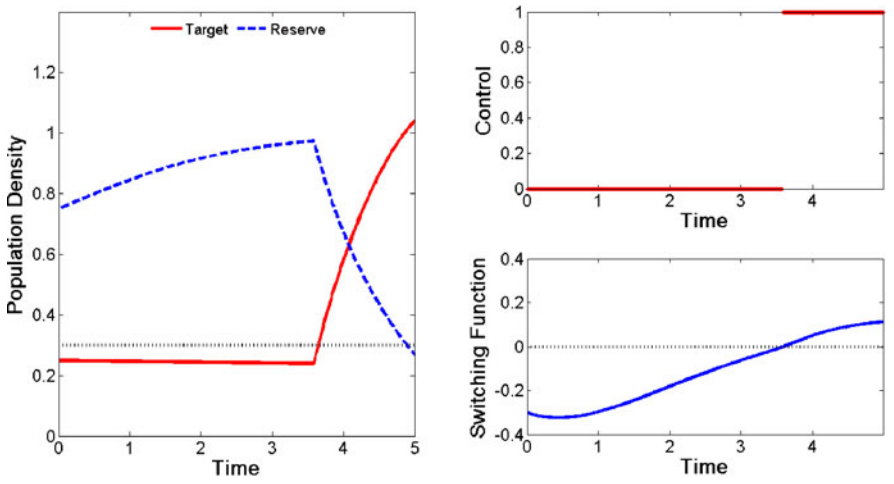
In the first scenario we take $s = 0.7$ and $B = 0$ (see Fig. 3(a)), and in the second scenario we take $s = 1.2$ and $B = 0$ (see Fig. 3(b)). Since $B = 0$ in each case, no importance is given to maximizing the reserve population at the final time. In each scenario, the qualitative strategy for augmentation is to do nothing for an initial interval of time, then to apply maximum control until the final time. However, when the intrinsic growth rate of the reserve population is lower, the optimal strategy switches from no augmentation to applying maximum augmentation sooner. When $s = 0.7$, the time of the switch is $t = 2.97$, whereas when $s = 1.2$ the time of the switch is $t = 3.58$. When the intrinsic growth rate is lower, the reserve population cannot replenish its population as quickly when being harvested. Thus, in order to maximize the target population, harvesting of the reserve population must start sooner in the case when the reserve intrinsic growth rate is lower.

Notice also that in both scenarios, the reserve population is harvested such that it falls below its minimum threshold for growth, $b = 0.3$ (see black dotted line in Fig. 3), by the final time. Additionally, in both scenarios, the target population is augmented such that it rises above its carrying capacity at the final time (i.e., $x(t_1) > 1$). We address the problem of over-harvesting the reserve population by increasing the importance of maximizing the reserve population at the final time, i.e. increasing the value of B.

Thus, in this next set of scenarios we vary the intrinsic growth rate of the reserve population while giving importance to maximizing the reserve population at the final time. In one scenario we take $s = 0.7$ and $B = 0.75$ (see Fig. 4(a)), and in the other scenario we take $s = 1.2$ and $B = 0.75$ (see Fig. 4(b)). Since $B = 0.75$ in each case, maximizing the reserve population at the final time is 75 % as important as maximizing the target population at the final time.



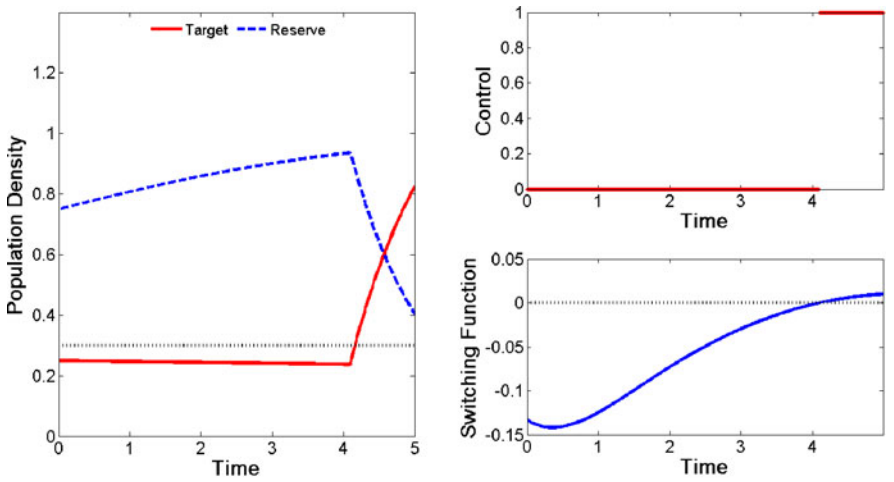
(a) Lower intrinsic growth rate of s = 0.7.



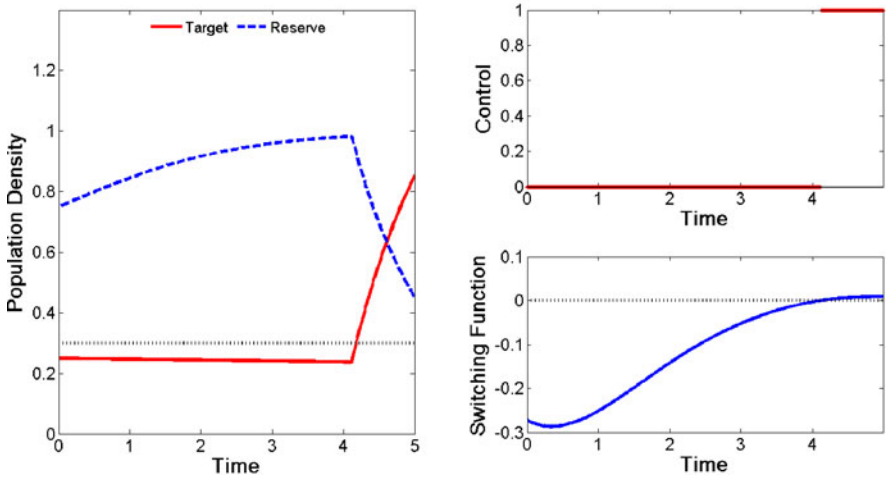
(b) Higher intrinsic growth rate of s = 1.2.

Fig. 3 Scenario for optimal control of augmentation with $B = 0$, and where $s = 0.7$ in (**a**) $s = 1.2$ in (**b**). The graphs on the left show the density of the target (*red*) and reserve (*blue dashed*) populations. The *black dotted line* represents the minimum threshold for growth, $a = b = 0.3$. The *top right graphs* show the optimal control over time, and the *bottom right graphs* show the switching function over time

Again, we see the qualitative augmentation strategy in each case is to do nothing for an initial interval of time, then to apply full control until the final time. With these two scenarios, we do not see a great difference in the time at which the optimal strategy switches from no augmentation to maximum augmentation. When $s = 0.7$, the time of the switch is $t = 4.09$, whereas when $s = 1.2$, the time of the switch



(a) Lower intrinsic growth rate of s = 0.7.



(b) Higher intrinsic growth rate of s = 1.2.

Fig. 4 Scenario for optimal control of augmentation with $B = 0.75$, and where $s = 0.7$ in (**a**) $s = 1.2$ in (**b**). The graphs on the *left* show the density of the target (*red*) and reserve (*blue dashed*) populations. The *black dotted line* represents the minimum threshold for growth, $a = b = 0.3$. The *top right graphs* show the optimal control over time, and the *bottom right graphs* show the switching function over time

is 4.11. If the unit of time were one year (i.e. $t = 1$ is one year after the initial time), then the difference between $t = 4.09$ and $t = 4.11$ would be a difference of about 7 days.

Notice that in both scenarios, because of the increased importance of maximizing the reserve populations (as compared to the scenarios in Fig. 3), the times at

which the optimal strategy switches from no augmentation to maximum augmentation are later, and thus less of the reserve population is translocated into the target population. The result on the reserve population is that it stays above its minimum threshold for growth, $b = 0.3$ (black dotted line in Fig. 4), at the final time. The result on the target population is that it is not augmented such that it rises above its carrying capacity at the final time (i.e. $x(t_1) < 1$). Indeed this is an ideal scenario for both populations since both the target and reserve populations are between their minimum threshold for growth and their carrying capacities at the final time (i.e. $a < x(t_1) < 1$ and $b < y(t_1) < 1$).

It should be noted that, given the scenarios in Fig. 4, if B is increased to 1, then the optimal augmentation strategy becomes to do no augmentation for the entire time period. When $B = 1$, it is just as important to maximize the reserve population at the final time as it is to maximize the target population at the final time. As expected, in this case the reserve population increases over the time period while the target population declines towards extinction.

Using the generalized Legendre-Clebsch Condition, we were able to obtain necessary conditions for u_s being optimal over an interval (see inequalities (23)). It should be noted that in the numerical scenarios (those shown here and many others investigated), the switching function was never identically zero over a nontrivial interval. Thus, we did not encounter a parameter set where the singular control could have occurred. Again, we note that the range values of the optimal control are discrete, 0 or 1.

What is most important to conclude from the different parameter scenarios is that these numerical simulations can tell natural resource managers the best they can do given a certain scenario, and what augmentation strategy will yield that "best" outcome. In comparing different parameter scenarios, there were some optimal augmentation strategies that differed only by a few days in determining when to start applying maximum augmentation effort. In comparing other scenarios, we found completely different qualitative optimal augmentation strategies. It is important for natural resource managers to be able to explore what parameter scenarios lead to these drastically different augmentation strategies.

5 Example: Integrodifference Equations

Integrodifference equations are frequently used to model population with distinct growth and dispersal stages [22, 24, 31]. These equations are often used to model invasive species since the kernels for spatial dispersal can represent various types and speeds of spread. In this model, the time steps are discrete and the spatial variables are continuous.

Pest control is the application in this integrodifference example. We note that the pest population and invasive species dynamics and control are modeled by many types of equations [4, 13, 21, 45, 46]. We refer the reader to survey articles on invasive species and control using an economic viewpoint [9, 36]. The book by Hawkins and Cornell [15] gives basic background on approaches to biological control.

The integrodifference model with linear growth is:

$$N_{t+1}(x) = \int_{\Omega} k(x, y)\big(1 - \alpha_t(y)\big) r N_t(y)\, dy \tag{31}$$

where $t = 0, 1, \ldots, T-1$.

Our model represents a pest population with a dispersal kernel $k(x, y)$, and a harvest control is used to slow the spread of this population. The state variable N and the control α are represented by

$$N = N(\alpha) = \big(N_0(x), N_1(x), \ldots, N_T(x)\big), \quad \alpha = \big(\alpha_0(x), \alpha_1(x), \ldots, \alpha_{T-1}(x)\big),$$

where x is the spatial variable in a domain Ω. The growth rate r is a positive constant. The order of events in the model is: growth, harvest, and dispersal. It seems reasonable to harvest before dispersal. But other order of events are possible and would give different results [30].

Assume Ω is a bounded domain in R^n. The non-negative initial distribution $N_0(x)$ is given in $L^\infty(\Omega)$. Assume $\alpha_t(x)$ is Lebesgue measurable and $0 \le \alpha_t(x) \le M < 1$ for all $t = 0, 1, \ldots, T-1$ and $x \in \Omega$.

Our goal is to minimize the objective functional $J(\alpha)$,

$$J(\alpha) = \sum_{t=1}^{T} \int_{\Omega} N_t(y)\, dy + \sum_{t=0}^{T-1} \int_{\Omega} \frac{B_t}{2} \big(\alpha_t(y)\big)^2\, dy. \tag{32}$$

We seek to minimize the pest population and the cost of applying harvesting control through all time steps. Since the initial population N_0 is given, it is not included in the objective functional. Our cost function is non-linear, and we will be dealing with a simple quadratic cost. The coefficient B_t is a positive weight factor that balances the two parts of the objective functional. We look for the control α^* that minimizes J, i.e.:

$$J\big(\alpha^*\big) = \min_{\alpha \in U} J(\alpha)$$

where the control set is $U = \{\alpha \in (L^\infty(\Omega))^T | 0 \le \alpha_t(x) \le M, t = 0, 1, \ldots, T-1\}$ for $M < 1$.

Assume that the kernels are bounded and measurable such that

$$\int_{\Omega} k(x, y)\, dy \le 1$$

for all $x \in \Omega$, and

$$0 \le k(x, y) \le \Gamma$$

for $(x, y) \in \Omega \times \Omega$ and $\Gamma < 1$.

Note that integrodifference equations do not have boundary conditions on $\partial\Omega$ like in reaction-diffusion equations. No individuals enter the population from outside Ω. If x is near the $\partial\Omega$, the individuals who disperse outside $\partial\Omega$ are not counted in our population in Ω.

Also note that the boundedness of the kernel and the initial distribution gives that the states, N_t for $t = 1, \ldots, T$, are non-negative and bounded above.

The existence of an optimal control can be proved by using weak convergence of a minimizing sequence of controls and pointwise convergence of the corresponding state sequence. See [18] for similar results. The emphasis in presenting this example is showing how to find the necessary conditions.

5.1 Characterization of an Optimal Control

Differentiating the map $\alpha \to J(\alpha)$ is a key feature in deriving necessary conditions. To characterize an optimal control, we must differentiate the map $\alpha \to J(\alpha)$, which requires first the differentiation of the solution map $\alpha \to N(\alpha)$. The directional derivative of this solution map is called the sensitivity of the state with respect to the control.

Theorem 2 *The mapping $\alpha \in U \to N \in (L^\infty(\Omega))^{T+1}$ is differentiable in the following sense: For any $\alpha \in U$ and $l \in (L^\infty(\Omega))^T$, such that $(\alpha + \varepsilon l) \in U$ for ε small, where $N^\varepsilon = N(\alpha + \varepsilon l)$ and $N = N(\alpha)$, there exists a sensitivity $\psi \in (L^\infty(\Omega))^{T+1}$ such that*

$$\frac{N_t^\varepsilon(x) - N_t(x)}{\varepsilon} \rightharpoonup \psi_t(x)$$

weakly in $L^2(\Omega)$, as $\varepsilon \to 0$ for each t. Also ψ, depending on N, α and l, satisfies:

$$\begin{aligned} \psi_{t+1}(x) &= \int_\Omega rk(x,y)\big[\big(1-\alpha_t(y)\big)\psi_t(y) - l_t(y)N_t(y)\big]\,dy \\ \psi_0(x) &= 0, \end{aligned} \tag{33}$$

for $t = 0, 1, \ldots, T$ and $x \in \Omega$.

Proof We form the difference quotient for the directional derivative of N with respect to α in the direction l:

$$\begin{aligned} &\frac{N_{t+1}^\varepsilon(x) - N_{t+1}(x)}{\varepsilon} \\ &= \frac{1}{\varepsilon}\int_\Omega rk(x,y)\big[\big(1-\alpha_t(y)\big)\big(N_t^\varepsilon(y) - N_t(y)\big) - \varepsilon l_t(y)N_t^\varepsilon(y)\big]\,dy \\ &= \int_\Omega rk(x,y)\left[\big(1-\alpha_t(y)\big)\frac{(N_t^\varepsilon(y) - N_t(y))}{\varepsilon} - l_t(y)N_t^\varepsilon(y)\right]dy. \end{aligned}$$

Using $N_0^\varepsilon = N_0$ and $\psi_0 \equiv 0$, we have

$$\frac{N_1^\varepsilon(x) - N_1(x)}{\varepsilon} = -\int_\Omega rk(x,y)l_0(y)N_0(y)\,dy = -\psi_1(x). \tag{34}$$

So

$$\left|\frac{N_1^\varepsilon(x) - N_1(x)}{\varepsilon}\right| \le C_1 \quad \text{for all } x \in \Omega.$$

And then by iteration,

$$\left|\frac{N_t^\varepsilon(x) - N_t(x)}{\varepsilon}\right| \le C_t \quad \text{for all } x \in \Omega, t = 1, 2, \ldots, T.$$

From the a priori estimate, we have

$$\frac{N_t^\varepsilon(x) - N_t(x)}{\varepsilon} \rightharpoonup \psi_t(x) \quad \text{weakly in } L^2(\Omega).$$

$$\frac{N_2^\varepsilon(x) - N_2(x)}{\varepsilon}$$
$$= \int_\Omega rk(x, y)\big(1 - \alpha_1(y)\big)\frac{N_1^\varepsilon(y) - N_1(y)}{\varepsilon}\,dy - \int_\Omega rk(x, y) l_1(y) N_1^\varepsilon(y)\,dy.$$

From the uniform boundedness of the states, we have $N_t^\varepsilon \rightharpoonup N_t$ weakly in $L^2(\Omega)$ for any $t = 1, 2, \ldots, T$. Together with the L^2 boundedness of the kernel function and l_t, we can pass the limit and get pointwise convergence of the quotient,

$$\begin{aligned}\frac{N_2^\varepsilon(x) - N_2(x)}{\varepsilon} &\to \int_\Omega rk(x, y)\big(1 - \alpha_1(y)\big)\psi_1(y)\,dy - \int_\Omega rk(x, y) l_1(y) N_1(y)\,dy \\ &= \psi_2(x). \end{aligned} \tag{35}$$

We can easily prove that

$$\left|\frac{N_2^\varepsilon(x) - N_2(x)}{\varepsilon} - \psi_2(x)\right|^2 \le C \quad \text{and} \quad \left|\frac{N_2^\varepsilon(x) - N_2(x)}{\varepsilon} - \psi_2(x)\right|^2 \to 0.$$

From dominated convergence theorem, the pointwise convergence in (35) becomes strong L^2 convergence.

By iteration, we have $\frac{N_t^\varepsilon(x) - N_t(x)}{\varepsilon}$ converges pointwise, and also strongly in L^2, which gives us the existence of $\psi \in (L^\infty(\Omega))^{T+1}$ such that

$$\psi_0(x) = 0$$

and

$$\int_\Omega rk(x, y)\left[\big(1 - \alpha_t(y)\big)\frac{(N_t^\varepsilon - N_t)(y)}{\varepsilon} - l_t(y) N_t^\varepsilon(y)\right] dy$$
$$\to \int_\Omega rk(x, y)\big[\big(1 - \alpha_t(y)\big)\psi_t(y) - l_t(y) N_t(y)\big]\,dy.$$

Passing to the limit, we get

$$\psi_{t+1}(x) = \int_{\Omega} rk(x,y)\big[\big(1-\alpha_t(y)\big)\psi_t(y) - l_t(y)N_t(y)\big]\,dy,$$

for $t = 0, \ldots, T$. □

Now we differentiate the map $\alpha \to J(\alpha)$ to obtain a characterization of an optimal control.

Theorem 3 *Given an optimal control α^* and corresponding state solution $N^* = N(\alpha^*)$, there exists a solution $p \in (L^\infty(\Omega))^T$ satisfying the adjoint system*:

$$\begin{aligned} p_{t-1}(x) &= r\big(1-\alpha^*_{t-1}(x)\big)\int_{\Omega} p_t(y)k(y,x)\,dy + 1, \\ p_T(x) &= 1 \end{aligned} \tag{36}$$

where $t = T, \ldots, 2, 1$. Furthermore, for $t = 0, 1, 2, \ldots, T-1$;

$$\alpha^*_t(x) = \min\left(\max\left(\frac{rN^*_t(x)}{B_t}\int_{\Omega} p_{t+1}(y)k(y,x)\,dy, 0\right), M\right). \tag{37}$$

Proof Let α^* be an optimal control and $N^* = N(\alpha^*)$ be the corresponding state. For variation l with $(\alpha^* + \varepsilon l) \in U$ for $\varepsilon > 0$ sufficiently small, let N^ε be the corresponding solution of the state equation. Since the adjoint system is linear, there exists a solution p. We compute the directional derivative of the functional $J(\alpha)$ with respect to α in the direction l at α^*. Since $J(\alpha^*)$ is the minimum value, we have

$$\begin{aligned} 0 &\le \lim_{\varepsilon\to 0^+} \frac{J(\alpha^* + \varepsilon l) - J(\alpha^*)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon}\left\{\sum_{t=1}^{T}\int_{\Omega}\big[N^\varepsilon_t(y) - N^*_t(y)\big]dy + \sum_{t=0}^{T-1}\int_{\Omega}\frac{B_t}{2}\big[\big(\alpha^*_t + \varepsilon l_t\big)^2 - \big(\alpha^*_t\big)^2\big]dy\right\} \\ &= \lim_{\varepsilon\to 0^+} \sum_{t=1}^{T}\int_{\Omega}\frac{N^\varepsilon_t(y) - N^*_t(y)}{\varepsilon}\,dy + \sum_{t=0}^{T-1}\int_{\Omega}\left[\frac{B_t}{2}\varepsilon l_t^2 + B_t\alpha^*_t l_t\right]dy \\ &= \sum_{t=1}^{T}\int_{\Omega}\psi_t(y)\,dy + \sum_{t=0}^{T-1}\int_{\Omega} B_t\alpha^*_t(y)l_t(y)\,dy. \end{aligned}$$

We use the coefficient '1' of the ψ_t term as the non-homogeneous term in the adjoint system and transform that term:

$$
\begin{aligned}
&\sum_{t=1}^{T}\int_{\Omega}\psi_t(y)\,dy\\
&\quad=\sum_{t=1}^{T-1}\int_{\Omega}\Big[p_t(y)-r\big(1-\alpha_t^*(y)\big)\int_{\Omega}p_{t+1}(x)k(x,y)dx\Big]\psi_t(y)\,dy\\
&\qquad+\int_{\Omega}\psi_T(y)\,dy\\
&\quad=\sum_{t=1}^{T-1}\int_{\Omega}p_t(y)\psi_t(y)\,dy-\sum_{t=1}^{T-1}\int_{\Omega}r\big(1-\alpha_t^*(y)\big)\psi_t(y)\int_{\Omega}p_{t+1}(x)k(x,y)\,dx\,dy\\
&\qquad+\int_{\Omega}\psi_T(y)\,dy\\
&\quad=\sum_{t=0}^{T-1}\int_{\Omega}p_{t+1}(y)\psi_{t+1}(y)\,dy\\
&\qquad-\sum_{t=0}^{T-1}\int_{\Omega}p_{t+1}(x)\int_{\Omega}r\big(1-\alpha_t^*(y)\big)\psi_t(y)k(x,y)\,dy\,dx\\
&\quad=\sum_{t=0}^{T-1}\int_{\Omega}p_{t+1}(x)\Big[\psi_{t+1}(x)-\int_{\Omega}rk(x,y)\big(1-\alpha_t^*(y)\big)\psi_t(y)\,dy\Big]dx\\
&\quad=\sum_{t=0}^{T-1}\int_{\Omega}p_{t+1}(x)\Big[-r\int_{\Omega}k(x,y)l_t(y)N_t^*(y)\,dy\Big]dx
\end{aligned}
$$

where we used $p_T(x)\equiv 1$, $\psi_0(x)\equiv 0$, and the sensitivity equation (33). Substituting out for the first term from our quotient calculation,

$$
\begin{aligned}
0&\le\sum_{t=0}^{T-1}\int_{\Omega}p_{t+1}(x)\Big[-r\int_{\Omega}k(x,y)l_t(y)N_t^*(y)\,dy\Big]dx+\sum_{t=0}^{T-1}\int_{\Omega}B_t\alpha_t^*(y)l_t(y)\,dy\\
&=\sum_{t=0}^{T-1}\int_{\Omega}\Big[\Big(\int_{\Omega}-p_{t+1}(x)k(x,y)\,dx\Big)rN_t^*(y)+B_t\alpha_t^*(y)\Big]l_t(y)\,dy.
\end{aligned}
$$

For any $t=0,1,\ldots,T-1$, on the set $\{x:0<\alpha_t^*(x)<M\}$, the variation l_t can be taken with support on this set, and have any sign, because the optimal control can be modified a little up or down and still stay inside the bounds. Thus, on this set, the rest of the integrand must be zero, so

$$
\alpha_t^*(x)=\frac{rN_t^*(x)}{B_t}\int_{\Omega}p_{t+1}(y)k(y,x)\,dy.
$$

By taking the upper and lower bounds into account we now show

$$\alpha_t^*(x) = \min\left(\max\left(\frac{rN_t^*(x)}{B_t}\int_\Omega p_{t+1}(y)k(y,x)\,dy, 0\right), M\right).$$

We now show how we handle the bounds. For any $t = 0, 1, \ldots, T-1$, on the set $\{x : \alpha_t^*(x) = 0\}$, take non-negative l_t with support on this set, and

$$0 \le \sum_{t=0}^{T-1}\int_\Omega \left[\left(\int_\Omega -p_{t+1}(x)k(x,y)\,dx\right)rN_t^*(y)\right]l_t(y)\,dy,$$

that indicates

$$\frac{rN_t^*(x)}{B_t}\int_\Omega p_{t+1}(y)k(y,x)\,dy \le 0.$$

Hence on this set, we have

$$\alpha_t^*(x) = \min\left(\max\left(\frac{rN_t^*(x)}{B_t}\int_\Omega p_{t+1}(y)k(y,x)\,dy, 0\right), M\right) = 0.$$

On the other hand, on the set $\{x : \alpha_t^*(x) = M\}$, then l_t with support on this set can only be non-positive, and

$$0 \le \sum_{t=0}^{T-1}\int_\Omega \left[\left(\int_\Omega -p_{t+1}(x)k(x,y)\,dx\right)rN_t^*(y) + B_tM\right]l_t(y)\,dy,$$

that indicates

$$\frac{rN_t^*(x)}{B_t}\int_\Omega p_{t+1}(y)k(y,x)\,dy \ge M.$$

Hence on this set,

$$\alpha_t^*(x) = \min\left(\max\left(\frac{rN_t^*(x)}{B_t}\int_\Omega p_{t+1}(y)k(y,x)\,dy, 0\right), M\right).$$

So our characterization of an optimal control is shown. □

5.2 Numerical Illustrations

We obtain uniqueness of the optimal control under the assumption of largeness of the cost coefficients, B_t, using a strict convexity argument. See [18] for similar arguments.

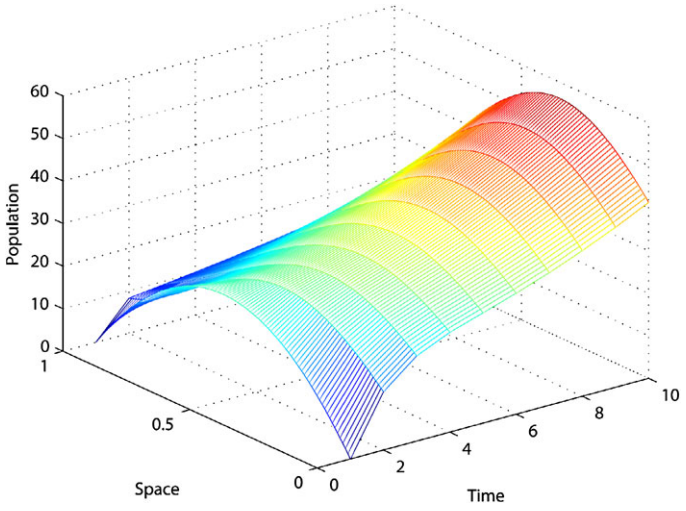


Fig. 5 Simulation of pest population without control, using a normal distribution kernel with $\beta = 5$ and grid spacing of 0.01

We show some numerical results to illustrate the optimal harvesting strategies for pest control.

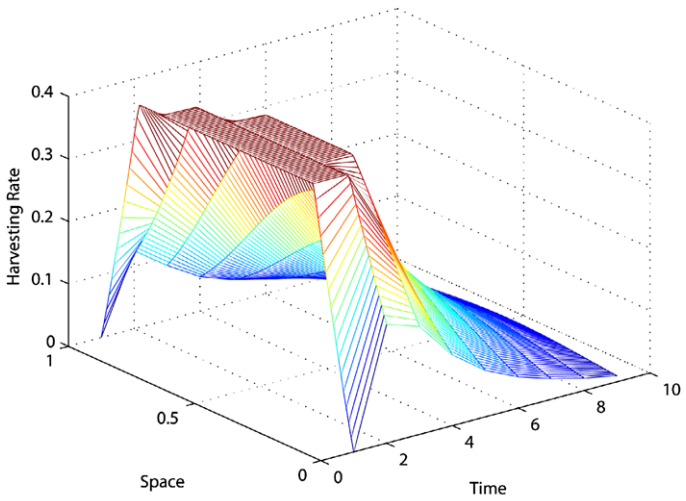
Starting with a given initial population distribution and a guess for the control, an iterative method is used to solve the optimality system. Given initial condition of the state and an initial guess of the control, we start with solving the state equations forward. Using the new state value, we solve backwards the adjoint equations, and calculate the characterization. We then update the control by taking a convex combination of the old control value and the new value from control characterization. The iteration stops when convergence occurs between successive iterates. The trapezoidal rule is used here to get integral approximations. We note that the trapezoidal rule requires C^2 regularity in space which holds for our examples. If x is near the $\partial\Omega$, then the part of the dispersal that would go outside Ω is not included in the integral. See [12, 14, 18, 19] for more details on such a numerical method.

We use a linear growth function with a growth rate r of 1.8 and the spatial domain is Ω. We study the harvesting strategy over a one dimensional space with size 1 during 10 time steps. Here the space grid size is 0.01. We use a parabola curve $100x(1-x)$ for the initial population. Possible maximum harvesting rate is 0.4 and the weights in the objective functional B_t to be a constant 1000 for each time step.

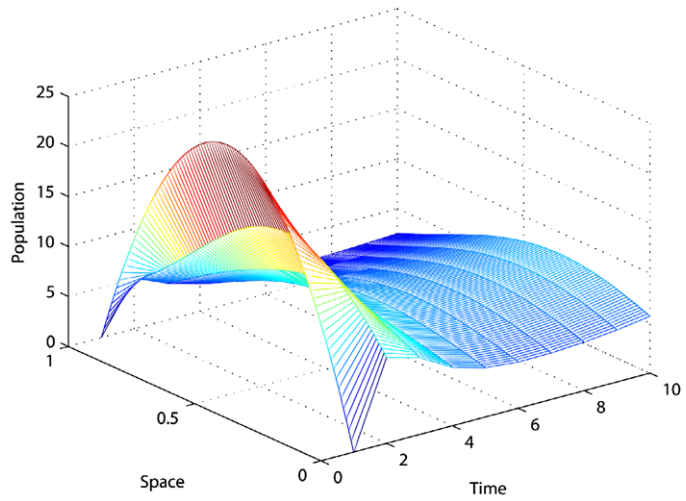
We illustrate numerical results for two commonly used kernels [23, 34].

The normal distribution kernel is

$$k(x, y) = \sqrt{\frac{\beta}{\pi}} \exp\left(-\beta(x-y)^2\right),$$



(a) Optimal harvesting rates



(b) Pest population

Fig. 6 Simulation of pest population using optimal harvesting rates in (**a**) with a normal distribution kernel with $\beta = 5$, and grid spacing of 0.01

and the finite range kernel is

$$k(x, y) = \begin{cases} 0, & \text{if } x \leq y - R \\ \frac{\pi}{4R} \cos[\frac{\pi}{2R}|x - y|], & \text{if } y - R < x < y + R \\ 0, & \text{if } x \geq y + R. \end{cases}$$

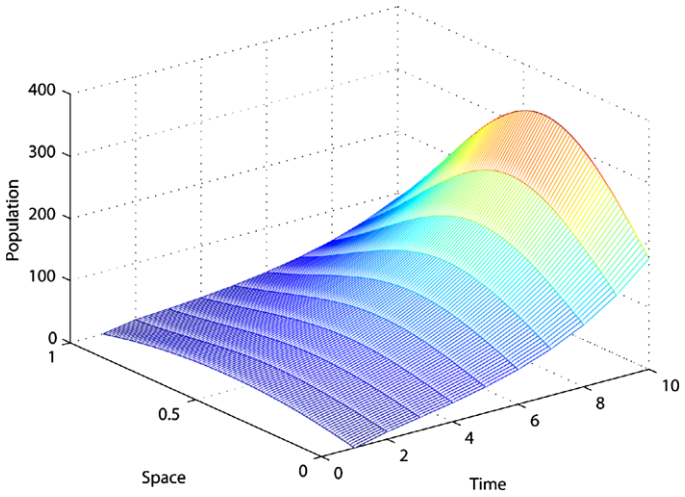


Fig. 7 Simulation of pest population without control, using a finite range kernel with $R = 0.5$ and grid spacing of 0.01

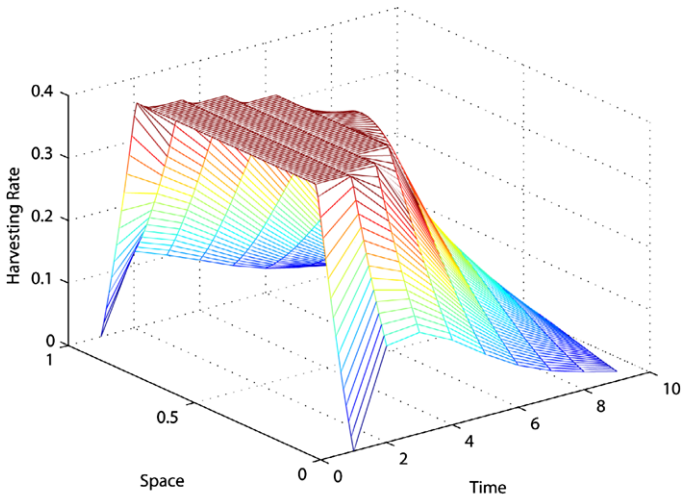
Figure 5 shows the pest populations at each time step when no harvesting is applied.

Figure 6(a) shows the optimal harvesting rates at each time step and Fig. 6(b) shows the corresponding populations after control is applied.

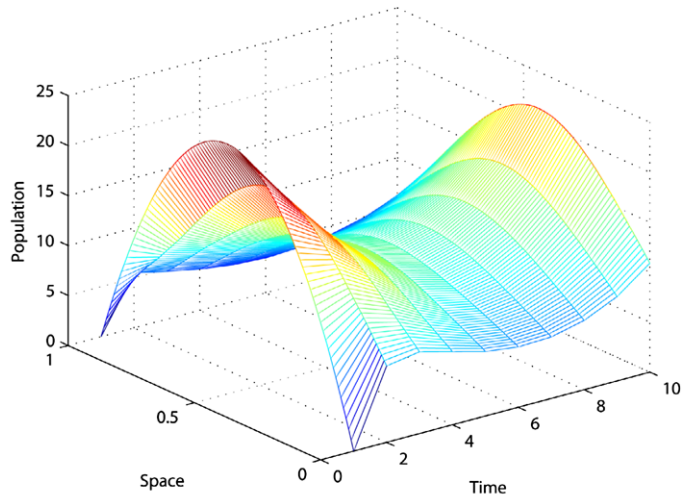
We use a finite dispersal kernel in Figs. 7, 8(a) and 8(b). The dispersal range is $R = 0.5$, which is half the size of the spatial region.

Dispersal close to the edge of the region causes population loss. Comparing Figs. 5 and 7, the pest population grow to a much higher scale in the latter case, due to less population loss from dispersal. Consequently, comparing the two optimal controls shown in Figs. 6(a) and 8(a), the harvesting rate for the latter case hits the ceiling for a longer time. The two pest populations after control shown in Figs. 6(b) and 8(b), are very similar to each other during the first four time steps. However, later on as the harvesting rates drop down, the population in the latter case recovers much faster.

The figures for both kernels show that the optimal harvesting strategy is to put more effort in the beginning time steps to cause an immediate drop of the pest population. We also perform harvesting at higher rates in the center of the region than on the edge. As shown in Figs. 6(a) and 8(a), the optimal harvesting rates are at the maximum level in the center of the region during the first few time steps. For both kernels, Figs. 6(b) and 8(b) show significant decrease in pest populations after harvesting is performed. Furthermore, the pest population starts to grow again slowly after we reduce harvesting effort during later time steps.



(a) Optimal harvesting rates



(b) Pest population

Fig. 8 Simulation of pest population using optimal harvesting rates in (**a**) using a finite range kernel with $R = 0.5$, and grid spacing of 0.01

6 Conclusion

We have presented an introduction of optimal control problems with discrete and continuous features and illustrated the use of adjoint functions in solving these problems. Our techniques used ideas related to the Pontryagin's Maximum Principle and extensions. Our numerical results used the forward-backward sweep method, an it-

erative algorithm to solve the optimality system, consisting of the state and adjoint equations and the optimal control characterization. For other solution approaches and numerical methods, see [1, 3, 17, 43].

Biological models are becoming more and more complex, and optimal control and optimization tools are useful in investigating management strategies in many complex models [1, 17, 32, 44]. However, some new tools do need to be developed, like tools for controlling individual-based models [38].

Acknowledgements Lenhart's work is partially supported by the National Institute for Mathematical and Biological Synthesis funded through the National Science Foundation EF0832858. We would like to thank David Reister and Louis Gross for some assistance.

References

1. Anita, S., Arnăutu, V., and Capasso, V.: An Introduction to Optimal Control Problems in Life Sciences and Economics. Birkhäuser, Basel (2011)
2. Behncke, H.: Optimal control of deterministic dynamics. Optim. Control Appl. Methods **21**, 269–285 (2000)
3. Betts, J.: Practical Methods for Optimal Control Using Nonlinear Programming. SIAM, Philadelphia (2001)
4. Blackwood, J., Hastings, A., and Costello, C.: Cost-effective management of invasive species using linear-quadratic control. Ecol. Econ. **69**, 519–527 (2010)
5. Bodine, E. N., Gross, L. J., and Lenhart, S.: Optimal control applied to a model for species augmentation. Math. Biosci. Eng. **4**, 669–680 (2008)
6. Bryson, A. E., Jr, and Ho, Y.-c.: Applied Optimal Control. Ginn, Waltham (1969)
7. Clayton, T., Duke-Sylvester, S., Gross, L. J., Lenhart, S., and Real, L. A.: Optimal control of a rabies epidemic model with a birth pulse. J. Biol. Dyn. **4**, 43–58 (2010).
8. dePillis, L. G., Gu, W., Fister, K. R., Head, T., Maples, K., Neal, T., Murugan, A., and Yoshida, K.: Chemotherapy for tumors: an analysis of the dynamics and a study of quadratic and linear optimal controls. Math. Biosci. **209**, 292–315 (2007)
9. Epanchin-Neill, R. S., and Hastings, A.: Controlling established invaders: integrating economics and spread dynamics to determine optimal management. Ecol. Lett. **13** (4), 528–541 (2010)
10. Fleming, W., and Rishel, R.: Deterministic and Stochastic Optimal Controls. Springer, Berlin (1975)
11. Fister, K. R., and Panetta, J. C.: Optimal control applied to competing chemotherapeutic cell-kill strategies. SIAM J. Appl. Math. **63**, 1954–1971 (2003)
12. Gaff, H., Joshi, H.R., and Lenhart, S.: Optimal harvesting during an invasion of a sublethal plant pathogen. Environ. Dev. Econ. J. **12**, 673–686 (2007)
13. Georgescu, P., Dimitriu, G., and Sinclair, R.: Impulsive control of an integrated pest management model with dispersal between patches. J. Biol. Syst. **18**(3), 535–569 (2010)
14. Hackbusch, W.: A numerical method for solving parabolic equations with opposite orientations. Computing **20**, 229–240 (1978)
15. Hawkins, B. A., and Cornell, H. V.: Theoretical Approaches to Biological Control. Cambridge University Press, Cambridge (1999)
16. Heinricher, A., Lenhart, S., and Solomon, A.: The application of optimal control methodology to a well-stirred bioreactor. Nat. Resour. Model. **9**, 61–80 (1995)
17. Hof, J., and Bevers, M.: Spatial Optimization in Ecological Applications. Columbia University Press, New York (2002)

18. Joshi, H. R., Lenhart, S., and Gaff, H.: Optimal harvesting in an integrodifference population model. Optim. Control Appl. Methods **27**, 135–157 (2006)
19. Joshi, H. R., Lenhart, S., Gaff, H., and Lou, H.: Harvesting control in an integrodifference population model with concave growth term. Nonlinear Anal. Hybrid Syst. **3**, 417–429 (2007)
20. Krabs W., and Pickl, S.: Modelling, Analysis and Optimization of Biosystems. Springer, Berlin (2007)
21. Kern, D. L., Lenhart, S., Miller, R., and Yong, J.: Optimal Control applied to native-invasive population dynamics. J. Biol. Dyn. **1** (4), 379–393 (2007)
22. Kot, M., and Schaffer, W. M.: Discrete-Time Growth-Dispersal Models. Math. Biosci. **80**, 109–136 (1986)
23. Kot, M.: Discrete-time travelling waves: ecological examples. J. Math. Biol. **30**, 413–436 (1992)
24. Kot, M., Lewis, M. A., and van den Driessche, P.: Dispersal data and the spread of invading organisms. Ecology **77**, 2027–2042 (1996)
25. Kot, M. : Elements of Mathematical Ecology. Cambridge University Press, Cambridge (2001)
26. Krener, A. J.: The high order maximal principle and it's application to singular extremals. SIAM J. Control Optim. **15**, 256–293 (1977)
27. Ledzewicz, U., and Schattler, H.: Second order conditions for extremum problems with non-regular equality constraints. J. Optim. Theory Appl. **86**, 113–144 (1995)
28. Ledzewicz, U., Brown, T., and Schattler, H.: A comparison of optimal controls for a model in cancer chemotherapy with L_1- and L_2- type objectives. Optim. Methods Softw. **19**, 351–359 (2004)
29. Lenhart, S., and Workman, J. T.: Optimal Control of Biological Models. Chapman and Hall/CRC Publishers, London/Boca Raton (2007)
30. Lenhart, S., and Zhong, P.: Investigating the order of events in optimal control of integrodifference equations. In Systems Theory: Modeling, Analysis and Control Proceedings Volume, 89–100 (2009). Presses Universitaires de Perpignan, France
31. Lewis, M. A., and Van Kirk, R. W.: Integrodifference models for persistence in fragmented habitats. Bull. Math. Biol. **59**, 107–137 (1997)
32. Li, H., and Yong, J.: Optimal Control Theory for Infinite Dimensional Systems. Birkhauser, Boston (1995)
33. Mesterson-Gibbons, M.: A primer on the Calculus of Variations and Optimal Control Theory. Student Mathematical Library **50**, AMS, Providence (2009)
34. Neubert, M., Kot, M., and Lewis, M. A.: Dispersal and pattern formation in a discrete-time predator-prey model. Theor. Popul. Biol. **48**, 7–43 (1995)
35. Neubert, M.: Marine reserves and optimal harvesting. Ecol. Lett. **6**, 843–849 (2003)
36. Olson, L.: The economics of terrestrial invasive species: a review of the literature. Agric. Resour. Econ. Rev. **35**, 178–194 (2006)
37. Pontryagin, L. S., Boltyanskii, V. G., Gamkrelidze, R. V., and Mishchenko, E. F.: The Mathematical Theory of Optimal Processes. Wiley, New York (1967)
38. Railsback, S. F., and Grimm, V.: Agent-based and individual-based modeling: a practical introduction. Princeton University Press, Princeton (2012)
39. Reister, D. B.: A new wheel control system for the omnidirectional HERMIES-III robot. Robotica **10**, 351–360 (1992)
40. Reister, D. B., and Lenhart, S. M.: Time optimal paths for high-speed maneuvering. Int. J. Robot. Res. **14**, 184–194 (1995)
41. Salinas, R. A., Lenhart, S., and Gross, L. J.: Control of a metapopulation harvesting model for black bears. Nat. Resour. Model. **18**, 307–321 (2005)
42. Sethi, S. P., and Thompson, G. L.: Optimal Control Theory: Applications to Management Science and Economics. Kluwer Academic, Boston, 2nd edn. (2000)
43. Speyer, J. L., and Jacobson D. H.: Primer on Control Theory. Advances in Design and Control. SIAM, Philadelphia (2010)
44. Troelzsch, F.: Optimal Control of Partial Differential Equations: Theory, Methods, and Applications. Am. Math. Soc., Providence (2010)

45. Whittle, A., Lenhart, S., and Gross, L. T.: Optimal control for management of an invasive plant species. Math. Biosci. Eng. **4**, 101–112 (2007)
46. Zhang, T., Meng, X., and Song, Y.: The dynamics of a high-dimensional delayed pest management model with impulsive pesticide input and harvesting prey at different fixed moments. Nonlinear Dyn. **64**(1–2), 1–12 (2011)

Index

R. Melnik, I. Kotsireas (eds.), *Advances in Applied Mathematics, Modeling, and Computational Science*, Fields Institute Communications 66,
DOI 10.1007/978-1-4614-5389-5, © Springer Science+Business Media New York 2013