Fields Institute Communications 79

The Fields Institute for Research in Mathematical Sciences

Roderick Melnik
Roman Makarov
Jacques Belair Editors

Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science

FIELDS

Springer

Fields Institute Communications

Volume 79

The Communications series features conference proceedings, surveys, and lecture notes generated from the activities at the Fields Institute for Research in the Mathematical Sciences. The publications evolve from each year's main program and conferences. Many volumes are interdisciplinary in nature, covering applications of mathematics in science, engineering, medicine, industry, and finance.

More information about this series at http://www.springer.com/series/10503

Roderick Melnik • Roman Makarov • Jacques Belair
Editors

Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science

FIELDS The Fields Institute for Research
in the Mathematical Sciences

Springer

Editors
Roderick Melnik
Department of Mathematics
and MS2Discovery Interdisciplinary
Research Institute
Wilfrid Laurier University
Waterloo, ON, Canada

Roman Makarov
Department of Mathematics
and MS2Discovery Interdisciplinary
Research Institute
Wilfrid Laurier University
Waterloo, ON, Canada

Jacques Belair
Departement de Mathematiques
Universite de Montreal
Montreal, QC, Canada

ISSN 1069-5265 ISSN 2194-1564 (electronic)
Fields Institute Communications
ISBN 978-1-4939-6968-5 ISBN 978-1-4939-6969-2 (eBook)
DOI 10.1007/978-1-4939-6969-2

Library of Congress Control Number: 2017943956

Mathematics Subject Classification (2010): 15-XX, 34-XX, 35-XX, 37-XX, 60-XX, 45-XX, 49-XX, 62-XX, 65-XX, 68-XX, 74-XX, 76-XX, 91-XX, 92-XX

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

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

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC
The registered company address is: 233 Spring Street, New York, NY 10013, U.S.A.

Preface

The application of mathematics and statistics in the age of computational science and engineering has transformed our society and has revolutionized the world we live in. Being some of the oldest cultural achievements of mankind, nowadays these disciplines are intrinsic part of our daily life through our activities and technologies, ranging from our banking and investment systems to new sophisticated electronic devices, and to our civil infrastructure and environment.

These disciplines continuously grow at their frontiers though new areas of applications, new theories, and new tools provided by mathematical and statistical models. As a result, they continue representing the core of human knowledge critical for new discoveries and innovation, our well-being, and our economic prosperity.

There has been a long and rich interplay between mathematics and statistics on the one hand and other disciplines on the other, resulting in their fruitful enrichments. With ever-expanding interdisciplinary horizons of applied mathematics and statistics, we see new progress and modern challenges in their development. This book is about such progress and challenges in applied mathematics, modelling, and computational science.

Today, mathematical and statistical models are applied in natural and social sciences, industry and technology, medicine and finance. They are at the heart of a multitude of human activities, allowing connecting such activities in a modern world, where our communication gets better, faster, and cheaper also due to mathematics-based models. They substantially contribute to our better understanding of complex systems and networks whose components interact in a dynamic manner. Furthermore, mathematics-based computational technologies enable us detailed simulations of complex systems in the areas where the knowledge about such systems has been limited until very recently. Many such systems are functioned in a competitive, and often uncertain, environment. Therefore, the development of mathematical and statistical based methodologies of uncertainty quantification, as well as addressing other associated challenges, is essential.

Along with more traditional applications of mathematics and statistics in physics and engineering, we are witnessing now substantial contributions of these disciplines to new breakthroughs in biology and medicine, finance, and social sciences.

Equally important, mathematical and statistical models allow us to develop new important insight and better understanding of environmental and ecological sustainability in our dynamic and complex world.

This book provides details on recent progress and challenges in selected areas of applied mathematics, modelling, and computational science. It contains 14 chapters which open to the reader details on state-of-the-art achievements in these selected areas. The book provides a balance between fundamental theoretical and applied developments, emphasizing interdisciplinary nature of modern trends in these areas.

Written by 27 experts in their respective fields, the book is aimed at researchers in academia, practitioners, and graduate students. It can serve as a reference in the diverse selected areas of applied mathematics, modelling, and computational science. The book promotes interdisciplinary collaborations in addressing new challenges in these areas.

We are thankful to the referees of this volume for their invaluable help and suggestions. We are also very grateful to the Springer editorial team, and in particular to Dahlia Fisch, for their highly professional support.

Waterloo, ON, Canada Roderick Melnik
Waterloo, ON, Canada Roman Makarov
Montreal, QC, Canada Jacques Belair

Contents

Part V Materials Science, Engineering, and New Technologies

Part VI Finance and Systemic Risk

Part VII Life and Environmental Sciences

Part VIII Number Theory and Algebraic Geometry in Cryptography and Other Applications

Part IX Sustainability and Cooperation

Part I
Modern Challenges and Interdisciplinary Interactions via Mathematical, Statistical, and Computational Models

Modern Challenges and Interdisciplinary Interactions via Mathematical, Statistical, and Computational Models

Roderick Melnik, Roman Makarov, and Jacques Belair

Abstract We live in an incredible age. Due to extraordinary advances in sciences and engineering, we better understand the world around us. At the same time, we witness profound changes in the technology, environment, societal organization, and economic well-being. We face new challenges never experienced by humans before. To efficiently address these challenges, the role of interdisciplinary interactions will continue to increase, as well as the role of mathematical, statistical, and computational models, providing a central link for such interactions.

1 The Role of Mathematical and Statistical Models

Since the dawn of human civilizations, technological innovations have been developing hand in hand with progress in mathematical and statistical sciences. Interactions and interdependence of mathematics, physics, engineering, and biology have been well elucidated in the literature with a number of excellent reviews and historical accounts (e.g., [12, 13, 19] and references therein). In the heart of these interactions and interdependence are mathematical and statistical models. Their role will continue to increase rapidly in both traditional (e.g., physics and engineering) and many emerging (e.g., health and life sciences) areas of their applications (e.g., [10, 15, 18] and references therein). Moreover, we are witnessing a dramatic increase in computing power and breathtaking advances in computational science and engineering which assist further in developing this trend.

Today, many other disciplines are catching up with this trend too. Indeed, mathematical and statistical models can be used to describe complex phenomena and systems such as stock markets, the internet traffic, logistics, supply, and demand

R. Melnik (✉) • R. Makarov
Department of Mathematics and MS2Discovery Interdisciplinary Research Institute, Wilfrid Laurier University, Waterloo, ON, Canada, N2L 3C5
e-mail: rmelnik@wlu.ca

J. Belair
Departement de Mathematiques, Universite de Montreal, Montreal, QC, Canada, H3C 3J7

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_1

of industrial networks, as well as climate change dynamics. Many complex systems that appear in nature, engineering applications, and society have components that interact in a remarkably dynamic manner in competitive, and often uncertain, environments. In order to understand them better, there is a need to develop new mathematical, including stochastic, models, as well as new methods for uncertainty quantification.

New challenges in of the modern world and our society require researchers working on many problems in economics and finance, social, environmental, and management sciences look for the development of quantitative models based on mathematical and statistical theories, methods, and tools.

As a result, new scientific, technological, and societal challenges we face in the twenty-first century can only be efficiently addressed in close collaboration with mathematicians and statisticians developing such quantitative models. At the same time, such challenges will stimulate the development of new concepts and theories in mathematical and statistical sciences, leading to many new breakthroughs in these two-way interactions between mathematics and statistics on the one hand and other disciplines on the other.

2 Application Areas and State-of-the-Art Developments

From a wide and increasing spectrum of applications of mathematical, statistical, and computational models, we selected some representative areas of these applications. Thus, the rest of the book consists of eight sections based on these areas. They contain state-of-the-art chapters, written by leading specialists from all over the world.

In selecting our areas for this book we intended to open to the reader a rich field of interdisciplinary interactions between many different disciplines with their unifying thread via mathematical and statistical models. The book provides details on theoretical advances in these selected areas of applications, as well as representative examples of modern problems from such applications. It also exposes the reader to open and emerging problems, and to challenges that lie ahead in addressing such problems.

Following this introductory section, each remaining section with its chapters stands alone as an in-depth research or a survey within a specific area of application of mathematical, statistical and computational modeling. Next, we highlight the main features of each such chapter within remaining sections of this book.

2.1 Large Deviation Theory and Random Perturbations of Dynamical Systems with Applications

Large deviation theory provides an important framework for modern statistical mechanics and stochastic system/process modelling. It allows us to describe the asymptotic behaviour of remote tails of sequences of probability distributions. Within this framework, many concepts of equilibrium statistical mechanics, such as entropy or free energy, can be considered as large deviation rate functions and can be generalized to the non-equilibrium case. Moreover, today this theory is considered to be one of the major tools for our better understanding of statistical information about complex dynamical systems, including the information on their most probable states, rare events, extremes, attractors, and typical fluctuations.

It is well known that long-time evolution of dynamical systems can be seriously affected by small random perturbations, leading to a lasting effect on their evolution. They can bring metastability, pronounced in transitions between otherwise stable equilibria [28], the phenomenon that is observed in a wide range of applications such as fluid dynamics, chemical reactions, population dynamics, and neuroscience. The mechanism of this phenomenon can also be explored with large deviation theory. A key element in the application of large deviation theory in this case is the path of maximum likelihood of such transitions. Moreover, the path itself can be computable through a numerical optimization problem as the minimizer of a certain objective function, known as action.

This section of the book, written by T. Grafke, T. Schafer, and E. Vanden-Eijnden, provides a review of theoretical foundation of large deviation theory that led to the rate function minimization problem. In particular, the authors are focusing on the geometric variant of this problem that is fundamental to the geometric minimum action method. They have proposed a new algorithm that simplifies this latter method. The authors demonstrate a considerable potential of their developed algorithm for a range of applied problems, including examples ranging from fluid dynamics and materials science to reaction kinetics and climate modelling.

2.2 Nonlinear waves, Hyperbolic Problems, and their Applications

For centuries the development of mathematical models and studies of waves have fascinated many researchers. Already Pythagoras analyzed waves through the relation of pitch and length of string in musical instruments. Today, the role of wave equations in the modern science and engineering is hard to overestimate. They are applied in classical and quantum mechanics, materials science and biology, medicine and finance, climate studies and social science. It is also an active area of theoretical research which includes the development of analytical and computational techniques and important connection to other areas of mathematics [23].

Originally presented in the context of physics applications, coherent states play an important role in nonlinear wave equations. In particular, such states are considered as quasi-classical states in quantum mechanical applications. Today the concept has been generalized to a number of other areas of mathematical physics and beyond (e.g., [5, 27]).

The first chapter of this chapter, written by E. Kirr, starts from a general Hamiltonian formulation applicable to a large class of models related to wave propagation. Apart from the classical wave equation, this includes mathematical models based on Schrödinger's, Hartree's, Dirac's, Klein-Gordon's, Kortweg-de-Vries' equations. The author demonstrates that while large coherent structures can be found via variational methods (e.g., as minimizers of the energy, subject to a fixed value on the second conserved property), this is not the case for the problems where all coherent states are required. To address this challenge the author proposes to apply the analytical global bifurcation theory for finding all coherent states, as well as for analyzing orbital stability of such states. Within this framework, the author provides details on how to study asymptotic stability of coherent states, as well as on long-time behaviour of nearby solutions, and identifies some open problems in this field. For example, despite the recent progress in asymptotic stability near an orbitally stable coherent state, in the general case we still do not know how to determine the full dynamic picture near a bifurcation point.

Hyperbolic equations are in the heart of discussion in the second chapter of this section, written by R. Abgrall, who deals with both linear and non-linear problems. The main focus is on the development of parameter-free methods for scalar hyperbolic equations that satisfy a local maximum principle. The author presents a systematic methodology for constructing higher order finite element type methods satisfying this principle. The results are not limited to the problems with regular solutions only. A detailed analysis of conditions that guarantee the convergence of the developed numerical scheme to weak solutions under stability assumptions has been provided. Furthermore, the author has provided the conditions that guarantee an arbitrary order of accuracy of the developed scheme. Generalizations of the proposed methodology have also been discussed in the context to its extensions to systems, including Euler's equations and the Navier-Stokes model. Among the remaining challenges the author highlights the importance of a better design of the filtering parameter.

2.3 Group-Theoretical Approaches to Conservation Laws and Their Applications

Numerical integrators, where we preserve exactly one or more properties of the original differential-equation-based mathematical model, has been a subject of interest for a long time, with a number of excellent reviews, books, and journal special issues published (e.g., [3, 11, 26] and references therein). Given that geometric properties of the exact flow of the underlying differential equation are typically preserved in such cases, we call the associated integrators structure-preserving or geometric numerical integrators. While this type of methodologies has largely been developed for ordinary differential equations, there are important results in the development of these ideas to partial differential equations too (e.g., [4, 20, 21]). These methodologies, applied to both deterministic and stochastic systems, have been developed in parallel, and often independently, from energy-conserving methods (e.g., [1, 16, 25] and references therein). Such methods are typically derived for the variational formulation of the problem and can be applied to both Lagrangian and Hamiltonian dynamics.

The underlying success of variational integrators, leading to their numerous applications, lies with their group-theoretical foundation and Lie group analysis which has been well elucidated in the literature (e.g., [14] and references therein). For example, applied to Lagrangian dynamical systems, they preserve a discrete multisymplectic form, as well as momenta associated to symmetries of the Lagrangian via Noether's theorem. As it was pointed out in [17], a prerequisite of obtaining variational integrators is the existence of a variational formulation for the considered dynamical system. Not all systems in applications are of this type. Examples of non-variational mathematical models based on partial differential equations can be found in such areas as plasma physics, fluid dynamics, as well as in magnetohydrodynamics, to name just a few. As a result, there is an increasing interest to a generalization of Noether's theorem to handle such cases too. Recent attempts in this direction include a discrete version of the Noether theorem for formal Lagrangians that yields the discrete momenta preserved by the resulting numerical schemes [17]. The method, based on the embedding of a dynamical system into a Lagrangian system by doubling the number of variables, has been applied to Vlasov-Poisson and magnetohydrodynamic systems, as well as to non-canonical Hamiltonian systems.

This section is a comprehensive review, written by S. Anco, discussing other generalizations of Noether's theorem to non-variational mathematical models based on partial differential equations. One of the major concepts is that related to multipliers, the expressions whose summed products with a PDE-based system yields a local divergence identity. The latter is associated with a continuity equation involving a conserved density and a spatial flux for solutions of the underlying PDE. The author demonstrates that when the underlying model is non-variational, such multipliers are an adjoint counterpart to infinitesimal symmetries. Moreover, the local divergence identity, that relates a multiplier to a conserved integral, appears to be an adjoint generalization of the variational identity that underlies Noether's theorem. A procedure for computation of multiplies has been described in detail.

2.4 Materials Science, Engineering, and New Technologies

Computer-aided innovation of new materials is an important area of research in materials science and engineering. The development of computationally efficient approaches and modelling in this field has been a subject of immense research interest ever since our advances in computational power [7, 22]. This includes innovative superhard materials, as well as smart materials such as superelastic and shape memory alloys [6]. A large class of such materials are binary alloys. For binary alloys, the most accurate energy calculations are typically done via the density functional theory, and as any ab initio calculations this methodology is computationally very expensive.

The first chapter in this section, written by J. Kristensen, I. Bilionis, and N. Zabaras, discusses viable alternatives to the above methodology. They argue that in this area of applications it is important to devise new schemes for the automatic and maximally informative selection of simulations. They provide a detailed description of their developed information acquisition policy for learning the ground state of binary alloys. Starting from the surrogate modelling technique and presenting the energy computation scheme, the authors describe their theoretical approach, based on a Bayesian interpretation of the cluster expended energy. Their developed framework for selecting structures has been extended to account for the effect of alloy structure costs. By comparisons with other structure acquisition algorithms, it has been concluded that optimal information acquisition policies should balance the maximization of the expected improvement of the ground state line and the minimization of the size of the simulated structure. The developed approach has been validated for a number of important binary alloys, including NiAl and TiAl. Once a probabilistic surrogate of the relevant thermodynamic potential is constructed, the proposed policies can be directly applied to the discovery of generic phase diagrams.

The second chapter of this section, written by P. Fischer, M. Schmitt, and A. Tomboulides, presents a comprehensive overview of spectral element methods for an important class of fluid dynamics problems. Their focus is on incompressible and low-Mach-number flows in domains with moving boundaries. From applications of these mathematical models, it is well known that moving boundaries introduce new sources of nonlinearity and stiffness [9]. For example, in fluid-structure interaction problems, one of the reasons for that lies with disparate time scales between the fluid and solid responses. A similar situation holds for other coupled problems. The authors pay special attention to recent developments addressing these moving-domain challenges, while keeping the computational efficiency required for turbulent flow simulations. One of the important features in their discussion is an arbitrary Lagrangian-Eulerian formulation for low-Mach-number flows that includes an evolution equation for the background thermodynamic pressure. A rich selection of numerical examples has also been provided to illustrate main theoretical results.

The concluding chapter of this section, written by C. Budd, offers an exciting journey into mathematical foundations of new technologies. The author reviews eight such technologies, identified by the UK government as those that would act as a focus for future scientific research and funding. They are

(a) Big Data
(b) Satellites and space
(c) Robotics and autonomous systems
(d) Synthetic biology
(e) Regenerative medicine
(f) Agricultural science
(g) Advanced materials
(h) Energy and its storage.

Based on a historical account and recent progress, the author demonstrates that mathematics lies at the heart of all of them, linking them all together. A number of challenges, both mathematical and interdisciplinary, have been determined and discussed in the context of future development of these technologies.

2.5 Finance and Systemic Risk

This section contains three chapters. It is opened by a chapter written by Q. Feng and C. Oosterlee addressing the problem of credit valuation adjustment. The authors pointed out that this quantity is required in the Third Basel Accord - a global framework on bank capital adequacy, stress testing, and market liquidity risk - that came around in the wake of the credit crisis. In calculating credit valuation adjustment, exposure is a key element. This characteristic is defined as the potential future loss on a financial contract due to a default event. The chapter describes a backward-dynamics-based general framework for calculating exposure profiles for different options, enabling us to analyze the sensitivity of the model to such options. The authors focus on two models, the Heston and Heston-Hull-White asset dynamic models, which they consider under European, Bermudan, and barrier options. In particular, for these models and options they describe their generalization of the Stochastic Grid Bundling Method for the computation of exposure profiles and sensitivity for asset dynamics. This generalization provides a flexible valuation framework for credit valuation adjustment. Details are given for the most important features of the developed methodology, including the choice of the basis functions for the local regression, the convergence of the direct and path estimators with respect to an increased number of bundles, and the associated accuracy. A series of numerical tests presented in this chapter demonstrated these features in practice, also showing that the computational efficiency of the developed methodology is connected to the number of bundles used in the Stochastic Grid Bundling Method. A drastic reduction in computational time is achieved with a parallel implementation of the developed algorithm.

The next chapter is this section, written by T. Bielecki, J. Jakubowski,a nd M. Nieweglowski, is devoted to recent progress in the emerging theory and practice of structured dependence between stochastic processes. It is argued that our success in this field will be dependent on our ability to construct different types of Markov copulae [2]. The authors of this chapter present a new result on independence copula for conditional Markov chains. A copula can be used to describe the dependence between random variables. It is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform. The authors of this chapter describe how to construct the conditionally independent Markov copula (or the conditionally independent multivariate Markov coupling) for a family of conditional Markov chains. While the reported result is important in finance applications, e.g. in modelling credit rating migrations, there is a range of its possible applications in other areas. One of the challenges in this field, identified by the authors, is to effectively construct weak Markov copulae and weak conditional Markov chains.

This section is concluded by the chapter devoted to financial systemic risk models. Written by T. Hurd, its main result is in providing essential foundations needed to prove rigorous percolation bounds and cascade mapping in assortative networks. The main premises of the author's approach are based on the fact that the network of interbank counterparty relationships can be described as a directed random graph. When cascade models of financial systemic risk are used, the structure of this graph (or the skeleton of a financial network) can be thought as a medium through which financial contagion is propagated. The author focuses on a particular general class of random graphs—the assortative configuration model. A new approximate Monte Carlo simulation algorithm for assortative configuration graphs has been described in detail, and challenges for efficient simulations of such graphs have been highlighted.

2.6 Life and Environmental Sciences

Many problems in biology and life sciences require consideration of environmental effects. One class of such problems is related to the analysis of coexistence of interacting populations in uncertain environment.

This section, written by S. Schreiber, focuses on this class of problems accounting for random fluctuations due to both environmental and demographic stochasticities which are experienced by all populations. It is argued that demographic stochasticity can be represented by Markovian models with a countable number of states where quasi-stationary distributions of these models characterize metastable patterns of the system behaviour connected to long-term transients. At the same time, the effects of environmental stochasticity on population dynamics can be modelled with stochastic difference equations. The author explains that for these models, stochastic persistence would correspond to empirical measures placing arbitrarily little weight on arbitrarily low population densities. Sufficient and necessary conditions for such persistence are based on a weighted combination of Lyapunov exponents. The theory has been developed for both single-species and multi-species models.

The author has provided the reader with a range of interesting examples to support the developed theory. This includes the quantification of climatic variability effects on the dynamics of Bay checkerspot butterflies, the persistence of coupled sink populations, coexistence of competitors through the storage effect, and stochastic rock-paper-scissor communities. The chapter contains a comprehensive list of open problems and challenges in this field.

2.7 *Number Theory and Algebraic Geometry in Cryptography and Other Applications*

The topic of elliptic curves has an important place in mathematics and its applications. In the domain of applications this topic came to its new prominence in the late 1970s of the twentieth century when public key cryptography and cryptosystems become important for private and secure electronic communication [8]. In 1993 it became also known that elliptic curves were used in Andrew Wiles' proof of Fermat's Last Theorem. With the astounding growth of the Internet and new security challenges of the twenty-first century, there is all evidence to expect increasing importance of elliptic curves in a number of application areas. Nowadays, the topic represents a combination of important challenges in practical/algorithmic issues and the underlying mathematical beauty with a range of open problems.

This section covers both computational/algorithmic and theoretical aspects of elliptic curves. Written by M. Bennett and A. Rechnitzer, the section provides a good introduction to ubiquitous nature of these structures in mathematical sciences, particularly in number theory and algebraic geometry. From a practical viewpoint, the authors focus on the problem of generating/tabulating elliptic curves with desired properties. Along with a comprehensive overview of state-of-the-art in the area, they provide details of an algorithm for computing models for all elliptic curves with integer coefficients and given conductor. The latter quantity has been studied since A. Ogg and A. Weil in the late 1960s of the previous century, and is often considered to be an integral ideal analogous to the Artin conductor of a Galois representation. In the context of the current section, the authors define it as an invariant that provides information about how a given elliptic curve behaves over finite fields. Based on extensive comparisons to existing data, they demonstrate that although their approach is based on classical ideas, it leads to a very efficient computational algorithm. Furthermore, given multiple examples and data provided in this section, the authors challenge the reader with new problems in this exciting area.

2.8 Sustainability and Cooperation

In many applications we have to deal with dynamic interactions of several entities, agents, or players, in such a way that we can achieve a long-term cooperation, stability, and sustainability. Many problems in economics, social, engineering, and management sciences are of this type. Additional examples include the interaction between economic and ecological dynamic systems or other systems where cooperation and negotiation on the amount and the allocation of investment could lead to more sustainable use of natural resources [24]. In many such cases, the dynamic coalition-formation process can be modelled via a self-organized transition from unilateral action (Nash equilibria) to multilateral cooperation (Pareto optima).

When we have only a few agents/players with interdependent playoffs, cooperating/competing repeatedly over time for resources under uncertainty (e.g., in demand), the general framework of dynamic games played over event trees is often most suitable way to formalize such problems mathematically.

In this section, written by G. Zaccour, all principle components of the theory of dynamic games played over event trees have been reviewed. Starting from a review of the literature pertinent to the sustainability of cooperation in dynamic games, the author moves to the details of the approach to achieve a node-consistent outcome in dynamic games played over event trees. This approach is illustrated by the node-consistent Shapley value, as well as by the node-consistent core. In a methodologically consistent manner, the author has demonstrated how sustainable cooperative solutions can be constructed. The chapter has also highlighted several open problems. For example, can cooperation still be sustained if the cores in some of subgames are empty? Another interesting problem is the analysis of node consistency for dynamic games played over event trees in the case when the end of the horizon is random.

3 Conclusions

In this section we highlighted a selection of areas, representing part of a broad spectrum of the interdisciplinary interface where mathematical, statistical, and computational models play a central role. Such models provide an indispensable tool for scientific discoveries and innovation in the areas ranging from physics and biology to economics and finance, from security and defense to sustainability studies.

References

1. Asher, U. M., Surprising Computations, Applied Numerical Mathematics, 62 (10), SI, 1276-1288, 2012.
2. Bibbona, E., Sacerdote, L., and Torre, E., A Copula-Based Method to Build Diffusion Models with Prescribed Marginal and Serial Dependence, Methodology and Computing in Applied Probability, 18 (3), 765-783, 2016.
3. Blanes, S. and Casas, F., A Concise Introduction to Geometric Numerical Integration, Taylor and Francis, CRC Press, 2016.
4. Budd, C. J. and Piggott, M. D., The geometric integration of scale-invariant ordinary and partial differential equations, Journal of Computational and Applied Mathematics, Vol. 128, Issues 1–2, 399–422, 2001.
5. Clerc, M. G., Coulibaly, S., Ferre, M. A., Garcia-Nustes, M. A., and Rojas, R. G., Chimera-type states induced by local coupling, Phys. Rev. E, 93 (5), 052204, 2016.
6. Dhote, R. P., Gomez, H., Melnik, R. V. N., and Zu, J., 3D coupled thermo-mechanical phase-field modeling of shape memory alloy dynamics via isogeometric analysis, Computers and Structures, 154, 48—58, 2015.
7. Elliott, J. A., Novel approaches to multiscale modelling in materials science, International Materials Reviews, 56 (4), 207—225, 2011.
8. **Enge**, A., Elliptic Curves and Their Applications to Cryptography: An Introduction, Springer, 1999.
9. Fluid Structure Interaction and Moving Boundary Problems IV, Chakrabarti, S. K. and Brebbia, C. A. (Eds.), WIT, 2007.
10. Ganusov, V. V., Strong Interference in Mathematical Modeling: a Method for Robust Science in the Twenty-First Century, Frontiers in Microbiology, 7, 1131, 2016.
11. Geometric Numerical Integration of Differential Equations, R. Quispel and R. McLachlan (Guest Editors), Journal of Physics A: Mathematical and General, Special Issue, 39 (19), pp. 5251–5652, 2006.
12. Hitchin, N., Interaction between mathematics and physics, ARBOR Ciencia, Pensamiento y Cultura CLXXXIII, 725, 427-432, 2007.
13. Hunter, P., Biology is the new physics, EMBO Rep., 11 (5), 350—352, 2010.
14. Ibragimov, N. H., Integration of dynamical systems admitting nonlinear superposition, J. Coupled Syst. Multiscale Dyn. 4, 91–106, 2016.
15. Ingalls, B. P., Mathematical Modeling in Systems Biology. An Introduction, The MIT Press, 2013.
16. Kent, J., Jablonowski, C., Thuburn, J., and Wood N., An energy-conserving restoration scheme for the shallow-water equations, Quarterly Journal of the Royal Meteorological Society, 142 (695), 1100—1110., Part B, 2016.
17. Kraus, M., & Maj, O., Variational integrators for nonvariational partial differential equations, Physica D: Nonlinear Phenomena, 310, 37-71, 2015.
18. Le Novère, N., Quantitative and logic modelling of molecular and gene networks, Nature Reviews Genetics, 16, 146–158, 2015.
19. Mackey, M. C. and Santillán, M., Mathematics, Biology, and Physics: Interactions and Interdependence, Notices of the AMS, 52 (8), 832—840, 2005.
20. Marsden, J. E., Patrick, G. W., Shkoller, S., Multisymplectic Geometry, Variational Integrators, and Nonlinear PDEs, Commun. Math. Phys., 199, 351 – 395, 1998.
21. Mclachlan, R. I. and Wilkins, M. C., The Multisymplectic Diamond Scheme, SIAM Journal on Scientific Computing, 37 (1), A369-A390, 2015.
22. Multiscale Paradigms in Integrated Computational Materials Science and Engineering Materials Theory, Modeling, and Simulation for Predictive Design Introduction, Deymier, P. A., Runge, K., and Muralidharan, K. (Eds.), Springer, 2016.
23. Nonlinear Wave Equations: Analytic and Computational Techniques, Curtis, C., Dzhamay, A., Hereman, W. A., Prinari, B. (Eds.), Contemporary Mathematics, Vol. 635, 2015.

24. Scheffran, J., The dynamic interaction between economy and ecology: Cooperation, stability and sustainability for a dynamic-game model of resource conflicts, Mathematics and Computers in Simulation, 53 (4-6), 371—380, 2000.
25. Tadmor, E., Review of Numerical Methods for Nonlinear Partial Differential Equations, Bulletin of the American Mathematical Society, 49 (4), 507—554, 2012.
26. Tao, M., Explicit symplectic approximation of nonseparable Hamiltonians: algorithm and long time performance, Phys. Rev. E, Vol. 94, Issue 4, Article Number 043303, 2016.
27. Torres, O. P. and Granados, M. A., Exact traveling wave solutions in the coupled plane-base rotator model of DNA, International Journal of Non-Linear Mechanics, 86, 8—14, 2016.
28. Valsson, O., Tiwary, P., and Parrinello, M., Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint, in Book Series: Annual Review of Physical Chemistry, Johnson, M. A. and Martinez, T. J. (Eds.), Annual Review of Physical Chemistry, Vol. 67, pp. 159-184, 2016.

Part II
Large Deviation Theory and Random Perturbations of Dynamical Systems with Applications

Long Term Effects of Small Random Perturbations on Dynamical Systems: Theoretical and Computational Tools

Tobias Grafke, Tobias Schäfer, and Eric Vanden-Eijnden

Abstract Small random perturbations may have a dramatic impact on the long time evolution of dynamical systems, and large deviation theory is often the right theoretical framework to understand these effects. At the core of the theory lies the minimization of an action functional, which in many cases of interest has to be computed by numerical means. Here we review the theoretical and computational aspects behind these calculations, and propose an algorithm that simplifies the geometric minimum action method to minimize the action in the space of arc-length parametrized curves. We then illustrate this algorithm's capabilities by applying it to various examples from material sciences, fluid dynamics, atmosphere/ocean sciences, and reaction kinetics. In terms of models, these examples involve stochastic (ordinary or partial) differential equations with multiplicative noise, Markov jump processes, and systems with fast and slow degrees of freedom, which all violate detailed balance, so that simpler computational methods are not applicable.

1 Introduction

Small random perturbations often have a lasting effect on the long-time evolution of dynamical systems. For example, they give rise to transitions between otherwise stable equilibria, a phenomenon referred to as metastability that is observed in a wide variety of contexts, e.g. phase separation, population dynamics, chemical reactions, climate regimes, neuroscience, or fluid dynamics. Since the time-scale over which these transition events occurs is typically exponentially large in some control

T. Grafke (✉) • E. Vanden-Eijnden
Courant Institute, New York University, 251 Mercer Street, New York, NY 10012, USA
e-mail: grafke@cims.nyu.edu; eve2@cims.nyu.edu

T. Schäfer
Department of Mathematics, College of Staten Island 1S-215, 2800 Victory Blvd., Staten Island, NY 10314, USA

Physics Program at the CUNY Graduate Center, 365 5th Ave, New York, NY 10016, USA
e-mail: tobias.schaefer@csi.cuny.edu

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_2

parameter (for example the noise amplitude), a brute-force simulation approach to compute these events quickly becomes infeasible. Fortunately, it is possible to exploit the fact that the mechanism of these transitions is often predictable when the random perturbations have small amplitude: with high probability the transitions occur by their path of maximum likelihood (PML), and knowledge of this PML also permits to estimate their rate. This is the essence of large deviation theory (LDT) [20], which applies in a wide variety of contexts. For example, systems whose evolution is governed by a stochastic (ordinary or partial) differential equation driven by a small noise or by a Markov jump process in which jumps occur often but lead to small changes of the system state, or slow/fast systems in which the fast variables are randomly driven and the slow ones feel these perturbations through the effect fast variables only, all fit within the framework of LDT. Note that, typically, the dynamics of these systems fail to exhibit microscopic reversibility (detailed balance) and the transitions therefore occur out-of-equilibrium. Nevertheless, LDT still applies.

LDT also indicates that the PML is computable as the minimizer of a specific objective function (action): the large deviation rate function of the problem at hand. This is a non-trivial numerical optimization problem which calls for tailor-made techniques for its solution. Here we will focus on one such technique, the geometric minimum action method (gMAM, [26, 39, 39]), which is based on the minimum action method and its variants [17, 41, 44], and was designed to perform the action minimization over both the transition path location and its duration. This computation gives the so-called quasipotential, whose role is key to understand the long time effect of the random perturbations on the system, including the mechanism of transitions events induced by these perturbations. Our purpose here is twofold. First, we would like to briefly review the theoretical aspects behind LDT that lead to the rate function minimization problem and, in particular, to the geometric variant of it that is central in gMAM. Second, we would like to discuss in some details the computational issues this minimization entails, and remedy a drawback of gMAM, namely its somewhat complicated descent step that requires higher order derivatives of the large deviation Hamiltonian. Here, we propose a simpler algorithm, minimizing the geometric action functional, but requiring only first order derivatives of the Hamiltonian. The power of this algorithm is then illustrated via applications to a selection of problems:

1. the Maier-Stein model, which is a toy non-gradient stochastic ordinary differential equation that breaks detailed balance;
2. a stochastic Allen-Cahn/Cahn-Hilliard partial differential equation motivated by population dynamics;
3. the stochastic Burgers-Huxley PDE, related to fluid dynamics and neuroscience;
4. Egger's and Charney-DeVore equations, introduced as climate models displaying noise-induced transitions between metastable regimes;
5. a generalized voter/Ising model with multiplicative noise;

6. metastable networks of chemical reaction equations and reaction-diffusion equations;
7. a fast/slow system displaying transitions of the slow variables induced by the effects of the fast ones.

The remainder of this paper is organized as follows. In Sect. 2 we briefly review the key concepts of LDT that we will use (Sect. 2.1) and give a geometrical point of view of the theory that led to the action used in gMAM (Sect. 2.2). In Sect. 3 we discuss the numerical aspects related to the minimization of the geometric action, propose a simplified algorithm to perform this calculation, and compare it to existing algorithms. We also discuss further simplifications of the algorithm that apply in regularly occurring special cases, such as additive or multiplicative Gaussian noise. Finally, in Sect. 4 we present the applications listed above.

2 Freidlin-Wentzell Large Deviation Theory (LDT)

Here we first give a brief overview of LDT [20], focusing mainly on stochastic differential equations (SDEs) for simplicity, but indicating also how the theory can be extended to other models, such as Markov jump processes or fast/slow systems. Then we discuss the geometric reformulation of the action minimization problem that is used in gMAM.

2.1 Some Key Concepts in LDT

Consider the following SDE for $X \in \mathbb{R}^n$

$$dX = b(X)dt + \sqrt{\epsilon}\sigma(X)dW, \tag{1}$$

where $b : \mathbb{R}^n \to \mathbb{R}^n$ denotes the drift term, W is a standard Wiener process in $\mathbb{R}^n$, $\sigma : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ is related to the diffusion tensor via $a(x) = (\sigma\sigma^\dagger)(x)$, and $\epsilon > 0$ is a parameter measuring the noise amplitude. Suppose that we want to estimate the probability of an event, such as finding the solution in a set $B \subset \mathbb{R}^n$ at time T given that it started at $X(0) = x$ at time $t = 0$. LDT indicates that, in the limit as $\epsilon \to 0$, this probability can be estimated via a minimization problem:

$$\mathbb{P}^x\left(X(T) \in B\right) \asymp \exp\left(-\epsilon^{-1} \min_{\phi\in\mathscr{C}} S_T(\phi)\right). \tag{2}$$

Here $\asymp$ denotes log-asymptotic equivalence (i.e. the ratio of the logarithms of both sides tends to 1 as $\epsilon \to 0$), the minimum is taken over the set $\mathscr{C} = \{\phi \in C([0,T],\mathbb{R}^n) : \phi(0) = x, \phi(T) \in B\}$, and we defined the action functional

$$S_T(\phi) = \begin{cases} \int_0^T L(\phi, \dot{\phi})\, dt & \text{if the integral converges} \\ \infty & \text{otherwise.} \end{cases} \tag{3}$$

Here

$$L(\phi, \dot{\phi}) = \tfrac{1}{2}\langle \dot{\phi} - b(\phi), (a(\phi))^{-1} (\dot{\phi} - b(\phi))\rangle, \tag{4}$$

where we assumed for simplicity that $a(\phi)$ is invertible and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in $\mathbb{R}^n$. LDT also indicates that, as $\epsilon \to 0$, when the event occurs, it does so with X being arbitrarily close to the minimizer

$$\phi_* = \operatorname*{argmin}_{\phi \in \mathscr{C}} S_T(\phi) \tag{5}$$

in the sense that

$$\forall \delta > 0: \qquad \lim_{\epsilon \to 0} \mathbb{P}^x\Big(\sup_{0 \le t \le T} |X(t) - \phi_*(t)\,| \ < \delta \Big| X(T) \in B \Big) = 1$$

Thus, from a computational viewpoint, the main question becomes how to perform the minimization in (5). Note that, if we define the Hamiltonian associated with the Lagrangian (4)

$$H(\phi, \theta) = \langle b(\phi), \theta \rangle + \tfrac{1}{2}\langle \theta, a(\phi)\theta \rangle \tag{6}$$

such that

$$L(\phi, \dot{\phi}) = \sup_{\theta} \big(\langle \dot{\phi}, \theta \rangle - H(\phi, \theta) \big), \tag{7}$$

this minimization reduces to the solution of Hamilton's equations of motion,

$$\begin{cases} \dot{\phi} = H_\theta(\phi, \theta) = b(\phi) + a(\phi)\theta \\ \dot{\theta} = -H_\phi(\phi, \theta) = -(b_\phi(\phi))^T \theta + \tfrac{1}{2}\langle \theta, a_\phi(\phi)\theta \rangle, \end{cases} \tag{8}$$

where subscripts denote differentiation and we use the convention $(b_\phi)_{ij} = \partial b_i / \partial \phi_j$. What makes the problem nonstandard, however, is the fact that these equations must be solved as a boundary value problem, with $\phi(0) = x$ and $\phi(T) = y \in B$. We will come back to this issue below.

If the minimum of the action in (2) is nonzero, this equation indicates that the probability of finding the solution in B at time T is exponentially small in ϵ, i.e. it is a rare event. This is typically the case if one considers events that occur on a finite time interval, $T < \infty$ fixed. LDT, however, also permits to analyze the effects of the perturbations over an infinite time span, in which case they become ubiquitous. In this context, the central object in LDT is the quasipotential defined as

$$V(x,y) = \inf_{T>0} \min_{\phi\in\mathscr{C}_{x,y}} S_T(\phi), \tag{9}$$

where $\mathscr{C}_{x,y} = \{\phi \in C([0,T],\mathbb{R}^n) : \phi(0) = x, \phi(T) = y\}$. The quasipotential permits to answer several questions about the long time behavior of the system. For example, if we assume that the deterministic equation associated with (1), $\dot{X} = b(X)$, possesses a single stable fixed point, x_a, as unique stable structure, and that (1) admits a unique invariant distribution, the density associated with this distribution can estimated as $\epsilon \to 0$ as

$$\rho(x) \asymp \exp\left(-\epsilon^{-1} V(x_a, x)\right). \tag{10}$$

Similarly, if $\dot{X} = b(X)$ possesses two stable fixed points, x_a and x_b, whose basins of attraction have a common boundary, we can estimate the mean first passage time the system takes to travel for one fixed point to the other as

$$\mathbb{E}\tau_{a\to b} \asymp \exp\left(\epsilon^{-1} V(x_a, x_b)\right), \tag{11}$$

where

$$\tau_{a\to b} = \inf\{t : X(t) \in B_\delta(x_b), X(0) = x_a\}, \tag{12}$$

in which $B_\delta(x_b)$ denotes the ball of radius δ around x_b, with δ small enough so that this ball is contained in the basin of attraction of x_b. In this setup, we can also estimate the ratio of the stationary probabilities to find the system in the basins of attraction of x_a or x_b. Denoting these probabilities by p_a and p_b, respectively, we have

$$\frac{p_a}{p_b} \asymp \frac{\mathbb{E}\tau_{a\to b}}{\mathbb{E}\tau_{b\to a}} \asymp \exp\left(\epsilon^{-1}(V(x_a, x_b) - V(x_b, x_a))\right). \tag{13}$$

These statements can be generalized to many other situations, e.g. if $\dot{X} = b(X)$ possesses more than two stable fixed points, or attracting structures that are more complicated than points, such as limit cycles. They can also be generalized to dynamical systems other than (1), e.g. if this equation is replaced by a stochastic partial differential equation (SPDE), or for Markov jump processes in which the jump rates are fast but lead to small changes of the system's state [20, 35], or in slow/fast systems where the slow variables feels random perturbations through the effect the fast variables have on them [6, 19, 28, 29, 40]. In all cases, LDT provides us with an action functional like (3), but in which the Lagrangian is different from (4) if the system's dynamics is not governed by an S(P)DE. Typically, the theory yields an expression for the Hamiltonian (6), which may be non-quadratic in the momenta, or even such that the Legendre transform in (7) is not available analytically. This *per se* is not an issue, since we can in principle minimize the action by solving Hamilton's equations (8). However, these calculations face two difficulties. The first, already mentioned above, is that (8) must be solved as a boundary value problem. The second, which is specific to the calculation of the quasipotential in (9), is that the time span over which (8) are solved must be varied as well since (9) involves a minimization over T, and typically the minimum is reached as $T \to \infty$ (i.e. there is a minimizing sequence but no minimizer) which complicates matters even more. These issues motivate a geometric reformulation of the problem, which was first proposed in [25] and we recall next.

2.2 Geometric Action Functional

As detailed in [25] (see Proposition 2.1 in that paper), the quasipotential defined in (9) can also be expressed as

$$V(x,y) = \min_{\varphi \in \hat{\mathscr{C}}_{x,y}} \hat{S}(\varphi), \tag{14}$$

where $\hat{\mathscr{C}}_{x,y} = \{\varphi \in C([0,1],\mathbb{R}^n) : \varphi(0) = x, \varphi(1) = y\}$ and $\hat{S}(\varphi)$ is the geometric action that can be defined in the following equivalent ways:

$$\hat{S}(\varphi) = \sup_{\vartheta : H(\varphi,\vartheta)=0} \int_0^1 \langle \varphi', \vartheta \rangle ds \tag{15a}$$

$$\hat{S}(\varphi) = \int_0^1 \langle \varphi', \vartheta_*(\varphi,\varphi') \rangle ds \tag{15b}$$

$$\hat{S}(\varphi) = \int_0^1 \frac{1}{\lambda(\varphi,\varphi')} L(\varphi, \lambda\varphi') ds, \tag{15c}$$

where $\vartheta_*(\varphi,\varphi')$ and $\lambda(\varphi,\varphi')$ are the solutions to

$$H(\varphi, \vartheta_*(\varphi,\varphi')) = 0, \qquad H_\vartheta(\varphi, \vartheta_*(\varphi,\varphi')) = \lambda(\varphi,\varphi')\varphi' \quad \text{with } \lambda \geq 0. \tag{16}$$

The action $\hat{S}(\varphi)$ has the property that its value is left invariant by reparametrization of the path φ, i.e. it is an action on the space of continuous curves. In particular, one is free to choose arclength-parametrization for φ, e.g. $|\varphi'| = 1/L$ for $\int |\varphi'| \, ds = L$. This also means that the minimizer of (14) exists in more general cases (namely as long as the path has finite length), which makes the minimization problem easier to handle numerically, as shown next.

3 Numerical Minimization of the Geometric Action

From (14), we see that the calculation of the quasipotential reduces to a minimization problem, whose Euler-Lagrange equation is simply

$$D_\phi \hat{S}(\varphi) = 0, \tag{17}$$

where D_φ denotes the functional gradient with respect to φ. The main issue then becomes how to find the solution φ_* to (17) that minimize the action $\hat{S}(\varphi)$. In this section, we first briefly review how the gMAM achieves this task. We will then introduce a simplified variant of the gMAM algorithm that in its simplest form relies solely on first order derivatives of the Hamiltonian. Subsequently, we also analyze several special cases where the numerical treatment can be simplified even further.

3.1 Geometric Minimum Action Method

The starting point of gMAM is the following expression involving $D_\phi \hat{S}(\varphi)$ that can be calculated directly from formula (15b) for the action functional:

$$-\lambda H_{\vartheta\vartheta} D_\varphi \hat{S}(\varphi) = \lambda^2 \varphi'' - \lambda H_{\vartheta\varphi} \varphi' + H_{\vartheta\vartheta} H_\varphi + \lambda\lambda' \varphi'. \tag{18}$$

This is derived as Proposition 3.1 in Appendix E of [25], and we will show below how this expression can be intuitively understood. Since $H_{\vartheta\vartheta}$ is assumed to be positive definite and $\lambda \geq 0$, we can use (18) directly to compute the solution of (17) that minimizes $\hat{S}(\varphi)$ via a relaxation method in virtual time τ, that is, using the equation:

$$\begin{aligned} \frac{\partial \varphi}{\partial \tau} &= -\lambda H_{\vartheta\vartheta} D_\varphi \hat{S}(\varphi) \\ &= \lambda^2 \varphi'' - \lambda H_{\vartheta\varphi} \varphi' + H_{\vartheta\vartheta} H_\varphi + \lambda\lambda' \varphi'. \end{aligned} \tag{19}$$

This equation is the main equation used in the original gMAM. Note that the computation of the right hand-side of this equation requires the computation of H_φ, $H_{\vartheta\varphi}$ and $H_{\vartheta\vartheta}$, where the second derivatives of the Hamiltonian possibly become unsightly for more complicated systems that arise naturally when trying to use gMAM in practical applications. In Sect. 3.2 we propose a simplification of this algorithm that reduces the terms necessary to only first order derivatives of the Hamiltonian, H_ϑ and H_φ.

Coming back to (18), it can be intuitively understood by using the associated Hamiltonian system. Consider a reparametrization of the original minimizer $\varphi_*(s(t)) = \phi_*(t)$. In the following we are using a dot in order to denote partial derivatives with respect to time and a prime in order to denote a partial derivative with respect to the parametrization s, hence $\dot{v} \equiv \partial v/\partial t$ and $v' \equiv \partial v/\partial s$. With this notation, we find for $\lambda^{-1} = t'(s)$ that $\dot{\phi}_* = \lambda\varphi'_*$ as well as $\dot{\phi}_* = H_\theta, \dot{\theta}_* = -H_\phi$, and therefore

$$\begin{aligned}\ddot{\phi}_* &= H_{\theta\phi}\dot{\phi}_* + H_{\theta\theta}\dot{\theta}_*\\ &= \lambda H_{\theta\phi}\varphi'_* - H_{\theta\theta}H_\phi\end{aligned}$$

but also, since $\partial/\partial t = \lambda\,\partial/\partial s$,

$$\begin{aligned}\ddot{\phi}_* &= \partial(\lambda\varphi'_*)/\partial t\\ &= \lambda\lambda'\varphi'_* + \lambda^2\varphi''_*\end{aligned}$$

so in total

$$-\lambda\lambda'\varphi'_* + \lambda H_{\theta\phi}\varphi'_* - H_{\theta\theta}H_\phi - \lambda^2\varphi''_* = 0 = \lambda H_{\theta\theta}D_\varphi\hat{S}(\varphi),$$

i.e. indeed the gradient vanishes at the minimizer.

3.2 A Simplified gMAM

In contrast to the previous section, we start from the form (15a) of the geometric action. We want to solve the mixed optimization problem, i.e. find a trajectory φ_* such that

$$\varphi_* = \operatorname*{argmin}_{\varphi\in\hat{\mathscr{C}}_{x,y}} \sup_{\vartheta:H(\varphi,\vartheta)=0} E(\varphi,\vartheta), \tag{20}$$

where

$$E(\varphi,\vartheta) = \int_0^1 \langle\varphi',\vartheta\rangle\,ds. \tag{21}$$

Let

$$E_*(\varphi) = \sup_{\vartheta:H(\varphi,\vartheta)=0} E(\varphi,\vartheta) \tag{22}$$

and $\vartheta_*(\varphi)$ such that $E_*(\varphi) = E(\varphi,\vartheta_*(\varphi))$. This implies that ϑ_* fulfills the Euler-Lagrange equation associated with the constrained optimization problem in (22), that is,

$$D_\vartheta E(\varphi,\vartheta_*) = \mu H_\vartheta(\varphi,\vartheta_*), \tag{23}$$

where on the right-hand side $\mu(s)$ is the Lagrange multiplier added to enforce the constraint $H(\varphi, \vartheta_*) = 0$. In particular, at $\vartheta = \vartheta_*$, we have

$$\mu = \frac{\|D_\vartheta E\|^2}{\langle\langle D_\vartheta E, H_\vartheta \rangle\rangle} = \frac{\|\varphi'\|^2}{\langle\langle \varphi', H_\vartheta \rangle\rangle}, \tag{24}$$

where the inner product $\langle\langle\cdot,\cdot\rangle\rangle$ and its induced norm $\|\cdot\|$ can be chosen appropriately, for example as $\langle\cdot,\cdot\rangle$ or $\langle\cdot, H_{\vartheta\vartheta}^{-1}\cdot\rangle$.

At the minimizer φ_*, the variation of E_* with respect to φ vanishes. Using (23) we conclude

$$\begin{aligned} 0 = D_\varphi E_*(\varphi_*) &= D_\varphi E(\varphi_*, \vartheta_*) + \left[D_\vartheta E D_\varphi \vartheta\right]_{(\varphi,\vartheta)=(\varphi_*,\vartheta_*)} \\ &= -\vartheta_*' + \mu \left[H_\vartheta D_\varphi \vartheta\right]_{(\varphi,\vartheta)=(\varphi_*,\vartheta_*)} \\ &= -\vartheta_*' - \mu H_\varphi(\varphi_*, \vartheta_*), \end{aligned} \tag{25}$$

where in the last step we used $H(\varphi, \vartheta_*) = 0$ and therefore

$$H_\varphi(\varphi, \vartheta_*) = -H_\vartheta(\varphi, \vartheta_*) D_\varphi \vartheta.$$

Multiplying the gradient (25) with any positive definite matrix as pre-conditioner yields a descent direction. It is necessary to choose μ^{-1} as pre-conditioner to ensure convergence around critical points, where $\varphi' = 0$.

Summarizing, we have reduced the minimization of the geometric action into two separate tasks:

1. For a given φ, find $\vartheta_*(\varphi)$ by solving the constrained optimization problem

$$\vartheta_*(\varphi) = \underset{\vartheta, H(\varphi,\vartheta)=0}{\operatorname{argmax}} E(\varphi, \vartheta), \tag{26}$$

which is equivalent to solving

$$D_\vartheta E(\varphi, \vartheta_*) = \varphi' = \mu H_\vartheta(\varphi, \vartheta_*) \tag{27}$$

for (μ, ϑ_*) under the constraint $H(\varphi, \vartheta_*) = 0$. This can be done via

- gradient descent;
- a second order algorithm for faster convergence (e.g. Newton-Raphson, as employed in [25]);
- in many cases, analytically (see below).

2. Find φ_* by solving the optimization problem

$$\varphi_* = \underset{\varphi \in \hat{\mathscr{C}}_{x,y}}{\operatorname{argmin}} E_*(\varphi), \tag{28}$$

for example by pre-conditioned gradient descent, using as direction

$$-\mu^{-1}D_\varphi E_* = \mu^{-1}\vartheta_*'(\varphi) + H_\varphi(\varphi, \vartheta_*(\varphi)), \tag{29}$$

with μ^{-1} as pre-conditioner. The constraint on the parametrization, e.g. $|\varphi'| = \text{const}$, must be fulfilled during this descent (see below).

3.3 Connection to gMAM

The problem of finding $\vartheta_*(\varphi)$ is equivalent to (16) from gMAM and the same methods are applicable. In particular note that the Lagrange multiplier μ which enforces $H(\varphi_*, \vartheta_*) = 0$ is identical to λ^{-1}.

It is also easy to see that, at (φ_*, ϑ_*), the combined optimization problem $\{D_\vartheta E = \mu H_\vartheta, D_\varphi E_* = 0\}$ is identical to the geometric equations of motion,

$$\begin{cases} D_\vartheta E = \varphi' = \mu H_\vartheta \\ D_\varphi E_* = -\vartheta' - \mu H_\varphi = 0. \end{cases} \tag{30}$$

On the other hand, none of the formulas in the above section use higher derivatives of the Hamiltonian: Only H_φ and H_ϑ are needed, which is a significant simplification. This is obviously also true for the equations of motion (8) and their geometric variant (30), which is the basis for the efficiency of algorithms like [11, 21, 22].

3.4 Simplifications for SDEs with Additive Noise

For an SDE of the form

$$dX = b(X)dt + \sqrt{\epsilon}\, dW, \tag{31}$$

where $\sigma = \text{Id}$, the equations of gMAM become significantly simpler. In the following, we derive explicit expressions for this case, as it arises in numerous applications.

The corresponding Hamiltonian is given by

$$H(\varphi, \vartheta) = \langle b, \vartheta\rangle + \frac{1}{2}\langle \vartheta, \vartheta\rangle = 0 \tag{32}$$

and we find directly

$$H_\varphi = (b_\varphi)^T\vartheta, \qquad H_\vartheta = b + \vartheta.$$

In many cases, we consider exits from stable fixed points of the deterministic system where we have $H = 0$ which, if we also use $D_\vartheta E = \mu H_\vartheta$, permits to conclude that

$$|H_\vartheta|^2 = |b+\vartheta|^2 = |b|^2 + 2\langle b,\vartheta\rangle + \langle \vartheta,\vartheta\rangle = |b|^2 + 2H = |b|^2. \qquad (33)$$

As a result

$$\mu = \frac{|D_\vartheta E|}{|H_\vartheta|} = \frac{|\varphi'|}{|b+\vartheta|} = \frac{|\varphi'|}{|b|}, \qquad (34)$$

i.e. we can compute μ without the knowledge of ϑ. On the other hand (27) implies

$$\varphi' = \mu H_\vartheta = \mu(b+\vartheta) \quad \Rightarrow \quad \vartheta = \mu^{-1}\varphi' - b. \qquad (35)$$

The whole algorithm therefore reduces to the gradient descent

$$\frac{\partial\varphi}{\partial\tau} = \mu^{-1}\vartheta_*' + (b_\varphi)^T\vartheta_*, \qquad (36)$$

with μ, ϑ_* given by (34) and (35). Examples in this class will be treated in Sects. 4.1, 4.2, and 4.4 below.

3.5 Simplifications for General SDEs (Multiplicative Noise)

As a slightly more complicated case, consider the following SDE with multiplicative noise:

$$dX = b(X)\,dt + \sqrt{\epsilon}\sigma(X)\,dW, \qquad (37)$$

where $a(\varphi) = \sigma(\varphi)\sigma^\dagger(\varphi)$. Then the Hamiltonian reads

$$H(\varphi,\vartheta) = \langle b,\vartheta\rangle + \tfrac{1}{2}\langle\vartheta, a\vartheta\rangle \qquad (38)$$

and

$$H_\varphi = (b_\varphi)^T\vartheta + \tfrac{1}{2}\langle\vartheta,(a_\varphi)\vartheta\rangle, \qquad H_\vartheta = b + a\vartheta. \qquad (39)$$

Defining an inner product and norm induced by the correlation, $\langle u,v\rangle_a = \langle u, a^{-1}v\rangle$ and $|u|_a = \langle u,u\rangle_a^{1/2}$ yields, as before,

$$|H_\vartheta|_a = |b|_a \qquad \Rightarrow \qquad \mu = \frac{|\varphi'|_a}{|b|_a} \qquad (40)$$

and

$$\vartheta = a^{-1}(\mu^{-1}\varphi' - b). \tag{41}$$

In the case of multiplicative noise, the algorithm therefore reads

$$\frac{\partial\varphi}{\partial\tau} = \mu^{-1}\vartheta_*' + \left((b_\varphi)^T\vartheta_* + \tfrac{1}{2}\langle\vartheta_*, (a_\varphi)\vartheta_*\rangle\right), \tag{42}$$

with μ, ϑ_* given by (40) and (41). An example in this will be treated in Sect. 4.5.

It is also worth pointing out that we encounter difficulties as soon as the noise correlation a is not invertible. This is equivalent to stating that some degrees of freedom are not subject to noise and thus behave deterministically. The adjoint field ϑ has to be equal to zero on these modes, and they fulfill the deterministic equation $\varphi' = b$ exactly. This translates into additional constraints for the minimization procedure, which have to be enforced numerically.

3.6 Comments on Improving the Numerical Efficiency

To increase the numerical efficiency of the algorithm, some alterations are possible:

- Arc-length parametrization, $|\varphi'| = \text{const}$, can be enforced trivially and without introducing a stiff Lagrange multiplier term by interpolation along the trajectory every (or every few) iterations. As additional benefit of this method all terms of the relaxation dynamics which are proportional to φ' can be discarded, as they are canceled by the reparametrization. This is of particular use in applications that involve PDEs (see Sect. 3.7), as shown in examples below.
- Stability in the relaxation parameter can be greatly increased if one treats the stiffest term of the relaxation equation implicitly. In ODE systems, the stiffest term usually is $H_{\vartheta\vartheta}^{-1}\varphi''$, which is contained in ϑ'. For simplicity of implementation, it is sufficient to compute ϑ_* in the usual way, apply ϑ_*' in the descent step, but subtract $H_{\vartheta\vartheta}^{-1}\varphi_n''$ and add $H_{\vartheta\vartheta}^{-1}\varphi_{n+1}''$ here. This approach also extends to the case of general Hamiltonians, where the dependence of ϑ_* on φ' is less obvious.

  In our implementation, the relaxation step is conducted by computing

$$\varphi_{n+1} = \left(1 - h\mu^{-2}H_{\vartheta\vartheta}^{-1}\partial_s^2\right)^{-1} R_n, \tag{43}$$

  where

$$R_n = \left(\varphi_n + h(\mu^{-1}\vartheta_*'(\varphi_n) + H_\varphi(\varphi_n, \vartheta_*(\varphi_n)) - \mu^{-2}H_{\vartheta\vartheta}^{-1}\varphi_n'')\right).$$

This division into an implicit treatment of the stiffest term and explicit treatment of the rest is the simplest case of Strang splitting [36] and the implementation of (43) is only first order accurate. The splitting can be taken to arbitrary order [43] under additional computational cost.

Note that the above modification, while increasing efficiency, at the same time increases complexity, as the computation of the second derivative $H_{\vartheta\vartheta}$ becomes necessary. In practice, if the Hamiltonian is not too complex, we find that the benefits outweigh the implementation costs, and some problems, especially PDE systems, are not tractable at all with the inefficient but simpler choice of explicit relaxation. If the PDE system contains higher-order spatial derivatives, even more terms should possibly be treated with a stable integrator, as is discussed in the next section.

- Depending on the problem, it might be beneficial to choose a different scalar product in the descent. In case of traditional gMAM, the descent is done using $\langle \cdot, (\mu^2 H_{\vartheta\vartheta})^{-1} \cdot \rangle$, but other choices are also feasible. Note that it is possible to choose the metric such that at least one term at the right-hand side disappears, as it becomes parallel to the trajectory and is canceled by reparametrization, as outlined above.
- Some insight about the nature of the transition can be obtained by first finding the heteroclinic orbits defined geometrically as

$$\varphi' \parallel b(\varphi). \tag{44}$$

This calculation can be done very efficiently even for complicated problems via the string method [16]. Even though the heteroclinic orbit differs from the transition path for systems that violate detailed balance, it *does* correctly predict the transition from the saddle point onward (the “downhill” portion, which happens deterministically). The method put forward here can then be used to find the transition path up to the saddle (the “uphill” portion) only. If there are several saddles to be taken into account, it is not known *a priori* which one will be visited by the transition pathway. In this case, the strategy has to be modified accordingly, for example by computing one heteroclinic orbit per saddle. To highlight the relation between the string and the minimizer, we compute and compare the two in many of the applications below. We denote with “string” the heteroclinic orbits connecting the fixed points to the saddle point of relevance found via the string method.

3.7 SPDEs with Additive Noise

In this section, we discuss the application to SPDE systems. For simplicity, we focus on the case of SPDEs with additive noise that can be written formally as

$$U_t = B(U) + \sqrt{\epsilon}\,\eta(x,t), \tag{45}$$

where the drift term is given by the operator $B(U)$ and η denotes spatio-temporal white-noise. It is a non-trivial task to make mathematical sense of such SPDEs under spatially irregular noise due to the possible ill-posedness of non-linear terms, especially if the spatial dimension is higher than one. This may require to renormalize the equation, which can be done rigorously in certain cases using the theory of regularity structures [23]. The renormalization procedure typically involves mollifying the noise term on a scale δ, and adding terms in the equation that counterbalance divergences that may occur as one lets $\delta \to 0$. In the context of LDT, the main issue is whether these renormalizing terms subsist if we also let $\epsilon \to 0$. In [24], it was shown in the context of the stochastic Allen-Cahn equation in 2 or 3 spatial dimensions that the action of the mollified equation converges towards the action associated with the (possibly formal) equation in (45) in which the noise is white-in-space provided that ϵ is sent to zero fast enough as $\delta \to 0$. This action reads

$$S_T(\phi) = \frac{1}{2}\int_0^T \|\phi_t - B(\phi)\|_{L^2}^2 dt, \tag{46}$$

where $\|\cdot\|_{L^2}$ denotes the L^2-norm. This leads to expressions for the geometric action that are similar to those in (15) but with the Euclidean inner product replaced by the L^2-inner product. In the sequel we will not dwell further on these mathematical issues and always assume that (46) and the associated geometric action are the relevant one to study.

The gradient descent for the minimizer of this geometric action is similar to the one in (36) but with the term $(b_\varphi)^T$ replaced by the functional derivative of the operator B with respect to φ.

$$\frac{\partial\varphi}{\partial\tau} = \mu^{-1}\vartheta_*' + \left(D_\varphi B\right)^T \vartheta_* \,. \tag{47}$$

In practice, however, this equation needs to be rewritten in order to allow for numerical stability. This is due to the fact that the scheme will contain derivatives of high orders, and their corresponding stability condition (CFL condition) will limit the rate of convergence of the scheme. We therefore want to treat the most restrictive terms either implicitly or with exponential integrators. To this end, let us focus on the following class of problems where the drift B can be written as

$$B = L\varphi + R(\varphi), \tag{48}$$

where L is a linear self-adjoint operator containing higher-order derivatives that does not depend on time explicitly, and $R(\varphi)$ is the rest, possibly nonlinear. Recall that ϑ_* can be computed from φ' via

$$\vartheta_* = \mu^{-1}\varphi' - B = \mu^{-1}\varphi' - L\varphi - R(\varphi). \tag{49}$$

On the other hand, we have also a term proportional to L in

$$D_\varphi B = D_\varphi R + L \tag{50}$$

and, therefore, the relaxation formula (47) for φ actually contains a term $L^2\varphi$. If L contains higher-order derivatives, this term will likely be the most restrictive in terms of numerical stability. It is therefore advantageous to treat it separately. Introducing an auxiliary variable $\tilde{\vartheta}_*$ defined by

$$\tilde{\vartheta}_* = \mu^{-1}\varphi' - R(\varphi) = \vartheta_* + L\varphi \tag{51}$$

we can rewrite the relaxation formula as

$$\begin{aligned}
\frac{\partial\varphi}{\partial\tau} &= \mu^{-1}\vartheta_*' + \left(D_\varphi B\right)^T \vartheta_* \\
&= \mu^{-1}\tilde{\vartheta}_*' - \mu^{-1}L\varphi' + \left(D_\varphi R\right)^T \vartheta_* + L\vartheta_* \\
&= \tilde{\mu}^{-1}\vartheta_*' - \mu^{-1}L\varphi' + \left(D_\varphi R\right)^T \vartheta_* + L(\tilde{\vartheta}_* - L\varphi) \\
&= \mu^{-1}\tilde{\vartheta}_*' - \mu^{-1}L\varphi' + \left(D_\varphi R\right)^T \vartheta_* + L\tilde{\vartheta}_* - L^2\varphi \\
&= \mu^{-1}\tilde{\vartheta}_*' + \left(D_\varphi R\right)^T \vartheta_* - LR(\varphi) - L^2\varphi.
\end{aligned}$$

The term $L^2\varphi$ is now separated and can be treated independently. Since it is linear by definition, it can be treated very efficiently with an integrating factor by employing exponential time differencing (ETD) [5]. For an equation with a deterministic term of the form (48), multiplying by the integrating factor $e^{-L\tau}$ and integrating from τ_n to $\tau_{n+1} = \tau_n + h$, one obtains the *exact* formula

$$\varphi_{n+1} = e^{Lh}\varphi_n + e^{Lh}\int_0^h e^{-L\tau} R(\varphi(t_n+\tau))\,d\tau, \tag{52}$$

which can be approximated by

$$\varphi_{n+1} = e^{Lh}\varphi_n + (e^{Lh} - \mathrm{Id})L^{-1}R(\varphi_n), \tag{53}$$

when treating the linear part of the equation exactly and approximating the integral to first order. This scheme can be taken to higher order [12] and its stability improved [27], but a first order scheme proved to be sufficient for the examples given below. For the descent (47) we want to treat the stiffest part $-L^2\varphi$ with ETD, so the integrating factor here becomes $e^{-L^2\tau}$.

A complete relaxation step then consists of

1. compute ϑ_* and $\tilde{\vartheta}_*$ using the explicit formulas

$$\tilde{\vartheta}_* = \mu^{-1}\varphi' - R(\varphi), \qquad \vartheta_* = \tilde{\vartheta}_* - L\varphi\,;$$

2. compute the explicit step

$$\xi = \mu^{-1}\tilde{\vartheta}'_* + \left(D_\varphi R\right)^T \vartheta_* - LR(\varphi) - \mu^{-2} H_{\vartheta\vartheta}^{-1} \varphi''_n \,,$$

   where as in the SDE case, if needed, we can subtract the term $\mu^{-2}H_{\vartheta\vartheta}^{-1}\varphi''_n$ to treat it implicitly later;
3. perform an ETD step

$$\bar{\varphi} = e^{-L^2 h}\varphi_n - (e^{-L^2 h} - \mathrm{Id})(L^2)^{-1}\xi \,;$$

4. apply the second derivative in arc-length direction implicitly,

$$\varphi_{n+1} = (1 - h\mu^{-2} H_{\vartheta\vartheta}^{-1} \partial_t^2)^{-1} \bar{\varphi}.$$

Note that the integral factors $e^{-L^2 h}$ and $(e^{-L^2 h} - \mathrm{Id})(\mu L^2)^{-1}$ are possibly costly to compute, as they contain matrix-exponentials and inversions. However, the computation can be done once before starting the iteration, so that the associated computational cost becomes negligible. In contrast, this is not true in general for the implicit step 4, since $\mu^{-2}H_{\vartheta\vartheta}^{-1}$ might depend on the fields in a complicated way and has to be recomputed at every iteration.

4 Illustrative Applications

In what follows we apply our simplified gMAM to the series of examples listed in the introduction. These examples illustrate specific questions encountered in practical applications arising in a variety of fields, in which the computation of the rate and mechanism of transitions is of interest. Note that all these examples involve non-equilibrium systems whose dynamics break detailed balance, so that simpler methods of computation are not readily available.

In the following, we will break our notation convention and instead use the notation of the respective fields to minimize confusion.

4.1 Maier-Stein Model

Maier and Stein's model [31] is a simple system often used as benchmark in LDT calculations. It reads

$$\begin{cases} du = (u - u^3 - \beta u v^2)dt + \sqrt{\epsilon}dW_u \\ dv = -(1 + u^2)v dt + \sqrt{\epsilon}dW_v, \end{cases} \tag{54}$$

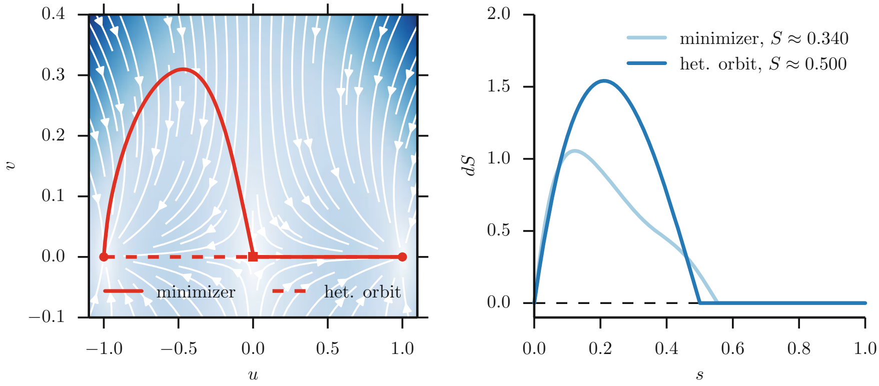


Fig. 1 Maier-Stein model, $\beta = 10$. Left: PML and heteroclinic orbit. The arrows denote the direction of the deterministic flow, the shading its magnitude. The solid line depicts the minimizer, the dashed line the heteroclinic orbit. Dots are located at the fixed points (circle: stable; square: saddle). Right: Action density along the minimizer and the heteroclinic orbit

where β is a parameter. For all values of β, the deterministic system has the two stable fixed points, $\varphi_- = (-1, 0)$ and $\varphi_+ = (1, 0)$, and a unique unstable critical point $\varphi_s = (0, 0)$. However it satisfies detailed balance only for $\beta = 1$. In this case, we can write the drift in gradient form, $b(\varphi) = \nabla_\varphi U(\varphi)$, and the minimizers of the geometric action that connects φ_- to φ_+ and *vice-versa* are the time-reverse of each other and lie on the location of the heteroclinic orbit where $\varphi' \parallel \nabla U$. Here, we use $\beta = 10$, in which case detailed balance is broken and the forward and backward transition pathways are no longer identical. Since the noise is additive, the system (54) falls into the category discussed in Sect. 3.4 and can be solved with the simplest variant of the algorithm. The minimizer of the action connecting φ_- to φ_+ and the value of the action along it are shown in Fig. 1. Since the system is invariant under the transformation $v \to -v$, there is also a minimizer with identical action in the $v < 0$ half-plane. Similarly, the paths from φ_+ to φ_- can be obtained via the transformation $u \to -u$. The numerical parameters used in these calculations were $h = 10^{-1}$, $N_s = 2^{10}$, where N_s denotes the number of configurations along the transition trajectory or the number of *images*.

4.2 Allen-Cahn/Cahn-Hilliard System

Pattern formation in motile micro-organisms is often driven by non-equilibrium forces, leading to visible patterns in cellular colonies [8, 34]. For example, *E. coli* in a uniform suspension separates into a bacteria-rich and a bacteria-poor phase if the swim speed decreases sufficiently rapidly with density [37]. Here we study a model inspired by these phenomena. We note that this model does not permit the thermodynamic mapping used in [37], so that understanding the non-equilibrium transitions in the model requires minimization of the geometric action of LDT.

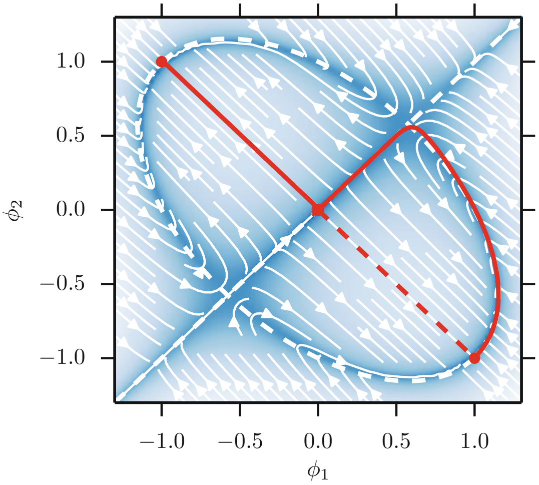


Fig. 2 Allen-Cahn/Cahn-Hilliard toy ODE model, $\alpha = 0.01$. The arrows denote the direction of the deterministic flow, the color its magnitude. The white dashed line corresponds to the slow manifold. The solid line depicts the minimizer, the dashed line the heteroclinic orbit. Markers are located at the fixed points (circle: stable; square: saddle)

4.2.1 Reduced Allen-Cahn/Cahn-Hilliard Model

Consider the SDE system

$$d\phi = \left(\frac{1}{\alpha}Q(\phi - \phi^3\right) - \phi)dt + \sqrt{\epsilon}dW \tag{55}$$

with $\phi = (\phi_1, \phi_2)$ and the matrix $Q = ((1, -1), (-1, 1))$. This system does not satisfy detailed balance, as its drift is made of two gradient terms with incompatible mobility operators (namely Q and Id). Model (55) can be seen as a 2-dimensional reduction to a discretized version of the continuous Allen-Cahn/Cahn-Hilliard model discussed later in Sect. 4.2.2.

The deterministic flowlines of (55) are depicted in Fig. 2. The deterministic dynamics has two stable fixed points, $\phi_A = (-1, 1)$ and $\phi_B = (1, -1)$, and an unstable critical point, $\phi_S = (0, 0)$, lying on the separatrix where $\phi_1 = \phi_2$ between the basins of ϕ_A and ϕ_B. The location of the heteroclinic orbits connecting ϕ_S to ϕ_A and ϕ_B is a straight line between these points. When α is small in (55), there exists a "slow manifold", comprised of all points where $Q(\phi - \phi^3) = 0$ which is shown as a white dashed line in Fig. 2. On this manifold, the deterministic dynamics are of order $O(1)$, which is small in comparison to the dynamics of the Q-term, which are of order $O(1/\alpha)$. This suggests that for small enough α the transition trajectory will follow this slow manifold on which the drift is small, rather than the heteroclinic orbit, to escape the basin of the stable fixed points. This is confirmed in Fig. 2 where we show the action minimizer connecting ϕ_B to ϕ_A. As can be seen, the minimizer first tracks the slow manifold, and it approaches the separatrix at a point far from ϕ_S. It then follows closely the separatrix towards ϕ_S (which has to be part of the transition) to cross into the other basin and then relax (deterministically) towards ϕ_A.

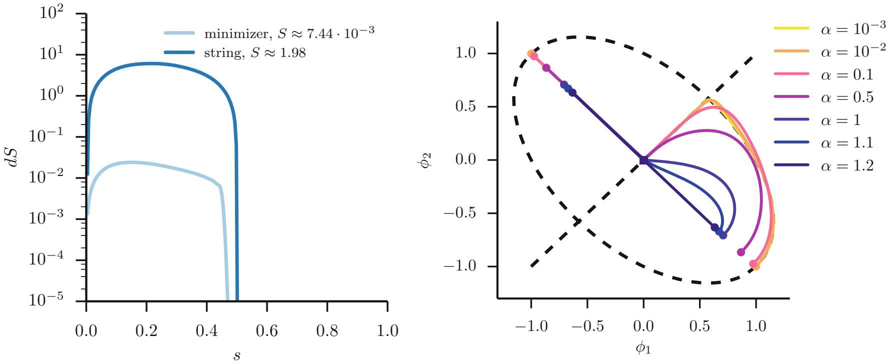


Fig. 3 Left: Action density along the path for the 2-dimensional reduced model. Path parameter is normalized to $s \in (0, 1)$. For the second half of the transition, the action density is zero. Right: Minimizers of the action functional for different values of α. For $\alpha \to 0$, the minimizer approaches the slow manifold. Note that the switch to a straight line minimizer happens at a finite value $\alpha \approx 1.12$

The action along the minimizer and the paths made of the heteroclinic orbits are depicted in Fig. 3 (left). Notably, due to its movement along the slow manifold, the action along the minimizer is smaller by a factor of order α. Minimizers for different values of α are shown in Fig. 3 (right). Note that in the opposite limit $\alpha \gg 0$ the switch to a straight line happens at a finite value $\alpha \approx 1.12$.

In these computations, we used $N_s = 2^{14}$, $h = 10^{-2}$.

4.2.2 Full Allen-Cahn/Cahn-Hilliard Model

Consider next the SPDE

$$\phi_t = \frac{1}{\alpha} P(\kappa \phi_{xx} + \phi - \phi^3) - \phi + \sqrt{\epsilon}\eta(x, t), \tag{56}$$

where P is an operator with zero spatial mean and $\eta(x, t)$ a spatio-temporal white-noise. This model is again of the form of two competing gradient flows with different mobilities:

$$\phi_t = -M_1 D_\phi V_1(\phi) - M_2 D_\phi V_2(\phi) + \sqrt{\epsilon} M_2^{1/2} \eta(x, t), \tag{57}$$

with

$$V_1(\phi) = \frac{1}{2}\kappa|\phi_x|^2 + \frac{1}{2}|\phi|^2 - \frac{1}{4}|\phi|^4, \quad M_1 = \frac{1}{\alpha}P \tag{58a}$$

$$V_2(\phi) = -\frac{1}{2}|\phi|^2, \qquad M_2 = \mathrm{Id}. \tag{58b}$$

For $P = -\partial_x^2$ the system is a mixture of a stochastic Allen-Cahn [2] and Cahn-Hilliard [7] equation. Here we will consider $P(\phi) = \phi - \fint \phi \, dx$, which is similar in most aspects discussed below but simpler to handle numerically. We are again interested in situations where α is small, and the time scales associated with V_1 and V_2 differ significantly. In this case it will turn out that transition pathways are very different from the heteroclinic orbits, in that the separatrix between the basins of attraction is approached far from the unstable critical point of the deterministic system. This behavior is reminiscent of the 2-dimensional example discussed above, but in an SPDE setting.

The fixed points of the deterministic ($\epsilon = 0$) dynamics of system (56) are the solutions of

$$P(\kappa\phi_{xx} + \phi - \phi^3) - \alpha\phi = 0. \tag{59}$$

The only constant solution of this equation is the trivial fixed point $\phi(x) = 0$, whose stability depends on α and κ. In the following, we choose $\alpha = 10^{-2}$ and $\kappa = 2 \cdot 10^{-2}$, in which case $\phi(x) = 0$ is unstable. The two stable fixed points obtained by solving (59) for these values of α and κ are depicted in Fig. 4 as ϕ_A and ϕ_B, with $\phi_A = -\phi_B$. An unstable fixed point configuration on the separatrix between ϕ_A and ϕ_B is also shown as ϕ_S.

For finite but small α, the deterministic part of (56) has a "slow manifold" made of the solutions of

$$P(\kappa\phi_{xx} + \phi - \phi^3) = 0. \tag{60}$$

On this manifold the motion is driven solely by changing the mean via the slow terms, $-\phi + \sqrt{\epsilon}\,\eta(x,t)$, on a time-scale of order $O(1)$ in α. After two integrations in space, (60) can be written as

$$\kappa\phi_{xx} + \phi - \phi^3 = \lambda, \tag{61}$$

where λ is a parameter. As a result the slow manifold can be described as one-parameter families of solutions parametrized by $\lambda \in \mathbb{R}$—in general there is more than one family because the manifold can have different branches corresponding to solutions of (59) with a different number of domain walls. The configuration labeled as ϕ_X in Fig. 4 shows the field at the intersection of one of these branches with the separatrix. Since the deterministic drift along the slow manifold is small compared to the $O(1/\alpha)$ drift induced by the Cahn-Hilliard term, one expects that the most probable transition pathway will use this manifold as channel to escape the basin of attraction of the stable fixed points ϕ_A or ϕ_B. This intuition is confirmed by the numerics, as shown next.

Figure 5 (left) shows the heteroclinic orbit connecting the two stable fixed points ϕ_A and ϕ_B to the unstable configuration ϕ_S. The mean is preserved along this orbit, which involves a nucleation event at the boundaries followed by domain wall motion through the domain. The unstable fixed point ϕ_s, denoted by S, which

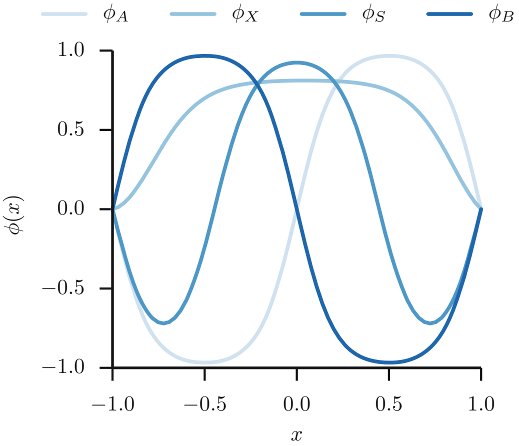


Fig. 4 The configurations A, B, S, X in space: ϕ_A and ϕ_B are the two stable fixed points, ϕ_S is the unstable fixed point on the separatrix in between. At point ϕ_X, the slow manifold intersects the separatrix

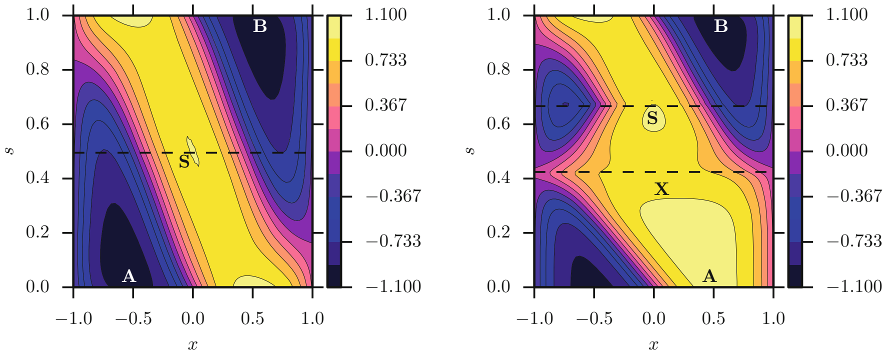


Fig. 5 Transition pathways between two stable fixed points of equation (56) in the limit $\epsilon \to 0$. Left: heteroclinic orbit, defining the deterministic relaxation dynamics from the unstable point S down to either A or B. Right: Minimizer of the geometric action, defining the most probable transition pathway from A to B, following the slow manifold up to X, where it starts to nearly deterministically travel close to the separatrix into S

also demarcates the position at which the separatrix is crossed, is the spatially symmetric configuration with a positive central region and two negative regions at the boundary. Locations A and B label the two stable fixed points ϕ_A and ϕ_B.

In contrast, Fig. 5 (right) shows the minimizer of the geometric action, which is the most probable transition path as $\epsilon \to 0$. It was computed via the algorithm outlined in Sect. 3.7, with $L = \frac{1}{\alpha}P\kappa\partial_x^2 - \mathrm{Id}$ and $R(u) = \frac{1}{\alpha}P(u - u^3)$. Starting at the fixed point A the minimizer takes a very different path than the heteroclinic orbit. It first moves the domain wall, at vanishing cost for $\alpha \to 0$, without nucleation. At the point X the motion changes, tracking closely the separatrix towards the unstable point S. From this point onward, $S \to B$, the transition path then follows the heteroclinic orbit, which is the deterministic relaxation path. In this respect, the SPDE model (56) resembles closely the 2-dimensional model (55).

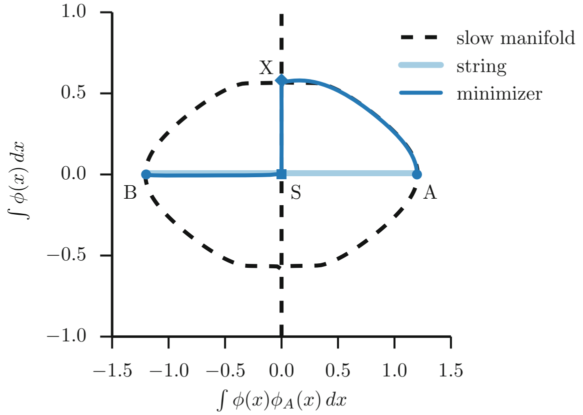


Fig. 6 Projection of the heteroclinic orbit and the minimizer of the action functional into a 2-dimensional plane. The x-direction is proportional to its component in the direction of the initial condition ϕ_A while the y-direction corresponds to its spatial mean. The stable fixed points are located at A and B, the unstable fixed point at S. The separatrix is the straight line $\int \phi(x)\phi_A(x)\,dx = 0$. The heteroclinic orbit (light) travels $A \to S \to B$ in a horizontal line with vanishing mean, while the minimizer (dark) travels first along the slow manifold (dashed) $A \to X$ and then tracks the separatrix from X to S

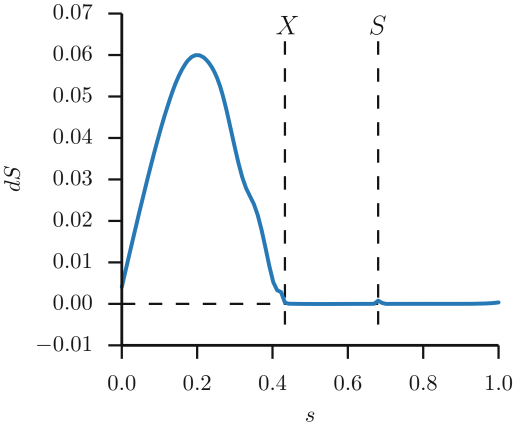
To further illustrate this resemblance, we choose to project the minimizer and the heteroclinic orbit onto two coordinates,

1. its mean $\int \phi(x)\,dx$, which resembles the direction $\phi_1 + \phi_2$ of the 2-dimensional model, and
2. its component in the direction of the initial (or final) state, $\int \phi(x)\phi_A(x)\,dx$, which corresponds to the direction $\phi_1 - \phi_2$ of the 2-dimensional model.

The transition path and the heteroclinic projected in these reduced coordinates are depicted in Fig. 6. Note that this figure is not a schematic, but the actual projection of the heteroclinic orbit and the minimizer of Fig. 5 according to (i) and (ii) above. The separatrix is the straight line $\int \phi(x)\phi_A(x)\,dx = 0$. The movement of the minimizer (dark) closely along the slow manifold (dashed), $A \to X$, and the separatrix, $X \to S$, (which is also part of the slow manifold) into S highlights its difference with the heteroclinic orbit (light). The configurations at the points A, B, S and X are depicted in Fig. 4, while Fig. 7 shows the action density dS along the transition path. Note that this quantity becomes close to zero already at X, because the minimizer follows closely the separatrix from X to S, and this motion is therefore quasi-deterministic.

The numerical parameters we used in these computations are $h = 10^{-1}$, $N_s = 100$, $N_x = 2^6$, where N_x denotes the number of spatial discretization points.

Fig. 7 Action along the minimizer. Note that the action is non-zero climbing up the slow-manifold, but diminishes to zero already at X when it approaches the separatrix, before it reaches S



4.3 Burgers-Huxley Model

As a second example involving an SPDE, we consider

$$u_t + \alpha u u_x - \kappa u_{xx} = f(u, x, t) + \sqrt{\epsilon}\eta(x, t). \tag{62}$$

where $\alpha > 0$ and $\kappa > 0$ are parameters, and we impose periodic boundary condition on $x \in [0, 1]$. Without the term $f(u, x, t)$, this is the stochastic Burgers equation which arises in a variety of fields, in particular in the context of compressible gas dynamics, traffic flow, and fluid dynamics. With the reaction term $f(u, x, t)$ added this equation is referred to as the (stochastic) Burgers-Huxley equation [42] , which has been used e.g. to describe the dynamics of neurons. The addition of a reaction term makes it possible to obtain multiple stable fixed points. As a particular case, we will consider (62) with

$$f(u, x, t) = -u(1 - u)(1 + u) \tag{63}$$

so that $u_+ = 1$ and $u_- = -1$ are the two stable fixed points of the deterministic dynamics. We are interested in the mechanism of the noise-induced transitions between these points.

When $\alpha = 0$, the system is in detailed balance and therefore the forward and backward reaction follow the same path. The potential associated with the reaction term (63) is symmetric under $u \to -u$, and both states are equally probable. In contrast, when $\alpha \neq 0$ it is not obvious *a priori* whether u_+ and u_- are equally probable, since the non-linearity breaks the spatial symmetry, leading to a steepening of negative gradients into shocks while flattening positive gradients. A computation of the minimizer of the geometric action in both directions, for $\kappa = 0.01$ and $\alpha = \frac{1}{4}$ reveals that indeed forward and backward reactions are equally probable, even though the transition paths do not coincide with the

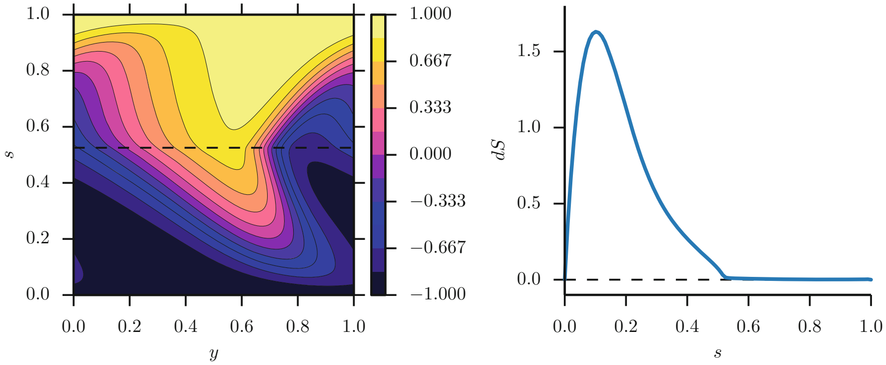


Fig. 8 Burgers-Huxley equation: Minimizer switching from $u_- = -1$ to $u_+ = 1$. Left: u-field. The saddle-point is marked with a dashed line. There is a noticeable kink in the dynamics switching from uphill ($s < s_{\text{saddle}}$) to downhill ($s > s_{\text{saddle}}$) dynamics. Right: Action density along the minimizer

heteroclinic orbits. The transition from u_- to u_+ is depicted in Fig. 8 (left). An intuitive explanation for the equal probability of u_+ and u_- is given by the fact that the backward reaction pathways is identical to the forward path under the transformation $u \to -u$, $x \to -x$. The action along this minimizer is depicted in Fig. 8 (right). The minimizer is computed via the algorithm lined out in Sect. 3.7, with $L = -\kappa\partial_x^2$ and $R(u) = \alpha u u_x + u(1-u)(1+u)$.

The numerical parameters were chosen as $N_s = 100$, $N_x = 2^8$, $h = 5 \cdot 10^{-3}$.

4.4 Noise-Induced Transitions Between Climate Regimes

Many climate systems exhibit metastability. Examples include the Kuroshio oceanic current off the coast of Japan, which can be in either a small or a large meander state and rarely switches between the two [9, 33], or the atmospheric mid-latitude circulation over the North-Atlantic, which makes rare transitions between a strongly zonal and a weakly zonal ("blocked") flow, characterized as "Grosswetterlagen" in [4]. In these and similar examples, the climate system stays trapped in the vicinity of the stable regimes most of the time. Random noise, originating either from physical stresses or from unresolved modes in truncated models, induces rare regime transitions, which can be captured by large deviation minimizers. The transition trajectory and their corresponding action allow to make statements about not only the relative probability of the different regimes and the transition rates, but also the exact transition pathway taken to switch between regimes.

We want to illustrate the feasibility of our numerical scheme for this particular field of application by investigating metastability in two simple climate models: A three-dimensional model for Grosswetterlagen proposed by Egger [18] and the six-dimensional Charney-DeVore model [10]. Due to their highly truncated nature, both models have very limited predictive power, but exemplify the phenomenon of metastability in climate patterns or regimes.

4.4.1 Metastable Climate Regimes in Egger's Model

Egger [18] introduces the following SDE system as a crude model to describe weather regimes in central Europe:

$$\begin{cases} da = kb(U - \beta/k^2)\,dt - \gamma a\,dt + \sqrt{\epsilon}dW_a, \\ db = -ka(U - \beta/k^2)\,dt + UH/k\,dt - \gamma b\,dt + \sqrt{\epsilon}dW_b, \\ dU = -bHk/2\,dt - \gamma(U - U_0)\,dt + \sqrt{\epsilon}dW_U. \end{cases} \tag{64}$$

When ϵ is small, these equation exhibit metastability between a "blocked state" and a "zonal state", shown in Fig. 9. We use our gMAM algorithm to compute the transition paths between these states. The system (64) falls into the category discussed in Sect. 3.4 and can be solved with the simplest variant of the algorithm. For $H = 12, \beta = 1.25, \gamma = 2, k = 2$ and $U_0 = 10.5$, the fixed points are approximately $(a, b, U) = (0.465, 1.65, 0.593)$ for the blocked, $(3.07, 0.392, 8.15)$ for the zonal and $(2.80, 1.35, 2.38)$ for the unstable fixed point (saddle). The minimizers of the action are show in Fig. 9(left) where they are compared to the heteroclinic orbits that connects the unstable critical points to the stable ones. The action density along the transition trajectories and the heteroclinic orbits is depicted in Fig. 9 (right).

The numerical parameters we used in these computations are $N_s = 2^8, h = 10^{-3}$.

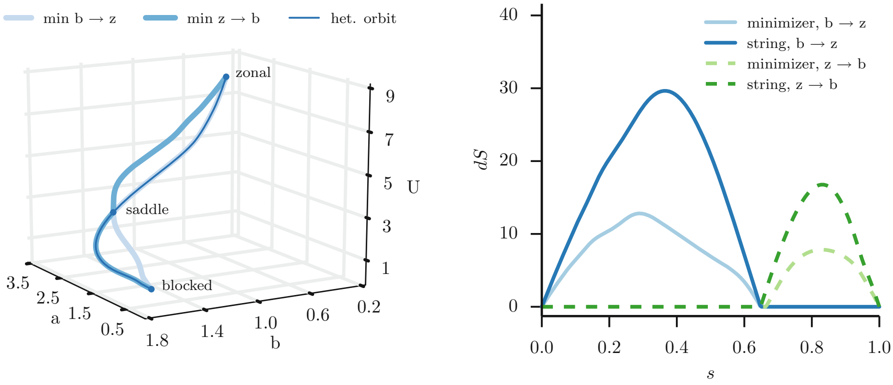


Fig. 9 Egger's model with $H = 12, \beta = 1.25, \gamma = 2, k = 2, U_0 = 10.5$ Left: Minimizers and deterministic relaxation paths. Right: Comparison of the action density

4.4.2 Metastable Climate Regimes in the Charney-DeVore Model

Egger's model retains no nonlinear interaction between different fluid modes, which is believed to be insufficient to explain the transitions between zonal and blocked states. A more sophisticated model, truncating the barotropic vorticity equation (BVE) with full nonlinear terms, was introduced by Charney and DeVore [10]. Their starting point is the two-dimensional BVE on the β-plane,

$$\frac{\partial}{\partial t}\omega = u \cdot \nabla\omega - C(\omega - \omega^*). \tag{65}$$

Here $\omega = \zeta + \beta y + \gamma h$ is the total vorticity, where γh is the topography in the β-plane, with $\beta = 2\Omega \cos(\theta)/R$ for planetary angular velocity Ω, radius R and latitude θ, and $\zeta = \Delta\psi$ is the relative vorticity for the stream-function ψ. The term $-C(\omega - \omega^*)$ accounts for Ekman damping with coefficient $C > 0$.

Charney-DeVore considered the vorticity equation (65) in the box $[0, 2\pi] \times [0, \pi b]$ with periodic boundary conditions in x-direction and no-slip boundary conditions in y-direction. They then projected this equation over 6 Fourier modes in total, using the following representation for the stream-function $\psi(x, y, t)$:

$$\psi(x, y, t) = \sum_{n,m} \psi_{nm}(t)\phi_{nm}(x, y), \tag{66}$$

where the sums run on $n \in \{-1, 0, 1\}$ and $m \in \{1, 2\}$ and

$$\phi_{0m}(y) = \sqrt{2}\cos(my/b), \qquad \phi_{nm}(x, y) = \sqrt{2}e^{inx}\sin(my/b). \tag{67}$$

Letting x_i, $i \in \{1, \dots, 6\}$ be defined as

$$\begin{aligned} x_1 &= \frac{1}{b}\psi_{01}, \quad & x_2 &= \frac{1}{\sqrt{2}b}\left(\psi_{11} + \psi_{-11}\right), \quad & x_3 &= \frac{i}{\sqrt{2}b}\left(\psi_{11} - \psi_{-11}\right), \\ x_4 &= \frac{1}{b}\psi_{02}, \quad & x_5 &= \frac{1}{\sqrt{2}b}\left(\psi_{12} + \psi_{-12}\right), \quad & x_6 &= \frac{i}{\sqrt{2}b}\left(\psi_{12} - \psi_{-12}\right), \end{aligned} \tag{68}$$

taking the following form for the topography

$$h(x, y) = \cos(x)\sin(y/b), \tag{69}$$

and choosing ω^* such that only two parameters x_1^* and x_4^* are free and the other are set zero, they arrived at the following six-dimensional model

$$
\begin{aligned}
dx_1 &= \left(\tilde{\gamma}_1 x_3 - C(x_1 - x_1^*)\right) dt + \sqrt{2\epsilon} dW_1, \\
dx_2 &= (-(\alpha_1 x_1 - \beta_1) x_3 - C x_2 - \delta_1 x_4 x_6)\, dt + \sqrt{2\epsilon} dW_2, \\
dx_3 &= ((\alpha_1 x_1 - \beta_1) x_2 - \gamma_1 x_1 - C x_3 + \delta_1 x_4 x_5)\, dt + \sqrt{2\epsilon} dW_3, \\
dx_4 &= \left(\tilde{\gamma}_2 x_6 - C(x_4 - x_4^*) + \eta(x_2 x_6 - x_3 x_5)\right) dt + \sqrt{2\epsilon} dW_4, \\
dx_5 &= (-(\alpha_2 x_1 - \beta_2) x_6 - C x_5 - \delta_2 x_3 x_4)\, dt + \sqrt{2\epsilon} dW_5, \\
dx_6 &= ((\alpha_2 x_1 - \beta_2) x_5 - \gamma_2 x_4 - C x_6 + \delta_2 x_2 x_4)\, dt + \sqrt{2\epsilon} dW_6,
\end{aligned}
\tag{70}
$$

where, for $m \in \{1, 2\}$,

$$
\begin{aligned}
\alpha_m &= \frac{8\sqrt{2}}{\pi} \frac{m^2}{4m^2 - 1} \frac{b^2 + m^2 - 1}{b^2 + m^2}, \\
\beta_m &= \frac{\beta b^2}{b^2 + m^2}, \\
\gamma_m &= \gamma \frac{\sqrt{2} b}{\pi} \frac{4m^3}{(4m^2 - 1)(b^2 + m^2)}, \\
\tilde{\gamma}_m &= \gamma \frac{\sqrt{2} b}{\pi} \frac{4m}{4m^2 - 1}, \\
\delta_m &= \frac{64\sqrt{2}}{15\pi} \frac{b^2 - m^2 + 1}{b^2 + m^2}, \\
\eta &= \frac{16\sqrt{2}}{5\pi}.
\end{aligned}
\tag{71}
$$

The original Charney-DeVore equation did not contain random forcing terms: here we added to each equations an independent white noise dW_i with amplitude $\sqrt{2\epsilon}$.

Choosing $b = \frac{1}{2}$, $C = \frac{1}{10}$, $\beta = \frac{5}{4}$, $\gamma = 1$, $x_1^* = \frac{9}{2}$, and $x_4^* = -\frac{9}{5}$, the 6-dimensional stochastic model above possesses two metastable states, shown in Fig. 10: a zonal state (left) and a blocked state (right). The transition paths from zonal to blocked and from blocked to zonal are different. They are shown in Figs. 11 and 12, respectively, and they were both calculated by minimizing the geometric action using our simplified gMAM algorithm. The actions along both paths are depicted in Fig. 13.

The numerical parameters in these computations were $N_s = 100$, $h = 10^{-3}$.

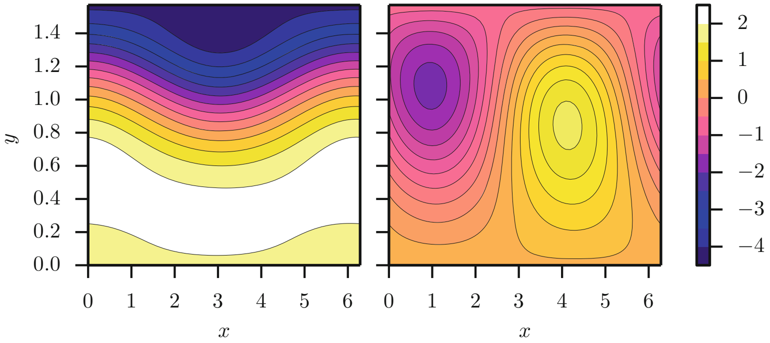


Fig. 10 Contours of the stream-function $\psi(x, y)$ of the two meta-stable configurations of the 6-dimensional CDV model. Left: Zonal state; Right: Blocked state

4.5 Generalized Voter/Ising Model

To analyze phase transitions in out-of-equilibrium systems, a Langevin equation was proposed in [1] that models critical phenomena with two absorbing states. This equation was constructed by requiring that it be symmetric under the transformation $\phi \to -\phi$ and have two absorbing states, arbitrarily chosen to be at ± 1. The presence of these absorbing states makes the noise multiplicative, with a scaling involving the square root of the distance to the absorbing boundaries, as suggested by the voter model [13, 15]. In order to account for Ising-like spontaneous symmetry breaking, the authors of [1] also added a bi-stable "potential"-term with $-V'(\phi) = (a\phi - b\phi^3)$ to the equation, which finally lead them to:

$$\phi_t = \big((1-\phi^2)(a\phi - b\phi^3) + D\phi_{xx}\big)\, dt + \sigma\sqrt{1-\phi^2}\,\eta(x,t). \tag{72}$$

In the absence of noise ($\epsilon = 0$) and for $a > 0$, the $\phi = 0$ state is locally unstable, but $b > 0$ ensures stable fixed points at $\phi = \pm\sqrt{a/b}$. In the limit $a/b \to 1$, these fixed points approach the absorbing boundaries, and we are interested in the noise induced transition between these states.

We stress that making mathematical sense of (72) is non-trivial (see the discussion in Sect. 3.7). In the present application, we are going to consider a finite truncation of this SPDE, where the question of spatial regularity disappears. Specifically, we transform (72) into a two-dimensional stochastic ODE model by discretizing the spatial direction via the standard 3-point Laplace stencil, and taking only $N_x = 2$ discretization points. This yields the stochastic ODE system

$$\begin{cases} d\phi_1 = \big((1-\phi_1^2)(a\phi_1 - b\phi_1^3) + D(\phi_1 - \phi_2)\big)\, dt + \sigma\sqrt{1-\phi_1^2}\, dW_x \\ d\phi_2 = \big((1-\phi_2^2)(a\phi_2 - b\phi_2^3) - D(\phi_1 - \phi_2)\big)\, dt + \sigma\sqrt{1-\phi_2^2}\, dW_y, \end{cases} \tag{73}$$

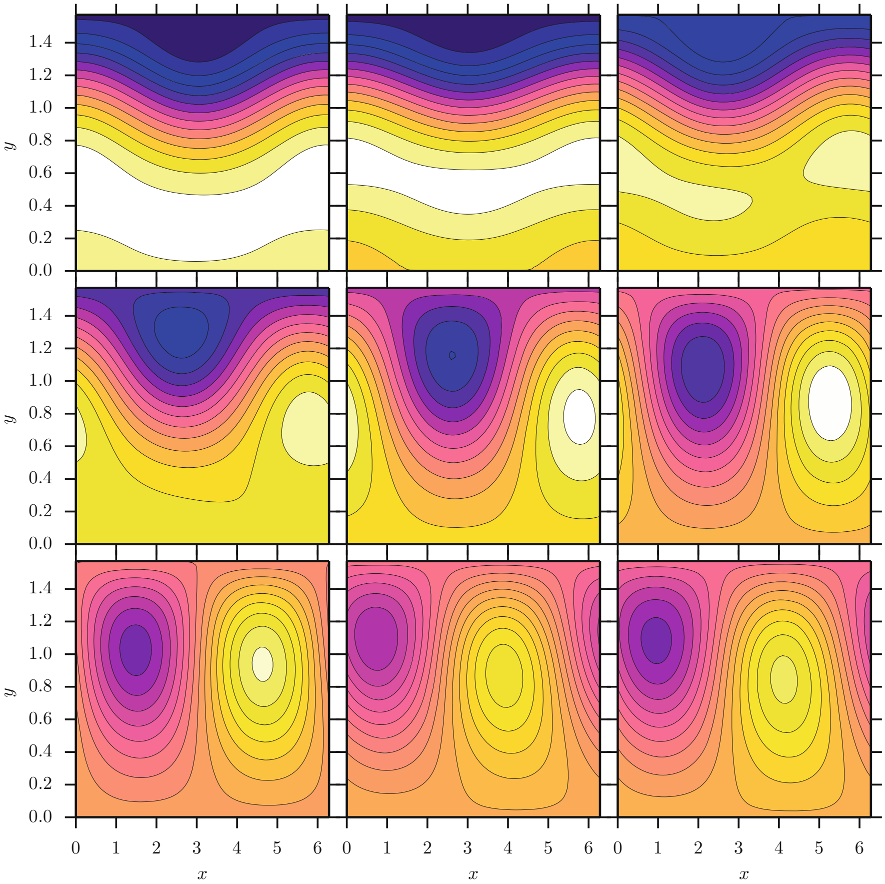


Fig. 11 Contours of the stream-function $\psi(x, y)$ along the transition trajectory from the zonal to the blocked meta-stable configuration for the CDV model. The arclength parameter increases in lexicographic order, with the top left plot being the initial state and the bottom right plot being the final state. The saddle point configuration is depicted in the center. The colormap is identical to Fig. 10

where the constant D couples the two degrees of freedom. This SDE poses an interesting test-case for our numerical scheme, since not only the noise is multiplicative, but also the computational domain must be restricted. The square defined by $1 = \max(|\phi_1|, |\phi_2|)$ marks the region in which the noise is defined (real), and the noise decreases towards zero as it approaches this absorbing barrier. Analog to the discussion in [1], the choice of the parameters (a, b) determines the dynamics, in particular if $a > 0, b > 0$ the model exhibits bi-stability: There is an unstable

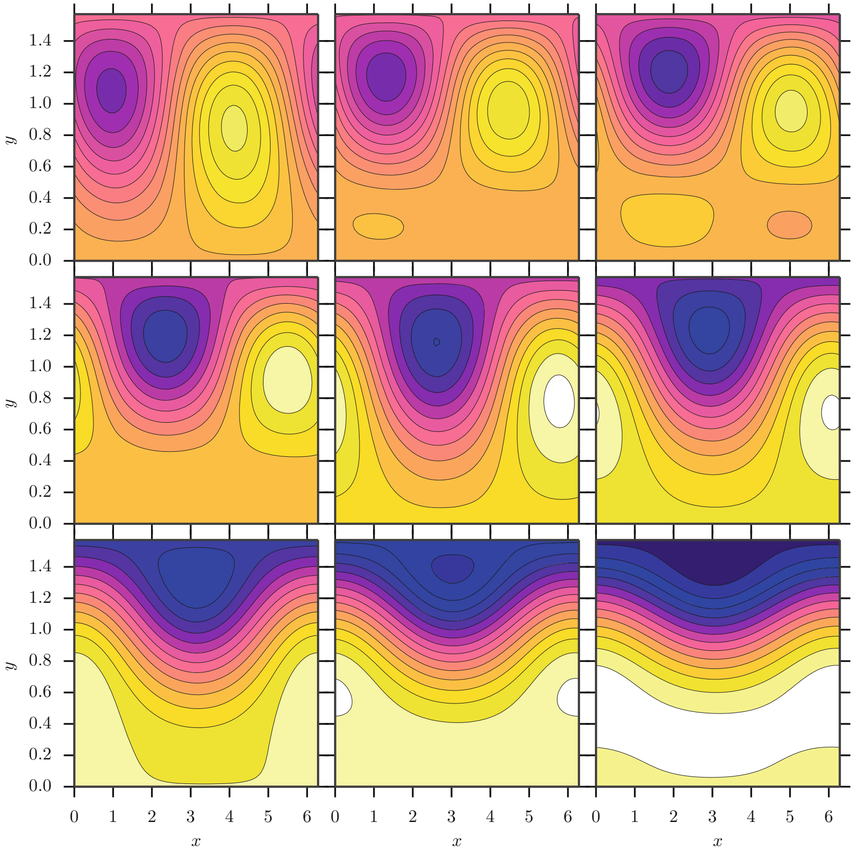


Fig. 12 Contours of the stream-function $\psi(x, y)$ along the transition trajectory from the blocked to the zonal meta-stable configuration for the CDV model. The arclength parameter increases in lexicographic order, with the top left plot being the initial state and the bottom right plot being the final state. The saddle point configuration is depicted in the center. The colormap is identical to Fig. 10

fixed point at $\phi = (0, 0)$ and stable fixed points at $\phi = \pm(\sqrt{a/b}, \sqrt{a/b})$. As long as $a < b$, these fixed points are inside the allowed region. For $a/b \to 1$ the two stable fixed points approach the absorbing boundary. Here, we take $b = 1, a = 1 - 10^{-4}, D = 0.4$, so that $\sqrt{a/b} \approx 0.99995$ is located close to the barrier at 1. The minimizer and corresponding action are shown in Fig. 14.

The numerical parameters were chosen as $N_s = 2^8, h = 10^{-3}$.

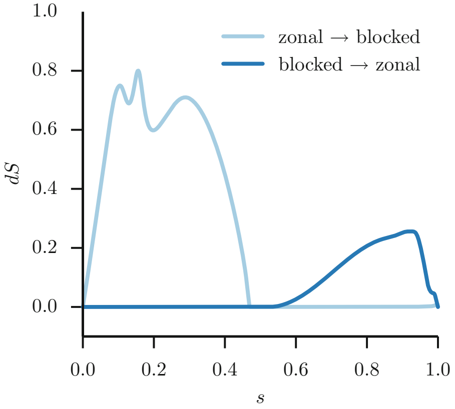


Fig. 13 Action density dS along the transition pathways from zonal to blocked (forward) and from blocked to zonal (backward). In both directions, after passing the saddle point, the action becomes zero since the motion is deterministic

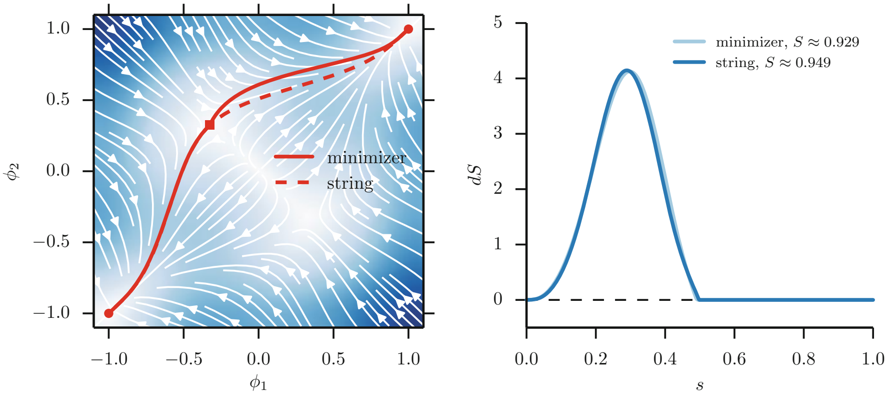


Fig. 14 Generalized voter/Ising model. Left: The arrows denote the direction of the deterministic flow, the shading its magnitude. The solid line depicts the minimizer, the dashed line the heteroclinic orbit. Markers are located at the fixed points (circle: stable; square: saddle). Right: Action density along the minimizers for the two trajectories, with normalized path parameter $s \in (0, 1)$

4.6 Bi-Stable Reaction-Diffusion Model

In the context of chemical reactions and birth-death processes, one considers networks of several reactants in a container of volume V which is considered well-stirred. As an example case, we consider the bi-stable chemical reaction network

$$A \underset{k_1}{\overset{k_0}{\rightleftharpoons}} X, \qquad 2X + B \underset{k_3}{\overset{k_2}{\rightleftharpoons}} 3X$$

with rates $k_i > 0$, and where the concentrations of A and B are held constant. This system was introduced in [32] as a prototypical model for a bi-stable reaction network. Its dynamics can be modeled as a Markov jump process (MJP) with generator

$$(L^R f)(n) = A_+(n)\,(f(n+1) - f(n)) + A_-(n)\,(f(n-1) - f(n)) \tag{74}$$

with the propensity functions

$$\begin{cases} A_+(n) = k_0 V + (k_2/V)n(n-1) \\ A_-(n) = k_1 n + (k_3/V^2)n(n-1)(n-2). \end{cases} \tag{75}$$

The model above satisfies a large deviation principle in the following scaling limit: Denote by $c = n/V$ the concentration of X, and normalize it by a typical concentration, $\rho = c/c_0$. Now, in the limit of a large number of particles per cell $\Omega = c_0 V$ and simultaneously rescaling time by $1/\Omega$, we obtain

$$(L^R_\epsilon f)(\rho) = \frac{1}{\epsilon}\Big(a_+(\rho)\,(f(\rho+\epsilon) - f(\rho)) + a_-(\rho)\,(f(\rho-\epsilon) - f(\rho))\Big), \tag{76}$$

where $\epsilon = 1/\Omega$ is a small parameter. Here, we defined $k_i = \lambda_i (c_0)^{1-i}$, and

$$\begin{cases} a_+(\rho) & = \lambda_0 + \lambda_2\rho^2 \\ a_-(\rho) & = \lambda_1\rho + \lambda_3\rho^3. \end{cases} \tag{77}$$

The large deviation principle for (76) can be formally obtained via WKB analysis, that is, by setting $f(\rho) = e^{\epsilon^{-1}G(\rho)}$ in (76) and expanding in ϵ [14]. To leading order in ϵ, this gives an Hamilton-Jacobi operator associated with an Hamiltonian that is also the one rigorously derived in LDT [35]. It reads

$$H(\rho, \vartheta) = a_+(\rho)(e^{\vartheta} - 1) + a_-(\rho)(e^{-\vartheta} - 1). \tag{78}$$

This is an example of a system whose Hamiltonian is not quadratic in the conjugate momentum ϑ. Therefore the computation of ϑ_* by (26) can not be performed explicitly in general. For parameters $\lambda_0 = 0.8, \lambda_1 = 2.9, \lambda_2 = 3.1, \lambda_3 = 1$, the system has two stable fixed points $\rho_\pm$ and a saddle ρ_s at $\rho_+ = \frac{8}{5}, \rho_- = \frac{1}{2}, \rho_s = 1$.

Since transitions in 1D are fairly trivial, we want to consider the case of N neighboring reaction compartments, each well-stirred, but with random jumps possible between neighboring compartments. This situation was analyzed in [38] via direct sampling, but we are interested in the computation of the transition trajectory. Denote by ρ_i the concentration in the i-th compartment and refer to the vector ρ as the complete state, $\rho = \sum_{i=0}^N \rho_i \hat{e}_i$. In this case, we obtain a diffusive part of the generator, L^D, coupling neighboring compartments. For a diffusivity D, it is

$$(L^D f)(\rho) = \frac{D}{\epsilon}\sum_{i=1}^{N} \rho_i \left(f(\rho - \epsilon\hat{e}_i + \epsilon\hat{e}_{i-1}) + f(\rho - \epsilon\hat{e}_i + \epsilon\hat{e}_{i+1}) - 2f(\rho)\right). \tag{79}$$

The process associated with this generator also admits a large deviation principle with Hamiltonian

$$H^D(\rho,\vartheta) = D\sum_{i=1}^{N}\rho_i\left(e^{\vartheta_{i-1}-\vartheta_i} + e^{\vartheta_{i+1}-\vartheta_i} - 2\right). \tag{80}$$

Therefore, the full Hamiltonian becomes $H(\rho,\vartheta) = H^D(\rho,\vartheta) + \sum_{i=1}^{N} H^R(\rho_i,\vartheta_i)$, where $H^R(\rho_i,\vartheta_i)$ is the reactive Hamiltonian in (78), which is summed up over all the compartments.

We used our new gMAM algorithm to minimize the geometric action and compute the transition paths between the stable fixed points for the simplest non-trivial case of $N = 2$ compartments. Shown in Fig. 15 are the forward and backward trajectories. Note that the backward transition $((\rho_+,\rho_+) \to (\rho_-,\rho_-))$ takes a special form: It climbs against the deterministic dynamics up to the maximum, then relaxes along the separatrix down to the saddle. Additionally, we compare these trajectories with the heteroclinic orbit obtained by the string method. The action along these trajectories is depicted in Fig. 16. Note how for the backward minimizer the action is zero already before it hits the saddle, as the movement from the maximum to the saddle happens deterministically.

The numerical parameters were chosen as $N_s = 2^9$, $h = 10$.

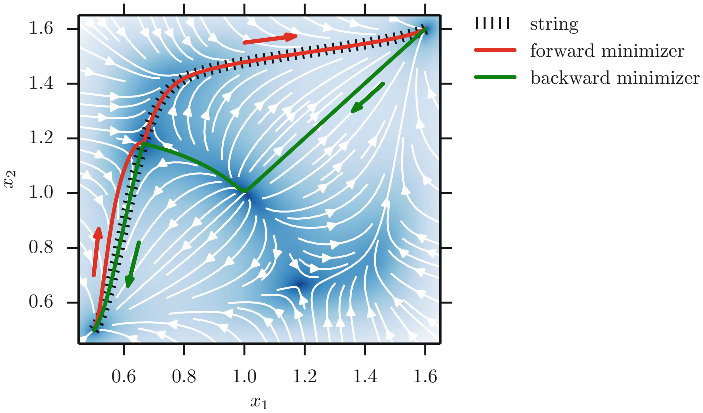


Fig. 15 Bi-Stable reaction-diffusion model with $N = 2$ reaction cells. Show are the forward (red) and backward (green) transitions between the two stable fixed points, in comparison to the heteroclinic orbit (dashed). The flow-lines depict the deterministic dynamics, their magnitude is indicated by the background shading

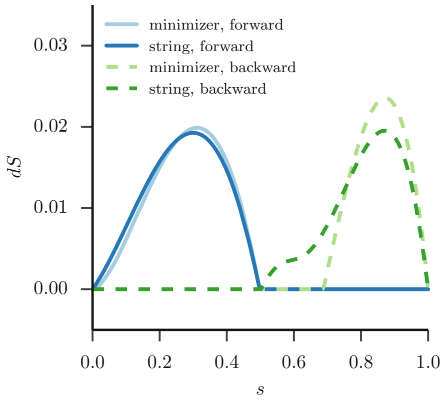


Fig. 16 Action densities for the bi-stable reaction-diffusion model. Depicted are the actions corresponding to the forward (solid) and backward (dashed) minimizer (dark) and heteroclinic (light) orbit

4.7 Slow-Fast Systems

In contrast to a large deviation principle arising in the limit of small noise or large number of particles, a different class of Hamiltonians arises for systems with a slow variable X evolving on a timescale $O(1)$ and a fast variable Y on a time scale $O(\alpha)$:

$$\dot{X} = f(X, Y) \tag{81a}$$

$$dY = \frac{1}{\alpha} b(X, Y)dt + \frac{1}{\sqrt{\alpha}} \sigma(X, Y)dW. \tag{81b}$$

Examples of systems with large timescale separation $\alpha \ll 1$ are ubiquitous in nature, and usually one is interested mostly in the long-time behavior of the slow variables. In particular, we are concerned with situations where the slow dynamics exhibits metastability. We want to use our algorithm to compute transition pathways in this setup for the limit of infinite time scale separation.

In the limit as $\alpha \to 0$, the fast variables reach statistical equilibrium before any motion of the slow variables, and these slow variables only experience the average effect of the fast ones. This behavior can be captured by the following deterministic limiting equation which is akin to a law of large numbers (LLN) in the present context and reads

$$\dot{\bar{X}} = F(\bar{X}) \quad \text{where} \quad F(x) = \lim_{T\to\infty} \frac{1}{T} \int_0^T f(x, Y_x(\tau))\, d\tau. \tag{82}$$

Here $Y_x(t)$ is the solution of (81b) for $X(t) = x$ fixed [3, 6, 19, 30]. For small but finite α, the slow variables also experience fluctuations through the fast variables. In particular, the statistics of $\xi = (X - \bar{X})/\sqrt{\alpha}$ on $O(1)$ time scales can be described by a central limit theorem (CLT) as small Gaussian noise on top of the slow mean $\bar{X}$. The CLT scaling, however, is inappropriate to describe the fluctuations of the slow variables that are induced by the effect of the fast variables on longer time scales and may, for example, lead to transitions between stable fixed points of the limiting equation in (82). In particular, the naive procedure of constructing an SDE out of the LLN and CLT to then compute its LDT fails. Instead, the transitions in the limit of $\alpha \to 0$ are captured by an LDP with the Hamiltonian

$$H(x, \vartheta) = \lim_{T\to\infty} \frac{1}{T} \log \mathbb{E} \exp\left(\vartheta \int_0^T f(x, Y_x(t))\, dt\right). \tag{83}$$

Except for the special case $f(x, y) = r(x) + s(y)y$ (linear dependence on the fast variable), the Hamiltonian (83) is non-quadratic in θ. As a consequence no S(P)DE with Gaussian noise exists for the slow variable which has an LDP to describe the transitions correctly.

The implicit nature of the Hamiltonian (83), in particular containing an expectation, complicates numerical procedures to compute its associated minimizers. Yet, in the non-trivial case of a quadratic dependence of the slow variable on the fast ones, for example,

$$\begin{cases} \dot{X} = Y^2 - \beta X \\ dY = -\dfrac{1}{\alpha}\gamma(X)Y\, dt + \dfrac{\sigma}{\sqrt{\alpha}} dW, \end{cases} \tag{84}$$

one indeed does obtain an explicit formula for the Hamiltonian (83) (as derived in [6])

$$h(x, \vartheta) = -\beta x \vartheta + \tfrac{1}{2}\left(\gamma(x) - \sqrt{\gamma^2(x) - 2\sigma^2\vartheta}\right). \tag{85}$$

This example is interesting for our purpose not only because the Hamiltonian is non-quadratic, but furthermore because of the existence of a forbidden region $\vartheta > \gamma^2/(2\sigma)$ where the Hamiltonian is not defined.

Additionally increasing the number of degrees of freedom by combining two independent multi-stable slow-fast systems and coupling them by a spring with spring constant D, the full system reads

$$\begin{cases} \dot{X}_1 = Y_1^2 - \beta_1 X_1 - D(X_1 - X_2) \\ \dot{X}_2 = Y_2^2 - \beta_2 X_2 - D(X_2 - X_1) \\ dY_1 = -\dfrac{1}{\alpha}\gamma(X_1)Y_1 dt + \dfrac{\sigma}{\sqrt{\alpha}} dW_1 \\ dY_2 = -\dfrac{1}{\alpha}\gamma(X_2)Y_2 dt + \dfrac{\sigma}{\sqrt{\alpha}} dW_2. \end{cases} \tag{86}$$

The Hamiltonian for the LDT for this system is

$$H(x_1, x_2, \vartheta_1, \vartheta_2) = h(x_1, \vartheta_1) + h(x_2, \vartheta_2) + \langle -\nabla U(x_1, x_2), \vartheta \rangle, \tag{87}$$

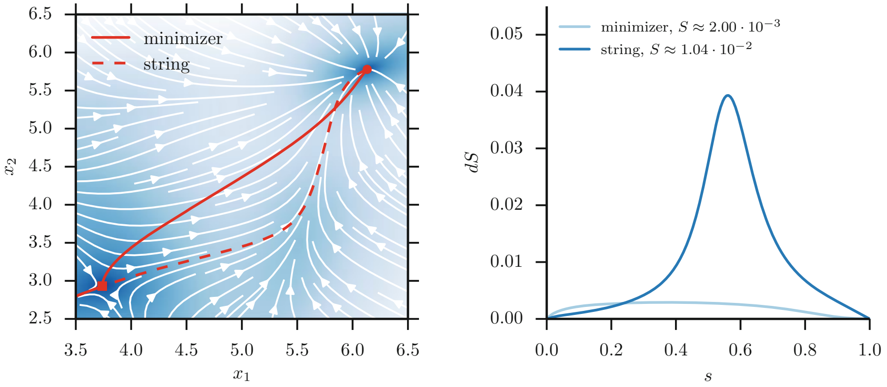


Fig. 17 Coupled slow-fast system ODE model for $D = 1.0$. Left: The arrows denote the direction of the deterministic flow, the shading its magnitude. The solid line depicts the minimizer, the dashed line the relaxation paths from the saddle. Markers are located at the fixed points (circle: stable; square: saddle). Right: Action density along the minimizers for the two trajectories up to the saddle, with normalized path parameter $s \in (0, 1)$

for $U(x, y) = \frac{1}{2}D(x - y)^2$ and $h(x, \vartheta)$ defined as in equation (85). The choice $\gamma(X) = (X-5)^2 + 1$ ensures two stable fixed points. The deterministic dynamics of this system (i.e. the evolution of the averaged slow variables) are depicted as white arrows in Fig. 17 (left). To stress the important portion of the transition trajectory, the plot is focused only on the initial state up to the saddle. Compared are the minimizer and the heteroclinic orbits connecting the stable fixed points to the saddle point. The corresponding actions are shown in Fig. 17 (right). The specific choice of model parameters for this computation is $\beta_1 = 0.6, \beta_2 = 0.3, D = 1.0$ and $\sigma^2 = 10$.

The numerical parameters were chosen as $N_s = 2^{10}$, $h = 10^{-2}$.

5 Concluding Remarks

We have discussed numerical schemes to compute minimizers of large deviation action functionals, which are based on the geometric minimum action method. The basis of these schemes is the minimization of a geometric action on the space of arc-length parametrized curves, which makes it possible to perform the double minimization over transition time T and action S_T that is required to compute the LDT quasipotential. In particular, transitions between metastable fixed points of a system, which generally involve $T \to \infty$ and which are not tractable with non-geometric minimum action methods can be naturally analyzed in this setup.

A simplified gMAM algorithm was proposed here which is based on a particular formulation of the geometric action leading to a mixed optimization problem. This new formulation of the gMAM algorithm is easier to implement than the original method: In its simplest form, only first order derivatives of the Hamiltonian

$H(\varphi, \vartheta)$ are needed. The algorithm is applicable to a large class of systems, and does not rely on an explicit formula of the large deviation rate function—only the Hamiltonian of the theory is needed. We derived specific reductions that are possible in regularly occurring special cases, such as SDEs with additive or multiplicative noise. Furthermore, we discussed optimizations for SPDEs with additive noise and commented on how to improve numerical efficiency.

The performances of the new gMAM algorithm were illustrated in a series of applications arising from different fields and involving different types of models, like S(P)DEs with additive and multiplicative Gaussian noises, Markov jump processes, or slow-fast systems.

Acknowledgements We would like to thank M. Cates, A. Donev, and F. Bouchet for helpful discussions. This research was supported in part by the NSF grants DMR-1207432 (T. Grafke), DMS-1108780 and DMS-1522737 (T. Schäfer), and DMS-1522767 (E. Vanden-Eijnden).

References

1. Al Hammal O, Chaté H, Dornic I, Muñoz MA (2005) Langevin Description of Critical Phenomena with Two Symmetric Absorbing States. Physical Review Letters 94(23):230,601
2. Allen S, Cahn J (1972) Ground state structures in ordered binary alloys with second neighbor interactions. Acta Metallurgica 20(3):423–433
3. Bakhtin VI (2003) Cramér's asymptotics in systems with fast and slow motions. Stochastics and Stochastic Reports 75(5):319–341
4. Baur F (1951) Extended range weather forecasting. Compendium of meteorology pp 814–833
5. Beylkin G, Keiser JM, Vozovoi L (1998) A New Class of Time Discretization Schemes for the Solution of Nonlinear PDEs. Journal of Computational Physics 147(2):362–387
6. Bouchet F, Grafke T, Tangarife T, Vanden-Eijnden E (2016) Large Deviations in Fast–Slow Systems. Journal of Statistical Physics pp 1–20
7. Cahn JW, Hilliard JE (1958) Free Energy of a Nonuniform System. I. Interfacial Free Energy. The Journal of Chemical Physics 28(2):258–267
8. Cates ME, Marenduzzo D, Pagonabarraga I, Tailleur J (2010) Arrested phase separation in reproducing bacteria creates a generic route to pattern formation. Proceedings of the National Academy of Sciences 107(26):11,715–11,720
9. Chao SY (1984) Bimodality of the Kuroshio. Journal of Physical Oceanography 14(1):92–103
10. Charney JG, DeVore JG (1979) Multiple Flow Equilibria in the Atmosphere and Blocking. Journal of the Atmospheric Sciences 36(7):1205–1216
11. Chernykh AI, Stepanov MG (2001) Large negative velocity gradients in Burgers turbulence. Physical Review E 64(2):026,306
12. Cox S, Matthews P (2002) Exponential Time Differencing for Stiff Systems. Journal of Computational Physics 176(2):430–455
13. Dickman R, Tretyakov AY (1995) Hyperscaling in the Domany-Kinzel cellular automaton. Physical Review E 52(3):3218–3220
14. Doering CR, Sargsyan KV, Sander LM, Vanden-Eijnden E (2007) Asymptotics of rare events in birth–death processes bypassing the exact solutions. Journal of Physics: Condensed Matter 19(6):065,145
15. Dornic I, Chaté H, Chave J, Hinrichsen H (2001) Critical Coarsening without Surface Tension: The Universality Class of the Voter Model. Physical Review Letters 87(4):045,701

16. E W, Ren W, Vanden-Eijnden E (2002) String method for the study of rare events. Physical Review B 66(5):052,301
17. E W, Ren W, Vanden-Eijnden E (2004) Minimum action method for the study of rare events. Communications on Pure and Applied Mathematics 57(5):637–656
18. Egger J (1981) Stochastically Driven Large-scale Circulations with Multiple Equilibria. Journal of the Atmospheric Sciences 38(12):2606–2618
19. Freidlin MI (1978) The averaging principle and theorems on large deviations. Russian Mathematical Surveys 33(5):117–176
20. Freidlin MI, Wentzell AD (2012) Random perturbations of dynamical systems, vol 260. Springer
21. Grafke T, Grauer R, Schäfer T, Vanden-Eijnden E (2014) Arclength Parametrized Hamilton's Equations for the Calculation of Instantons. Multiscale Modeling & Simulation 12(2):566–580
22. Grafke T, Grauer R, Schindel S (2015) Efficient Computation of Instantons for Multi-Dimensional Turbulent Flows with Large Scale Forcing. Communications in Computational Physics 18(03):577–592
23. Hairer M (2014) A theory of regularity structures. Inventiones mathematicae 198(2):269–504
24. Hairer M, Weber H (2015) Large deviations for white-noise driven, nonlinear stochastic PDEs in two and three dimensions. Annales de la facultédes sciences de Toulouse Mathématiques 24(1):55–92
25. Heymann M, Vanden-Eijnden E (2008) The geometric minimum action method: A least action principle on the space of curves. Communications on Pure and Applied Mathematics 61(8):1052–1117
26. Heymann M, Vanden-Eijnden E (2008) Pathways of maximum likelihood for rare events in nonequilibrium systems: application to nucleation in the presence of shear. Phys Rev Lett 100(14):140,601
27. Kassam A, Trefethen L (2005) Fourth-Order Time-Stepping for Stiff PDEs. SIAM Journal on Scientific Computing 26(4):1214–1233
28. Kifer Y (1992) Averaging in dynamical systems and large deviations. Inventiones Mathematicae 110(1):337–370
29. Kifer Y (2004) Averaging principle for fully coupled dynamical systems and large deviations. Ergodic Theory and Dynamical Systems 24(03):847–871
30. Kifer Y (2009) Large deviations and adiabatic transitions for dynamical systems and Markov processes in fully coupled averaging. Memoirs of the American Mathematical Society 201(944):0–0
31. Maier RS, Stein DL (1996) A scaling theory of bifurcations in the symmetric weak-noise escape problem. Journal of Statistical Physics 83(3–4):291–357
32. Schlögl F (1972) Chemical reaction models for non-equilibrium phase transitions. Zeitschrift für Physik 253(2):147–161
33. Schmeits MJ, Dijkstra HA (2001) Bimodal behavior of the Kuroshio and the Gulf Stream. J Phys Ocean 31:3435
34. Shapiro JA (1995) The significances of bacterial colony patterns. BioEssays 17(7):597–607
35. Shwartz A, Weiss A (1995) Large Deviations For Performance Analysis: QUEUES, Communication and Computing. CRC Press
36. Strang G (1968) On the Construction and Comparison of Difference Schemes. SIAM Journal on Numerical Analysis 5(3):506–517
37. Tailleur J, Cates ME (2008) Statistical Mechanics of Interacting Run-and-Tumble Bacteria. Physical Review Letters 100(21):218,103
38. Tănase-Nicola S, Lubensky DK (2012) Exchange of stability as a function of system size in a nonequilibrium system. Physical Review E 86(4):040,103
39. Vanden-Eijnden E, Heymann M (2008) The geometric minimum action method for computing minimum energy paths. Jour Chem Phys 128:061,103
40. Veretennikov AY (2000) On large deviations for SDEs with small diffusion and averaging. Stochastic Processes and their Applications 89(1):69–79

41. Wan X (2011) An adaptive high-order minimum action method. Journal of Computational Physics 230(24):8669–8682
42. Wang XY, Zhu ZS, Lu YK (1990) Solitary wave solutions of the generalised Burgers-Huxley equation. Journal of Physics A: Mathematical and General 23(3):271
43. Yoshida H (1990) Construction of higher order symplectic integrators. Physics Letters A 150(5–7):262–268
44. Zhou X, Ren W, E W (2008) Adaptive minimum action method for the study of rare events. J Chem Phys 128:104,111

Part III
Nonlinear Waves, Hyperbolic Problems, and their Applications

Long Time Dynamics and Coherent States in Nonlinear Wave Equations

E. Kirr

Abstract We discuss recent progress in finding all coherent states supported by nonlinear wave equations, their stability and the long time behavior of nearby solutions.

1 Introduction

Since the discovery of wave equations two and a half centuries ago, many scientist and mathematicians have tried to understand their most striking feature, the coherent structures. An exact definition of coherent structures will be given in Sect. 2 but, formally, they are solutions which propagate without changing shape (or with periodic change in shape). In the linear case, they are the eigenfunctions of the corresponding wave operator and, via spectral decomposition, the actual dynamics becomes a superposition of these coherent states (eigenfunctions) and a projection onto the remaining (continuous) spectrum. Once the latter had been analyzed via dispersive estimates or scattering wave operators, it became clear that the following asymptotic completeness conjecture is true for the linear case.

Asymptotic Completeness Conjecture: Any initial data evolves towards a superposition of coherent structures plus a part that radiates (scatters) to infinity.

In the nonlinear case, some coherent structures are known either as minimizer or mountain pass type critical points of the energy subject to certain constraints, see Sect. 3.1. Other coherent states (sometimes the same ones) can be found via bifurcations from already known solutions such as the trivial one, see Sect. 3.2. In many situations the coherent states undergo symmetry breaking phenomena, see for example [3, 20, 23, 27], which are very important in practical applications. But none of these results, nor the mathematical methods they rely on can claim that they can actually identify all coherent states supported by a given nonlinear wave equation. Consequently, the asymptotic completeness conjecture is wide open with

E. Kirr (✉)
Department of Mathematics, University of Illinois, Urbana-Champaign, IL 61801, USA
e-mail: ekirr@illinois.edu

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_3

two notable exceptions: the case of completely integrable systems (such as the cubic Schrödinger equation in one dimension) where a scattering transform renders the problem linear, or the case of weakly nonlinear regimes (i.e. small initial data) where at most two (small) coherent states are present and one of them is selected after a long transitional time, see [51–54].

This paper aims to present recent results and new ideas for finding *all coherent states* (solitary waves, breathers, kinks, vortices, etc) supported by a given nonlinear wave equations. As shown in Sect. 3.2, coherent states can be viewed as zeroes of a map between Banach spaces. Then, the global bifurcation theory, see [8, 45], allows us to organize them in smooth manifolds which either form loops or can be extended to the boundary of the domain inside which the linearization of the map is Fredholm. The new results and ideas concern finding all the limit points of such manifolds on the boundary of the Fredholm domain. Hence, from these limit points, the manifolds of coherent states *can be found and traced* both theoretically and numerically inside the Fredholm domain. Moreover, by comparing the spectrum of the linearized operator at the two "end points" of these manifolds we can deduce whether eigenvalues cross zero which is equivalent with the existence of bifurcations in this Fredholm region. Once discovered, the bifurcations can be studied using local bifurcation techniques to determine the new branches (manifolds) emerging from them and their dynamical stability. Global bifurcation theory now implies that the new branches have "end points" on the boundary of the Fredholm domain. Consequently, we are able to find the bifurcations along the new branches and the process iterates until all these branches are discovered and matched with all possible limit points.

Results and open problems regarding the orbital stability of the manifolds of coherent states are discussed in Sect. 3.3 while Sect. 4 discusses their asymptotic stability. The latter brings us closer to a resolution of the *Asymptotic Completeness Conjecture* but, unfortunately, it only describes the dynamics in a neighborhood of the coherent state manifolds. The last section is reserved for concluding remarks.

2 General Hamiltonian Formulation

Most models related to wave propagation, in particular the Schrödinger, Hartree, Dirac, Klein-Gordon, Korteweg-de-Vries and the classical wave equation, can be cast in the following general framework, see [18]. The evolution of the quantity of interest u is given by:

$$\frac{du}{dt} = JD_u\mathscr{E}(u), \qquad t \in \mathbb{R}, \tag{1}$$

where X is a real Hilbert space, the energy $\mathscr{E} : X \mapsto \mathbb{R}$ is a C^2 functional, D_u denotes the Frechet derivative with respect to the variable u, and $J : D(J) \subseteq X^* \mapsto X$ is a skew-symmetric operator $J^* = -J$, defined on a dense subset of the dual of X. Note that even though X is a Hilbert space, its dual is not necessarily identified with X. The reason is twofold: the applications have a physically important larger Hilbert space Y for which $X \hookrightarrow Y = Y^* \hookrightarrow X^*$, where all the embeddings are dense, and the mathematical analysis of the operators appearing in the applications rely on the larger Hilbert space Y.

Besides being time independent, the energy is in general invariant under additional groups of symmetries. These symmetries can be modeled by one or more (strongly) continuous groups of unitary operators. In what follows we will focus on one such group of symmetries $T(s) : X \mapsto X,\ s \in \mathbb{R}$, which leave the energy functional invariant and commutes with the J operator:

$$\mathscr{E}(T(s)u) = \mathscr{E}(u), \quad T(s)J = JT^*(-s), \quad \text{for all } s \in \mathbb{R},\ \text{and } u \in X.$$

By Noether's Theorem, see for example [5], the Hamiltonian dynamics has, besides energy, a second conserved quantity:

$$Q(u) = \frac{1}{2}\langle Bu, u\rangle, \quad u \in X \tag{2}$$

provided that there exists a bounded, self-adjoint, linear operator $B : X \mapsto X^*$ such that JB extends the infinitesimal generator of the continuous group: $T'(0)$.

The coherent states are solutions of (1) of the type:

$$u(t) = T(\omega t)\phi_\omega$$

where $\phi_\omega \in D(T'(0)) \subseteq X,\ \omega \in \mathbb{R}$ are fixed. In applications ϕ_ω usually gives the shape of u, so these are solutions which do not change their shape as they propagate. By plugging in (1) one finds that:

$$JD_\phi\mathscr{E}(\phi_\omega) = J\omega D_\phi Q(\phi_\omega).$$

Consequently the solutions in $D(T'(0))$ of the stationary equation:

$$D_\phi\mathscr{E}(\phi) = \omega D_\phi Q(\phi) \tag{3}$$

generate coherent states and they are the only possible coherent structures if J is one-to-one.

Coherent states are orbitally stable if any solution of (1) starting close to the orbit $T(s)\phi_\omega,\ s \in \mathbb{R}$, of the coherent state, remains close to it at all times. More precisely for any ε there exists a δ such that

$$\inf_{s\in\mathbb{R}} \|u(0) - T(s)\phi_\omega\| < \delta \quad \text{implies} \quad \sup_{t\in\mathbb{R}} \inf_{s\in\mathbb{R}} \|u(t) - T(s)\phi_\omega\| < \varepsilon.$$

Asymptotic stability means certain convergence of the solutions to the orbit of a coherent structure and usually takes the form: there exists a Banach space Z, $X \hookrightarrow Z$ densely, and $\delta > 0$ such that if $\inf_{s\in\mathbb{R}} \|u(0) - T(s)\phi_\omega\|_{X\bigcap Z^*} < \delta$ then there exists a coherent structure $T(\omega_+ t)\phi_{\omega_+}$ (close to $T(\omega t)\phi_\omega$) with the property:

$$\lim_{t\to\infty} \inf_{s\in\mathbb{R}} \|u(t) - T(s)\phi_{\omega_+}\|_Z = 0.$$

For example, in the case of the nonlinear Schrödinger equation (NLS) we have $X = H^1(\mathbb{R}^n)$, the Sobolev space of complex valued functions but with the real Hilbert space structure, $X^* = H^{-1},\ Jv = -iv,\ T(s)u = e^{-is}u$,

$$\mathscr{E}(u) = \frac{1}{2}\int_{\mathbb{R}^n} |\nabla u(x)|^2 dx + \frac{1}{2}\int_{\mathbb{R}^n} V(x)|u(x)|^2 dx + \frac{\gamma}{p+2}\int_{\mathbb{R}^n} |u(x)|^{p+2} dx, \quad (4)$$

$$Q(u) = \frac{1}{2}\int_{\mathbb{R}^n} |u(x)|^2 dx, \quad (5)$$

hence the evolution equation (1) becomes:

$$i\frac{\partial u}{\partial t} = (-\Delta + V(x))u(t,x) + \gamma|u|^p u(t,x), \quad t \in \mathbb{R},\ x \in \mathbb{R}^n, \quad (6)$$

while the coherent structures are solutions of the form $u(t,x) = e^{iEt}\phi_E(x),\ E = -\omega \in \mathbb{R},\ \phi_E \in H^1(\mathbb{R}^n)$, and satisfy the equation:

$$F(\phi_E, E) = (-\Delta + V + E)\phi_E + \gamma|\phi_E|^p\phi_E = 0. \quad (7)$$

Here $V : \mathbb{R}^n \mapsto \mathbb{R}$ is called the potential, $\gamma \in \mathbb{R}$ measures the strength of the nonlinear interaction while its sign classifies it into attractive for $\gamma < 0$, and repelling for $\gamma > 0$. In this context the coherent states are usually called bound states or, in the translation invariant case $V \equiv 0$, solitons. The Hartree Equation has exactly the same $X,\ J,\ T$ and Q but the superquadratic term in the energy becomes nonlocal:

$$\mathscr{E}(u) = \frac{1}{2}\int_{\mathbb{R}^n} |\nabla u(x)|^2 dx + \frac{1}{2}\int_{\mathbb{R}^n} V(x)|u(x)|^2 dx + \frac{\gamma}{4}\int_{\mathbb{R}^n} K(x,y)|u(x)|^2|u(y)|^2 dxdy, \quad (8)$$

where the kernel $K \geq 0$.

3 Coherent States

This section illustrates how one can find all coherent states of equations of type (1) i.e., all solutions of (3), and how one can determine their orbital stability. Traditionally, large coherent structures are found via variational methods, for example as minimizers of the energy subject to a fixed value of the second conserved quantity. However, as we shall see in the next subsection, the variational techniques are not capable of finding all coherent states. Instead we will show in Sect. 3.2 how bifurcation methods, in particular the analytical global bifurcation theory [8], can be enhanced to determine all coherent states, and their orbital stability, see Sect. 3.3.

3.1 Variational Methods: Existence and Stability of Ground States

The coherent states equation (3) coincides with the equation for the critical points of the energy $\mathscr{E}$ restricted to the level sets of the second conserved quantity Q i.e., it is the Euler-Lagrange equation for the energy subject to the constrain $Q = constant$. An important subset of the coherent states are the ground states which are minimizers of the energy under the constrain:

$$\mathscr{E}(\phi_\omega) = \min_{\phi \in X, Q(\phi)=\mu} \mathscr{E}(\phi), \quad \text{and} \quad \mu \in \mathbb{R} \tag{9}$$

If the energy (subject to the constrain) is bounded from below:

$$\exists m \in \mathbb{R} \text{ such that } \mathscr{E}(\phi) \geq m \ \forall \phi \in X \text{ with } Q(\phi) = \mu,$$

and coercive:

$$\lim_{\|\phi\|_X \to \infty, Q(\phi)=\mu} \mathscr{E}(\phi) = \infty,$$

then minimizing sequences $\{\phi_n\}_{n\in\mathbb{N}} \subset X$ are bounded and, due to the reflexivity of the Hilbert space X, have at least one weak limit, say $\phi_{n_k} \rightharpoonup \phi_0$. For ϕ_0 to be a ground state it must satisfy

$$Q(\phi_0) = \mu = \lim_{k\to\infty} Q(\phi_{n_k}), \quad \mathscr{E}(\phi_0) \leq \lim_{k\to\infty} \mathscr{E}(\phi_{n_k}) = \inf_{\phi\in X, Q(\phi)=\mu} \mathscr{E}(\phi) \tag{10}$$

Note that these conditions are not trivially satisfied as Q or $\mathscr{E}$ might not be weakly sequentially continuous even though they are both continuous with respect to the norm on X. These two issues are resolved by compactness arguments which show that weakly convergent minimizing sequences are actually strongly convergent i.e.,

$$\phi_{n_k} \overset{X}{\rightharpoonup} \phi_0 \quad \text{implies} \quad \lim_{k\to\infty} \|\phi_{n_k} - \phi_0\|_X = 0, \tag{11}$$

see for example [5, 12].

In NLS with confining potentials $V \geq 0$, $\lim_{|x|\to\infty} V(x) = \infty$, the potential restricts the domain of finite energy to a compact subspace. More precisely we have:

$$X = \left\{ \phi \in H^1(\mathbb{R}^n) \mid \int_{\mathbb{R}^n} V(x)|\phi(x)|^2 dx \right\} \tag{12}$$

is a Hilbert space with scalar product:

$$\text{with } \langle \phi, \psi \rangle_X = \langle \phi, \psi \rangle_{H^1} + \int_{\mathbb{R}^n} V(x)\overline{\phi}(x)\psi(x)dx,$$

which satisfies:

$$X \overset{compact}{\hookrightarrow} L^p(\mathbb{R}^n),\ 2 \leq p < \frac{2n}{n-2} \text{ if } n \geq 3,\ 2 \leq p < \infty \text{ if } n = 1, 2. \tag{13}$$

In particular, the above weakly convergent, minimizing subsequence:

$\phi_{n_k} \overset{X}{\rightharpoonup} \phi_0$ is strongly convergent in $L^p(\mathbb{R}^n)$ i.e., $\lim_{k\to\infty} \|\phi_{n_k} - \phi_0\|_{L^p} = 0$, for $2 \leq p < \dfrac{2n}{n-2}$.

Therefore

$$\mu = Q(\phi_{n_k}) = \frac{1}{2}\int_\Omega |\phi_{n_k}(x)|^2 dx \overset{k\to\infty}{\longrightarrow} \frac{1}{2}\int_\Omega |\phi_0(x)|^2 dx = Q(\phi_0)$$

and

$$\int_\Omega |\phi_{n_k}(x)|^{p+2} dx \overset{k\to\infty}{\longrightarrow} \int_\Omega |\phi_0(x)|^{p+2} dx.$$

Combining the above with the weak lower semicontinuity of the first two (kinetic and potential) terms in the energy which are convex, we deduce that (10) holds and ϕ_0 is a ground state. Moreover, the following inequality holds

$$\mathscr{E}(\phi_0) \geq \inf_{\phi \in X, Q(\phi)=\mu} \mathscr{E}(\phi) = \lim_{k\to\infty} \mathscr{E}(\phi_{n_k}),$$

which is opposite to (10). Therefore, on this minimizing subsequence, the kinetic and potential terms must be convergent which combined with the convergence of Q gives $\|\phi_{n_k}\|_X \to \|\phi_0\|_X$ in addition to $\phi_{n_k} \rightharpoonup \phi_0$. The strong convergence (11) now follows from the uniform convexity of the Hilbert space X.

Essential in the above argument is the compactness of the embeddings (13). Heuristically, one might think that $\int_{\mathbb{R}^n} V(x)|\phi(x)|^2 dx < \infty$ and $\lim_{|x|\to\infty} V(x) = \infty$ forces a "uniform decay at infinity" on $\phi \in X$ which does imply compactness, see [12, Section 1.7]. But this is not quite correct since if ϕ is a countable sum of smoothed characteristic functions of disjoint annuli with the same center and exterior and interior radius growing to infinity we have $\lim_{|x|\to\infty}\phi(x) \neq 0$, but $\phi \in X$ if the Lebesgue measure (volume) of the annuli converges to zero sufficiently fast. However, the argument in [12, Section 1.7] can be adapted to prove (13) as follows. Consider an arbitrary bounded sequence

$$\{\phi_n\}_{n\in\mathbb{N}} \subset X \text{ with } \|\phi_n\|_X \leq M, \text{ for all } n \in \mathbb{N}$$

Since X is Hilbert the sequence has a weakly convergent subsequence

$$\phi_{n_k} \overset{X}{\rightharpoonup} \phi_0 \in X.$$

Then, for each $\varepsilon > 0$ we can choose $R > 0$ such that

$$V(x) > 16\varepsilon^{-2} \max\left\{M^2, \int_{\mathbb{R}^n} V(x)|\phi_0(x)|^2 dx\right\}, \text{ for } |x| > R.$$

Consequently, we have

$$M^2 \geq \int_{\mathbb{R}^n} V(x)|\phi_{n_k}(x)|^2 dx \geq \int_{|x|>R} V(x)|\phi_{n_k}(x)|^2 dx \geq \frac{16M^2}{\varepsilon^2}\int_{|x|>R} |\phi_{n_k}(x)|^2 dx,$$

and

$$\begin{aligned}\int_{\mathbb{R}^n} V(x)\|\phi_0(x)\|^2 dx &\geq \int_{|x|>R} V(x)|\phi_0(x)|^2 dx \\ &\geq \frac{16\int_{\mathbb{R}^n} V(x)|\phi_0(x)|^2 dx}{\varepsilon^2}\int_{|x|>R} |\phi_0(x)|^2 dx,\end{aligned}$$

which imply

$$\begin{aligned}\left(\int_{|x|>R} |\phi_{n_k}(x) - \phi_0(x)|^2 dx\right)^{1/2} &\leq \left(\int_{|x|>R} |\phi_{n_k}(x)|^2 dx\right)^{1/2} \\ &+ \left(\int_{|x|>R} |\phi_0(x)|^2 dx\right)^{1/2} \leq \varepsilon/2.\end{aligned}$$

Now:

$$\begin{aligned}\|\phi_{n_k} - \phi_0\|_{L^2(\mathbb{R}^n)} &= \|\phi_{n_k} - \phi_0\|_{L^2(|x|<R)} + \|\phi_{n_k} - \phi_0\|_{L^2(|x|>R)} \\ &< \|\phi_{n_k} - \phi_0\|_{L^2(|x|<R)} + \varepsilon/2.\end{aligned}$$

But, by Rellich-Kondrachov Theorem, $H^1(|x| < R)$ is compactly embedded in $L^2(|x| < R)$ which means that the X hence H^1 weakly convergent sequence $\{\phi_{n_k}\}$

is strongly convergent in $L^2(|x| < R)$ and we can choose $k(\varepsilon) \in \mathbb{N}$ such that $\|\phi_{n_k} - \phi_0\|_{L^2(|x|<R} < \varepsilon/2$ for all $k > k(\varepsilon)$. All in all, for each $\varepsilon > 0$ we can find $k(\varepsilon) \in \mathbb{N}$ such that $\|\phi_{n_k} - \phi_0\|_{L^2(\mathbb{R}^n)} < \varepsilon$ for all $k > k(\varepsilon)$ i.e., ϕ_{n_k} converges strongly (in norm) in $L^2(\mathbb{R}^n)$. Moreover, ϕ_{n_k} bounded in X hence in $H^1(\mathbb{R}^n)$ also implies, via Sobolev embedding, that it is bounded in $L^{2n/(n-2)}(\mathbb{R}^n)$ and, by interpolation, convergent to ϕ_0 in L^p, $2 \leq p < 2n/(n-2)$. So, any bounded sequence in X has a convergent subsequence in $L^p(\mathbb{R}^n)$, $2 \leq p < 2n/(n-2)$ if $n \geq 3$, $2 \leq p < \infty$ if $n = 1, 2$, which implies (13).

However, in general, the verification of (10) requires *concentration compactness*, see [36, 37] or [12, Section 1.7]. This theory will be discussed in a different context in the next subsection. Suffices to say that in the NLS example it covers the case of non-confining potentials: $\lim_{|x|\to\infty} V(x) = 0$, (when the energy is bounded from below.)

The main difference between the ground states given by (9) and other solutions of (3) (called excited states) is that the former are in general stable under the dynamics:

Theorem 3.1 *Fix $\mu \in \mathbb{R}$ and assume the set of ground states,*

$$G = \{\phi_\omega \in X \mid \phi_\omega \textit{ solves } (9)\},$$

is non-empty. Fix $\phi_0 \in G$ and further assume that any minimizing sequence of (9) has a strongly convergent subsequence in X. Then for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $u_0 \in X$ with $\|u_0 - \phi_0\|_X < \delta$ we have that the solution $u(t)$ of the wave equation (1) with initial condition $u(0) = u_0$ remains within ε distance from G for all times.

Proof Suppose contrary, there is an $\varepsilon > 0$, a sequence $\{u_n\}_{n\in\mathbb{N}} \subset X$ with $\|u_n - \phi_0\|_X \to 0$ and a sequence of times $\{t_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ such that the solutions $u_n(t)$ of the wave equation (1) with initial condition $u(0) = u_n$ satisfy

$$\operatorname{dist}(u_n(t_n), G) = \inf\{\|u(t_n) - \psi\|_X \mid \psi \in G\} \geq \varepsilon.$$

By continuity of Q, we have $Q(u_n) \to Q(\phi_0) = \mu$, and, by using its bilinear form (2), we can find $\{\lambda_n\}_{n\in\mathbb{N}} \subset \mathbb{R}$ such that

$$Q(\lambda_n u_n) = \lambda_n^2 Q(u_n) = \mu \quad \text{and} \quad \lambda_n \to 1.$$

We now claim that $\{\lambda_n u_n(t_n)\}_{n\in\mathbb{N}} \subset X$ is a minimizing sequence for (9). Indeed, by conservation of Q along solutions of (1) we have $Q(u_n(t)) = Q(u_n)$ for all $t \in \mathbb{R}$ in particular:

$$Q(\lambda_n u_n(t_n)) = \lambda_n^2 Q(u_n(t_n)) = \lambda_n^2 Q(u_n) = \mu,$$

while by conservation of the energy we have:

$$\mathscr{E}(\lambda_n u_n(t_n)) = \mathscr{E}(\lambda_n u_n) \to \mathscr{E}(\phi_0) = \min_{\phi \in X, Q(\phi)=\mu} \mathscr{E}(\phi),$$

where the convergence follows from the continuity of $\mathscr{E}$ w.r.t. the norm in X.

Now, since $\{\lambda_n u_n(t_n)\}_{n\in\mathbb{N}} \subset X$ is a minimizing sequence for (9), it has, by hypothesis, a convergent subsequence to some $\phi_1 \in X$ i.e., $\|\lambda_n u_n(t_n) - \phi_1\|_X \to 0$. But, by continuity of Q and $\mathscr{E}$ we have:

$$Q(\phi_1) = \lim_{k\to\infty} Q(\lambda_{n_k} u_{n_k}(t_{n_k})) = \mu,\ \mathscr{E}(\phi_1) = \lim_{k\to\infty} \mathscr{E}(\lambda_{n_k} u_{n_k}(t_{n_k})) = \mathscr{E}(\phi_0) = \min_{\phi\in X, Q(\phi)=\mu} \mathscr{E}(\phi)$$

i.e. $\phi_1 \in G$. Therefore we have:

$$\operatorname{dist}(u_n(t_n), G) \le \|u_n(t_n) - \phi_1\|_X \le |1-\lambda_n|\|u_n\|_X + \|\lambda_n u_n(t_n) - \phi_1\|_X \to 0,$$

since $\lambda_n \to 1$ and $\|\lambda_n u_n(t_n) - \phi_1\|_X \to 0$. This contradicts our assumption that $\operatorname{dist}(u_n(t_n), G) \ge \varepsilon$ and finishes the proof of the theorem. □

Remark 3.1 Note that all examples discussed above satisfy the hypotheses of the Theorem. Moreover, with the exception of the case $V \equiv 0$ the set of ground states is unique up to the symmetries induced by the semigroup T :

$$G = \{T(s)\phi_0 \mid \text{for some } \phi_0 \text{ which solves (9) and all } s \in \mathbb{R}\}, \tag{14}$$

provided μ is small, see for example [3, 47].

Note that the invariance of both Q and $\mathscr{E}$ w.r.t T automatically implies $G \supseteq \{T(s)\phi_0 \ : \ s \in \mathbb{R}\}$ if ϕ_0 solves (9). However the equality between the two sets implies orbital stability:

Corollary 1 *Under the assumptions in Theorem 3.1, if, in addition,* (14) *holds, then the ground state ϕ_0 is orbitally stable.*

Proof By the theorem, for each $\varepsilon > 0$ there exists $\delta > 0$ such that for all $u_0 \in X$ with $\|u_0 - \phi_0\|_X < \delta$ we have:

$$\sup_{t\in\mathbb{R}} \operatorname{dist}(u(t), G) < \varepsilon.$$

But, in this case

$$\operatorname{dist}(u(t), G) = \inf_{s\in\mathbb{R}} \|u(t) - T(s)\phi_0\|_X$$

which implies orbital stability, see the definition below (3). □

Remark 3.2 More generally, the orbital stability result in the previous corollary holds even if the ground states are not unique (up to the action of T) provided that the orbit $T(s)\phi_0$ is separated from the orbits of the other ground states by a fixed distance $d > 0$. Just use $\varepsilon < d/2$ in the above proof and note that the set of points at distance less the $d/2$ from the orbit of ϕ_0 is disjoint from the set of all points at distance less the $d/2$ from the orbits of all other ground states while the map $\mathrm{dist}(u(t), G)$ is continuous in time.

In the case $V \equiv 0$, it is necessary to mode out the second (hidden) symmetry, namely the invariance of both energy and Q with respect to translations, see for example [12, Section 8.3]. One obtains:

$$\sup_{t\in\mathbb{R}} \inf_{s\in\mathbb{R}, y\in\mathbb{R}^n} \|u(t) - T(s)\phi_0(\cdot + y)\| < \varepsilon.$$

Another advantage of the global minimization problem (9) over other solutions of (3) is that certain manipulation of the functions $\phi \in X$, such as symmetrization, can lower the energy and provide information on the shape of the ground states. For example, in NLS with spherically symmetric potentials the ground states are also spherically symmetric, see [12, Chapter 8]. In [3], see also [20] for a related result, the authors show that the ground states for the Hartree equation must localize at global minima of the potential in the limit $\mu \to \infty$, and, consequently, the following symmetry breaking phenomena occurs:

Theorem 3.2 *Consider the Hartree example* (8) *with an attractive nonlinearity* $\gamma < 0$, *and a continuous, bounded potential V which is invariant under a finite group of Euclidian symmetries on* $\mathbb{R}^n$. *If the action of the group is nontrivial on any global minima of V then the are* $\mu_0 \le \mu_1$ *such that the ground states with* $\mu < \mu_0$ *are invariant under the group of Euclidian symmetries but the ground states with* $\mu > \mu_1$ *are not invariant.*

The disadvantages of the minimization problem (9) are that it cannot give all coherent states i.e., all solutions of (3), and it may have no solutions. This is the case when the energy is not be bounded from below (even when constrained to $Q = const$) which occurs in the NLS example with critical and supercritical nonlinearities $p \ge 4/n$. The issue is sometimes resolved by reformulating the problem i.e., by finding the global minimum of a functional different from the energy. In [47] the authors use the following reformulation:

$$\min_{\phi\in X, \phi\neq 0} J^E(\phi) = \frac{\int_{\mathbb{R}^n} |\nabla\phi(x)|^2 + V(x)|\phi(x)|^2 + E|\phi(x)|^2}{\left(\int_{\mathbb{R}^n} |\phi(x)|^{p+2} dx\right)^{\frac{2}{p+2}}}$$

which is equivalent to:

$$\min_{\phi\in X, \int_{\mathbb{R}^n} |\phi(x)|^{p+2} dx = \mu} \int_{\mathbb{R}^n} |\nabla\phi(x)|^2 + V(x)|\phi(x)|^2 + E|\phi(x)|^2, \qquad \mu \in \mathbb{R},$$

see [12, Chapter 8] for other possible reformulations. An important result in [47] is the existence of a solution to (7) for each $E > E_0$ where $-E_0$ is the lowest eigenvalue of $-\Delta + V$. However, the method is still limited to finding a subset of solutions of (7) and does not provide any direct information regarding their dynamical stability (the latter may be fixed by combining information on the Hessian of J^E at a minima with the techniques described in Sect. 3.3). One expects that these variational reformulations will also lead to information on the shape and localization of the ground states, consequently symmetry breaking results may be proven for critical and supercritical nonlinearities, see [20]. While other critical points of the energy or other associated functionals may be found via mountain pass techniques, see for example [42], the variational methods do not provide a systematic method to identify all solutions of (3).

3.2 Bifurcation Methods

This section discusses recent progress towards finding all coherent states of a nonlinear wave equation i.e., all solutions of equation (3). We will focus on the NLS example for which (3) becomes (7):

$$F(\phi_E, E) = (-\Delta + V + E)\phi_E + \gamma|\phi_E|^p\phi_E = 0,$$

where the potential $V : \mathbb{R}^n \mapsto \mathbb{R}$, is first assumed to be non-confining, $\lim_{|x|\to\infty} V(x) = 0$, $V \in L^q(\mathbb{R}^n) + L^\infty(\mathbb{R}^n)$ for some $q \geq 1$, $q > n/2$. The case of confining potentials is discussed in Remark 3.8. We will assume that the power of the nonlinearity satisfies $0 < p < \infty$, if $n = 1$ or 2 and $0 < p < 4/(n-2)$ if $n \geq 3$, which insures local well posedness of the time dependent equation (1) with initial data in the Sobolev space $H^1(\mathbb{R}^n)$. Of special interest are the ground states which, for the purpose of this presentation, will be defined as coherent states i.e., solutions of (7) that satisfy $\phi_E(x) > 0$, $\forall x \in \mathbb{R}^n$, modulo multiplication by a complex number of modulus one (modulo rotations). The coherent states can be viewed as zeroes of the map $F : H^1(\mathbb{R}^n) \times \mathbb{R} \mapsto H^{-1}(\mathbb{R}^n)$ where the Sobolev spaces are endowed with their *real* Hilbert space structure in order for F to be differentiable. Note that F is equivariant under rotations:

$$F(e^{i\theta}\phi, E) = e^{i\theta}F(\phi, E),\ \theta \in \mathbb{R},$$

hence the solution set for (7) is invariant under rotations.

We will study the solutions of (7) in the subdomain of $H^1(\mathbb{R}^n) \times \mathbb{R}$ where its linearization is Fredholm. We have

$$D_\phi F(\phi, E)[u + iv] =$$

$$\begin{bmatrix} -\Delta + V + E + \gamma(p+1)|\phi|^p - 2\gamma(\Im\phi)^2|\phi|^{p-2} & 2\gamma(\Re\phi)(\Im\phi)|\phi|^{p-2} \\ 2\gamma(\Re\phi)(\Im\phi)|\phi|^{p-2} & -\Delta + V + E + \gamma|\phi|^p + 2\gamma(\Im\phi)^2|\phi|^{p-2} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

where we separated the real and imaginary parts of the complex valued functions involved into the first and second component. So,

$$D_\phi F(\phi, E) = \begin{bmatrix} -\Delta + E & 0 \\ 0 & -\Delta + E \end{bmatrix} + \mathscr{V}(\phi)$$

where, for any $\phi \in H^1(\mathbb{R}^n)$,

$$\mathscr{V} = \begin{bmatrix} V + \gamma(p+1)|\phi|^p - 2\gamma(\Im\phi)^2|\phi|^{p-2} & 2\gamma(\Re\phi)(\Im\phi)|\phi|^{p-2} \\ 2\gamma(\Re\phi)(\Im\phi)|\phi|^{p-2} & V + \gamma|\phi|^p + 2\gamma(\Im\phi)^2|\phi|^{p-2} \end{bmatrix}.$$

is a relatively compact perturbation of the diagonal operator $-\Delta + E$ on $H^{-1} \times H^{-1}$ with domain $H^1 \times H^1$ (or on $L^2 \times L^2$ with domain $H^2 \times H^2$), see for example [46]. Since the latter has essential (continuous) spectrum the interval on the real line $[E, \infty)$ we get via Weyl's Theorem:

Lemma 3.1 *$D_\phi F(\phi, E)$ is Fredholm (of index zero) iff $E > 0$. At the left boundary, $E = 0$, zero is at the edge of its essential (continuous) spectrum while for $E < 0$ zero is inside the essential (continuous) spectrum.*

We will restrict ourselves to the domain $H^1 \times (0, \infty)$, i.e. $\{E > 0\}$, which will now be called the bifurcation diagram. Note that, for $E < 0$, by the limiting absorbtion principle, there are no nontrivial solutions of (7) under mild assumptions on their decay rates, see [6]. Obviously, $(\phi_E \equiv 0, E)$, $E \in \mathbb{R}$ solves (7).

For a while we will assume:

(SA) $-\Delta + V$ has at least one negative eigenvalue. The lowest will be denoted by $-E_0$.

Note that the assumption holds in space dimensions $n = 1, 2$ for non-trivial, negative potentials and requires potentials with sufficiently large negative parts in dimensions $n \geq 3$. As shown for example in [27, 29, 44], hypothesis (SA) leads to a pitchfork bifurcation at $(\phi_{E_0} \equiv 0, E_0)$, which creates exactly one curve (modulo rotations) of non-trivial ground states, see Fig. 1. Moreover, if V is invariant under a group of symmetries then so are the ground states on this branch. In particular, if V is a symmetric, double well potential, see Fig. 2 top panel, then the profile of the ground states is equally distributed between the two wells.

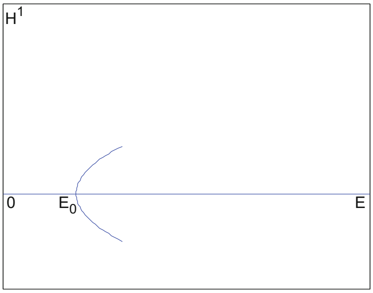


Fig. 1 Bifurcation diagram for bound states. Only the trivial zero coherent state and the small ground states are represented for the attractive case $\gamma < 0$. For the repelling case, $\gamma > 0$, the branch points the other way, i.e. $E < E_0$ along the branch

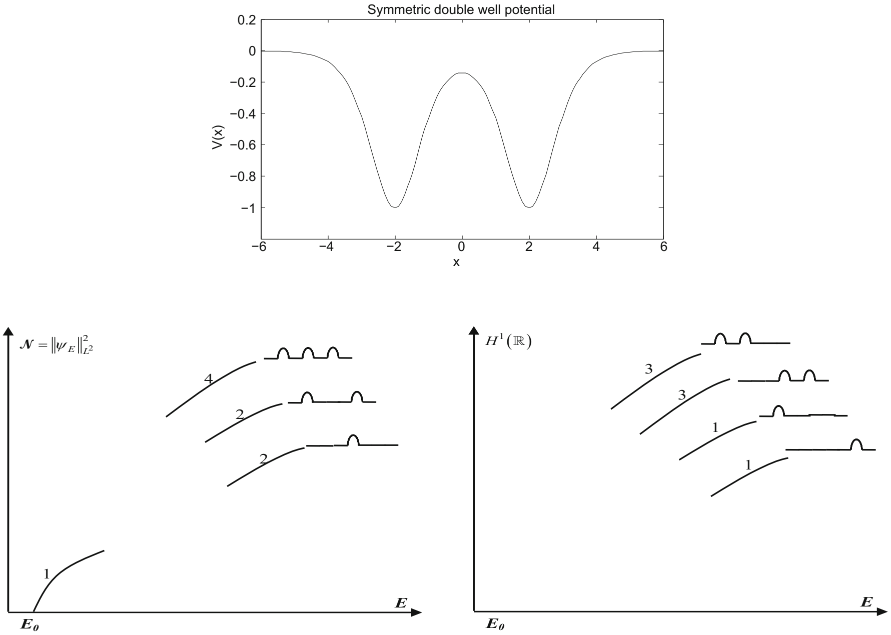


Fig. 2 An one dimensional even potential (top panel) and a sketch of the corresponding symmetric (bottom left panel) and asymmetric (bottom right panel) ground state branches for subcritical nonlinearity, $p < 2$. The number on top of each branch gives the number of negative eigenvalues for the linearized operator while the shape on the right shows the actual shape of the solution on the branch

We are going to rely on global bifurcation theory which, besides a Fredholm linearization, also requires compactness either of the solution set of (7) (in the analytic case, see [8]) or of the map $\tilde{F}$ (in the continuous case where degree theory is used, see [45]) where $\tilde{F}$ is obtained by transforming (7) into a fixed point problem, for example:

$$\phi = (-\Delta+1)^{-1}\psi,\ \psi=(1-E)\phi-V(x)\phi-\gamma|\phi|^p\phi \overset{def}{=} \tilde{F}(\psi,E),\ \tilde{F}: H^{-1} \times\mathbb{R}\mapsto H^{-1} \tag{15}$$

Note that for ϕ defined on a bounded domain, $\Omega \subset \mathbb{R}^n$, compactness of $\tilde{F}$ follows from compactness of Sobolev embeddings $H^1_0(\Omega) \hookrightarrow H^{-1}(\Omega),\ L^q(\Omega) \hookrightarrow H^{-1}(\Omega),\ 2 \leq q < 2n/(n-2)$, however on $\mathbb{R}^n$ the situation is much more delicate. For repelling nonlinearities, $\gamma > 0$, the problem is analyzed in [24] where the authors prove uniform bounds on the solutions of the inequality $|F(\phi(x),E)| \leq \Psi(x)$ to obtain the compactness necessary for defining a degree for $\tilde{F}$ given in (15). Then global bifurcation theory implies that from $(0,E)$, where $-E$ is any negative eigenvalue of $-\Delta + V$ with odd multiplicity, bifurcate branches of non-trivial solutions of (7) which“end up” either at the boundary of the bifurcation diagram, or at $(0, E_1)$ where $-E_1$ is a different eigenvalue of $-\Delta + V$. Note that the results cannot give any information on existence of other bifurcation points along this branches, or on existence of other branches that may not connect to the trivial solution, or on the exact region or point where each branch ends up. In the particular case of the ground states bifurcating from $(0, E_0)$, the authors of [25] use further comparison theorems to infer that the branch approaches the left boundary $\{E = 0\}$ of the bifurcation diagram, i.e. the part of the boundary where zero is at the edge of the continuous spectrum of the linearized operator, see Lemma 3.1 and Fig. 1 but note that for $\gamma > 0$ the branch points the other way.

To avoid the difficult issue of bifurcations from continuous spectrum let us focus first on the attractive case $\gamma < 0$. Suppose we assume that $D_\phi F(\phi_E, E),\ E > E_0$, is nonsingular along the ground state branch emerging from $(0, E_0)$. If we can now show that the branch can be uniquely continued for all $E > E_0$, i.e. it approaches the right boundary of the bifurcation diagram, and we can identify the limit point $\lim_{E\to\infty} \phi_E$, and if the linearization at the limit point must have two (or more) negative eigenvalues, then we have a contradiction since the linearization had only one negative eigenvalue near $(0, E_0)$. Hence the existence of a singular point along this branch is guaranteed and the resulting bifurcation can give us new branches of ground states. We have actually sketch a result that not only shows there are ground states for all $E > E_0$ but improves on the result of Theorem 3.2 by identifying a symmetry breaking bifurcation:

Theorem 3.3 *Consider an attractive nonlinearity $\gamma < 0$, and a potential V which is invariant under a finite group of Euclidian symmetries on $\mathbb{R}^n$. Assume that the action of the group is nontrivial on any critical point of $V(x)$ different from $x = 0$, and assume that $x = 0$ is a non-degenerate critical point of V different from a minima. Then the branch of ground states bifurcating from $(0, E_0)$ undergoes a second bifurcation past which the symmetric bound states become orbitally unstable. Moreover, one of the new branches emerging from the bifurcation point is made of asymmetric ground states which are generally orbitally stable.*

In particular, for a double well potential, the bifurcation is of pitchfork type and the emerging branch is made of ground states which localize in one of the two wells, see Fig. 3 bottom panels and reference [28].

The theorem is proven using three intermediate results which are important themselves because they determine all ground states provided a few remaining obstacles are surmounted, see Remarks 3.3, 3.4 and 3.5 below. The first result is:

Theorem 3.4 *If a C^1 branch of coherent states approaches the top or bottom boundary of the bifurcation diagram, i.e. $E \to E_*,\ 0 < E_* < \infty,\ \|\phi_E\|_{H^1} \to \infty$, then we have:*

$$Q(\phi_E) \to \infty, \qquad \frac{\|\phi_E\|_{L^{p+2}}^{p+2}}{Q(\phi_E)} \to 0.$$

If a C^1 branch of coherent states approaches the right boundary $E \to \infty$, then there exists $b > 0$ such that

$$\frac{\|\phi_E\|_{L^{p+2}}^{p+2}}{E^{2/p+1-n/2}} \to b, \quad \frac{Q(\phi_E)}{E^{2/p-n/2}} \to -\gamma\frac{(2-n)p+4}{2p+4}b, \quad \frac{\|\nabla\phi_E\|_{L^2}^2}{E^{2/p+1-n/2}} \to -\gamma\frac{npb}{2p+4}.$$

These estimates are obtained from the ordinary differential equation valid along these branches:

$$\frac{d\mathscr{E}}{dE}(\phi_E) = -E\frac{dQ}{dE}(\phi_E) \tag{16}$$

combined with the equation (7) and Pohozaev's identity (essentially the L^2-scalar product between (7) and $x \cdot \nabla\phi_E$), which leads to closed differential inequalities for $\|\phi_E\|_{L^{p+2}}^{p+2}$, see [28] and [35] for details.

Remark 3.3 The caveat is that the theorem does not yet cover the case of branches undergoing infinitely many bifurcations in all neighborhoods of the boundary (hence they cannot be parametrized by a C^1 map in E in any neighborhood of the boundary). However, most of these peculiar situations have been resolved in the sense that they lead to similar estimates which can be used in the next results, see [35].

The theorem is essential in finding the limit points of the bound state branches at the boundary of the bifurcation diagram. For example, at the $\{E \to E_*,\ 0 < E_* < \infty,\ \|\phi_E\|_{H^1} \to \infty\}$ part of the boundary, the estimate above, combined with the (15) form of the equation and the fact that $(-\Delta + 1)^{-1} : H^{-1} \mapsto H^1$ is an isomorphism imply that $\phi_E/\sqrt{Q(\phi_E)}$ converges in H^1 to a solution of $-\Delta\psi + E_*\psi = 0$. The latter has only the zero solution which contradicts $\|\phi_E/\sqrt{Q(\phi_E)}\|_{L^2} \equiv 1$.

A contradiction is also obtained at the $E = 0$ from hypothesis (SA) and comparison principles for the linearized (self-adjoint) operator. The comparison principle relies heavily on the fact that the nonlinearity is always negative, see [35].

At the $E \to \infty$ portion of the boundary the estimates imply that the change of variables:

$$\psi_E(x) = E^{-1/p}\phi_E(E^{-1/2}x + x_0)),\ x_0 \in \mathbb{R}^n \tag{17}$$

leads to a uniformly bounded curve $E \mapsto \psi_E$ in H^1 and transforms (7) into:

$$-\Delta\psi_E + E^{-1}V(E^{-1/2}x + x_0)\psi_E + \psi_E + \gamma|\psi_E|^p\psi_E = 0$$

which formally converges to

$$-\Delta\psi + \psi + \gamma|\psi|^p\psi = 0. \tag{18}$$

The rigorous result is:

Theorem 3.5 *There are no coherent states approaching $\{E = 0\}$ and $\{E > 0, \|\cdot\|_{H^1} \to \infty\}$ boundary of the bifurcation diagram. Ground states approaching $E \to \infty$ boundary converge in H^1, modulo the re-scaling* (17) *and rotations in the complex plane, to a superposition of positive solutions of* (18) *each localized at a critical point of the potential V.*

Note that the result at $E \to \infty$ has been conjectured in [47]. Our convergence argument uses concentration compactness [12, Section 1.7] combined with a rather delicate analysis of bifurcations from infinity, see [35] for details. For example, if splitting of profiles would occur then at least one of them must move towards infinity, and since $\lim_{|x|\to\infty} V(x) = 0$ we can show that the profile converges to a solution of the equation without potential. But we are dealing with ground states so this solution must be positive modulo rotations. It is known that, modulo translations, there is only one such solution and the properties of the linearized operator at this solution are also known. Using a Lyapunov-Schmidt decomposition based on the linearized operator we show that there are no bifurcations from solutions of the translation invariant problem concentrated at infinity into a solution of our problem with potential under mild hypotheses on the behavior of the potential at infinity.

Remark 3.4 New solutions of translation invariant NLS equation (18) may be discovered based on Theorem 3.5. Indeed, the corresponding result for excited states (i.e., solutions of (7) which are not ground states) is that as $E \to \infty$ the re-scaled ψ_E either converges strongly to a superposition of positive solutions of (18), some of them multiplied by -1, and each localized at a critical point of the potential V, or (18) must have solution that cannot be obtained from the positive one via translations or rotations in the complex plane. There are no such solutions in space dimension $n = 1$ (hence the theorem applies to all coherent states in one space dimension) but their existence/non-existence in higher dimensions is an open problem. Note that, in principle, the re-scaled ψ_E can be numerically traced along excited state branches at large E. If profiles that change sign emerge (instead of profiles in which the positive part drifts away from the negative part) then the profile is a new solution of (18). The algorithm can start from excited states of (7) which bifurcate from zero at the second and higher eigenvalues of the linear operator $-\Delta + V$. The existence of such eigenvalues is guaranteed for sufficiently negative potentials.

To obtain all limit points of the ground state branches at $E \to \infty$ we combine Theorem 3.5 with the local bifurcation result:

Theorem 3.6 *At $E = \infty$, from any superposition of positive or negative solutions of* (18) *each localized at distinct, non-degenerate, critical points of the potential V bifurcates, modulo the re-scaling* (17) *and rotations in the complex plane, exactly one curve of coherent states for* (7). *These coherent states have as many nodal points as the number of sign changes in the superposition. The number of negative* *eigenvalues of the linearization calculated at these coherent states can be computed with the formula: $k + n_1 + n_2 + \cdots + n_k$ where k is the number of profiles and n_j, $j = 1, 2, \ldots k$ is the number of negative directions for the Hessian of the potential calculated at the critical point where the j^{th} profile localizes.*

See Fig. 2 for an illustration of this theorem in the case of a double well potential.

The theorem is reminiscent of the result in [43], see also [2, 15], for the semiclassical limit. Note that we are not in the semiclassical limit, our re-scaled equation immediately below (17) differ by the E^{-1} factor in front of the potential making it an even more degenerate problem. Moreover, our method can be adapted to the semiclassical case and gives a stronger result by not only showing uniqueness of such solutions but also providing their parametrizations and spectral properties of the linearized operator. The extension of Theorem 3.6 to degenerate critical points is still open.

Remark 3.5 As of now Theorems 3.5–3.6 do not exclude or treat the case of multiple profiles localizing at the *same* critical point of V, see [42] for a related result. For ground states the phenomenon does not occur at local minima, but the other cases are still open.

The compactness argument at $E \to \infty$ can be extended in the interior of the bifurcation diagram domain to obtain:

Theorem 3.7 *Any set of ground-states (ϕ_E, E) which is bounded in $H^1(\mathbb{R}^n) \times (E_1, \infty)$, where $E_1 > 0$, is relatively compact and any limit point is a solution of* (7).

Now, Theorem 3.3 follows from a contradiction argument. Suppose that along the symmetric branch starting at $(0, E_0)$ no eigenvalues of the linearized operator cross zero. Then, by Lemma 3.1 and the implicit function theorem the branch can be continued and remains symmetric until it reaches the boundary of the bifurcation diagram. By Theorem 3.5 it will have $E \to \infty$ and, in this limit, it will converge, modulo re-scaling (17), to a superposition of positive solutions of (18) each localized at a critical point of V (some may localize at the same critical point). If the limit is localized at $x = 0$ then from Theorem 3.6 we deduce that the linearized operator along this branch at large E has at least two negative eigenvalues (one plus the number of negative directions for the Hessian of the potential at $x = 0$)in contradiction with the fact that it had only one negative eigenvalue near E_0. If the limit has a profile localizing at a non-zero critical point x_0, by symmetry it must have a profile (positive solution of (18)) at each point in the orbit of x_0 under the action of the Euclidian group. In this case the number of negative eigenvalues of the linearized operator at large E is at least the number of profiles, see Theorem 3.6, which is at least the number of points in the orbit. By hypothesis the latter is at least 2 and gives a contradiction. Consequently, there must be an E_*, $E_0 < E_* < \infty$, such that an eigenvalue of $D_\phi F(\phi_E, E)$ converges to zero as $E \nearrow E_*$. By Theorem 3.7 there is a limit point (ϕ_{E_*}, E_*) and local bifurcation theory can be used to analyze the branches emerging from this point.

More importantly, for analytic nonlinearities (p an even, positive integer), our Theorem 3.7 combined with global bifurcation theory imply that ground states not only organize themselves in smooth manifolds but the manifolds can also be smoothly continued past their singularity points (i.e. bifurcation points) until they either form loops or reach the boundary of the bifurcation diagram region $H^1(\mathbb{R}^n) \times (0, \infty)$, see [8, 35]. Note that if we somehow exclude the cases described in Remarks 3.3 and 3.5 then Theorems 3.5–3.6 give us all the limit points at the boundary and we can now trace back all ground-states.

For example consider the symmetric double well potential in one space dimension which has three critical points, see top panel in Fig. 2. We claim that all ground states of this problem are given by the bottom left panel in Fig. 3. A similar picture can be obtained for excited states with a fixed number of nodal points (zeroes). Indeed, by first restricting the analysis to the Banach subspace of even functions in H^1 we get via Theorem 3.6 (and modulo rotations) the three curves near $E = \infty$ in addition to the one given by (SA) near E_0, see the left panel of Fig. 2. Global bifurcation theory says that the latter connects smoothly with one of the former. Hence, we have three possibilities, two are presented in the upper panels of Fig. 3,

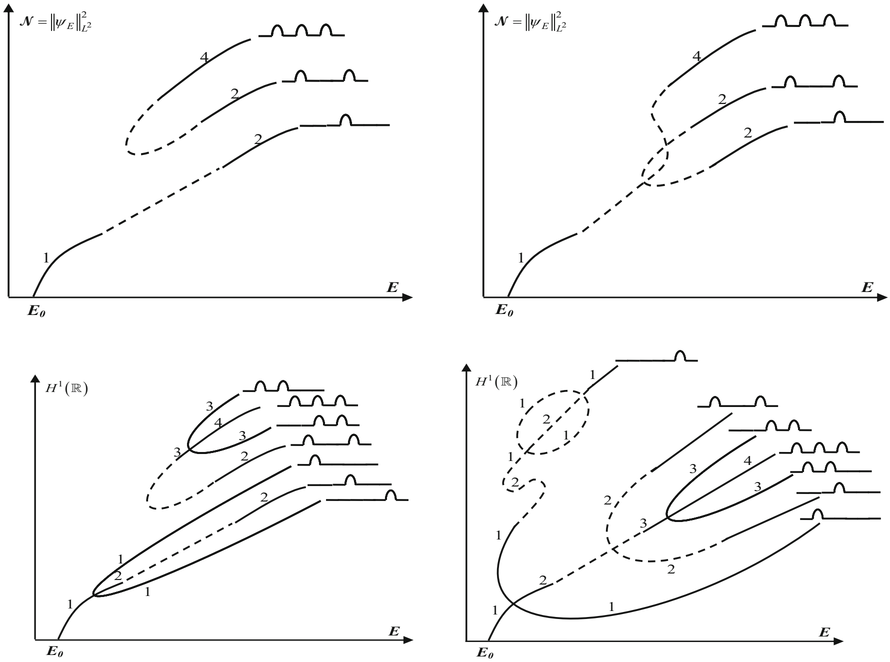


Fig. 3 Top panel shows two possible ways the even branches connect. The third is similar to the left panel. Bottom panel shows how the asymmetric branches will bifurcate from the symmetric ones in the two cases. In all figures the dotted lines show region where the branches are not completely understood, i.e. "snaking" or pitchfork like bifurcation may occur but the latter must lead to loops, see the top branch in the bottom right panel

the third is similar to the left panel. The remaining two curves of symmetric ground states must connect with each other since, again by Theorem 3.5, they cannot end up on top or left boundaries of the bifurcation diagram, neither can they end at $(0, E_0)$ due to the uniqueness of the branch emerging from this point. Sturm-Liouville theory allows only a simple eigenvalue to cross zero at each bifurcation point, so, to match the number of negative eigenvalue on the symmetric branches we find that, in the case described in the left panels of Fig. 3, one more bifurcation point is needed on each (the turning point is already a bifurcation on the top curve). Since each of these bifurcations correspond to an antisymmetric eigenvector in the kernel of the linearized operator, one asymmetric branch emerges from each, see bottom left panel in Fig. 3. They already match the curves of asymmetric ground states given by Theorem 3.6 at $E = \infty$, hence the picture is complete. A similar counting argument can be done for the right panels in Fig. 3. Moreover, the same analysis can now be performed for excited states with one nodal point (one change of sign), two nodal points, etc, since Theorem 3.5 applies to them in one dimension, see Remark 3.4.

The above analysis did not include the multi-profile ground states which, as $E \to \infty$, may have more than one profile localizing at the local maxima of the potential, see Remark 3.5. Recent numerical investigations in [34] show that they are present and the branch starting from one profile at each minima when $E = \infty$ turns back and connects to a branch with two profiles both localizing at the local maxima $x = 0$, while the branch starting from a profile at each critical point connects to the branch with three profiles all of them localizing at $x = 0$. A rigorous understanding of this phenomena is underway.

Remark 3.6 Theorems 3.4–3.7 are valid in any space dimension, however, to obtain all ground states, the counting argument described above needs to be adapted when non-simple eigenvalues of the linearized operator cross zero. In practical applications the multiplicity of these eigenvalue is due the Euclidian symmetries of the underlying phenomenon hence its Hamiltonian. The symmetries can be used to simplify the normal form of the local bifurcation, see [16]. A case by case study is underway, beginning with potentials invariant under finite group of symmetries (such as under reflection w.r.t. hyperplanes, or generated by rotations with a fixed angle) and finishing with potentials invariant under continuous group of symmetries such as spherical ones.

Remark 3.7 Bifurcations from continuous spectrum may occur in the absence of the spectral hypothesis *(SA)*. Indeed, the sketch of proof for Theorem 3.5 showed that *(SA)* is essential in excluding branches which approach the $\{E = 0\}$ boundary. While we can build the picture of all ground states starting now from the branches given by Theorem 3.6 at large E, some of these branches will end up at $\{E = 0\}$. To complete the picture we need to find all limit points on this boundary, in particular we need to understand bifurcations from the edge of the continuous spectrum, see Lemma 3.1. A summary of recent progress in such problems can be found in [56]. Repelling nonlinearities $\gamma > 0$ also fall in this category as preliminary calculations show that all branches of coherent states end up at the left boundary $\{E = 0\}$, see [25] for a different method applicable to ground-states only. More complicated nonlinearities may also push the coherent states towards this boundary. For example $-|\phi|^2\phi + |\phi|^4\phi$ formally behave like an attractive nonlinearity near $(0, E_0)$ but at large bound-states the repelling part dominates. Hence a turning point is formed on the branch starting at $(0, E_0)$ and the conjecture is that it eventually approaches $\{E = 0\}$, see [22]. The study of such nonlinearities and more general ones is in progress.

Remark 3.8 Confining potentials $\lim_{|x|\to\infty} V(x) = \infty$ allow for stronger results compared to Theorems 3.4–3.7. Indeed, the bound states now belong to the Banach space $\{\phi \in H^1 : \int_{\mathbb{R}^n} V(x)|\phi(x)|^2 dx < \infty\}$ which embeds compactly in L^2, see the previous subsection. This implies that the linearized operator has purely discrete spectrum, that the set of solutions of (7) is relatively compact, and that the map $\tilde{F}$, see (15) is compact. In particular Theorems 3.4–3.7 are valid for all coherent states. Based on this observation a rigorous study vortices in rotating but confined Bose-Einstein Condensates is underway, see [26] for a recent summary of open problems, results and applications.

Remark 3.9 Non-analytical nonlinearities require compactness results stronger than the one given by Theorem 3.7 i.e., valid also for approximate solutions of (7). Such compactness is needed to construct a degree for the map $\tilde{F}$ in (15) on which the global bifurcation theory relies, see [45]. Such results hold for confining potentials, see remark above, or repelling nonlinearities $\gamma > 0$, see [24, 25]. The problem for non confining potentials combined with attractive nonlinearities is open.

3.3 Orbital Stability

Two of the most cited results in orbital stability of coherent structures are the ones by Grillakis, Shatah and Strauss in [18, 19]. One of its refinements [17], which is applicable to the Schrödinger and Klein-Gordon equations because of the diagonal structure of their linearization, implies that, in the example presented in the previous subsection, all branches with more than one negative eigenvalue in the spectrum of the linearized operator are unstable while the ones with exactly one negative eigenvalue are stable provided their L^2 norm is strictly increasing in E, see Fig. 3. However, neither the results in [18, 19] nor their numerous refinements cover all possible cases. For example, in the Schrödinger case with attractive nonlinearity, the first excited state bifurcating from zero at the second eigenvalue of $-\Delta + V$ is outside the scope of the current orbital stability theory. Thanks to hundreds of pages of proofs based on asymptotic stability techniques, see [53] and [51], we now know that this branch is unstable in the weakly nonlinear regime provided a resonance condition is satisfied. Is there a simpler way to study the stability of such coherent states, one that will not rely on weak nonlinearities and resonance conditions?

In the general framework presented in Sect. 2 the theory uses the Lyapunov functional: $u \mapsto \mathscr{E}(u) - \omega Q(u)$ to study the stability of the coherent states (ϕ_ω, ω)

which are solutions of (3) hence critical points of the Lyapunov functional. The results in [18, 19] exploit the fact that Q is invariant under the dynamics and can be summarized as follows: if ϕ_ω is a local minimizer of the Lyapunov functional *restricted* to the manifold $Q(u) = Q(\phi_\omega)$ then ϕ_ω is orbitally stable; if ϕ_ω is a saddle point of the Lyapunov functional *restricted* to the manifold $Q(u) = Q(\phi_\omega)$ with an odd number of negative directions i.e., odd number of negative eigenvalues of the Hessian *restricted* to the tangent space of the manifold $Q(u) = Q(\phi_\omega)$ at ϕ_ω, then ϕ_ω is linearly and orbitally unstable. More precisely, for stability the Hessian can be nonnegative or it can have one negative eigenvalue over the entire space X which disappears when the domain is restricted to the tangent space of the codimension one manifold $Q(u) = Q(\phi_\omega)$ which turns out to be equivalent with the condition $\partial_\omega Q(\phi_\omega) < 0$ at the particular ω under study. However, the theory leaves open the cases when the Hessian has more than one negative eigenvalue over the entire space X and an even number of them remain when restricting to the tangent space of $Q(u) = Q(\phi_\omega)$. The example in the above paragraph is in the case with two negative eigenvalues and no recent refinements of the theory can cover it. Also note that none of the refinements applies to the general framework described in Sect. 2 but to rather particular cases.

Is it possible to show that if the coherent state ϕ_ω is a saddle point of the Lyapunov functional *restricted* to the manifold $Q(u) = Q(\phi_\omega)$ then it is orbitally unstable? Note that the Lyapunov functional is actually invariant under the dynamics, hence initial data on the manifold with energy just below the energy of the coherent state evolve on a level set that takes it far away from the coherent state. More precisely, there is a fixed neighborhood of the orbit of the coherent state which is left by all orbits with initial data approaching the coherent state from a negative direction of the Hessian. This idea has been partially exploited in [18] but there the negative direction turns out to be an unstable direction of the linearized dynamics $\frac{dv}{dt} = JD^2E(\phi_\omega)[v]$ i.e., an eigenfunction of a positive eigenvalue of $JD^2E(\phi_\omega)$, hence an exponential growth of the distance between orbits leads to instability. This is not the case in the example discussed in the first paragraph of this subsection and in many others. But the point is that even in the absence of unstable directions for the linearized dynamics, the presence of negative direction for the Hessian suffices to prove a (much weaker) linear growth of distance between certain orbits in a small neighborhood of the coherent state which still implies instability. This work is still in progress.

Note that, if the answer to the above question is affirmative then the theory becomes a characterization of orbital stability i.e., in the general framework of (1) that is only invariant under the action $T(s)$ of a *one dimensional Lie group*, $s \in \mathbb{R}$, the coherent state ϕ_ω given by (3) is orbitally stable if and only if it is a local minimizer of the Lyapunov functional: $u \mapsto \mathscr{E}(u) - \omega Q(u)$ *restricted* to the manifold $Q(u) = Q(\phi_\omega)$ i.e., the Hessian over the whole space X of the Lyapunov functional can have at most one negative eigenvalue (counting multiplicity) and if it has one then $\partial_\omega Q(\phi_\omega)$ must be negative at the particular ω under study. Since the Hessian is basically the linearization of the equation for coherent structures (3), its eigenvalues change continuously along manifolds of solutions and cross zero only at bifurcation points. Therefore, the stability properties can now be deduced directly form the bifurcation diagrams, see for example Fig. 3.

4 Asymptotic Stability of Coherent States

When the techniques proposed in the previous Section lead to a new branch of orbitally stable coherent states for (1) or a bifurcation point involving both stable and unstable branches, the question is whether the dynamics of solutions starting near the branch can be described in detail. In particular, asymptotic stability would mean that the solutions converge to certain coherent structures on the branch but in a weaker norm corresponding to a space $Z,\ X \hookrightarrow Z$, see the discussion at the end of Sect. 2. The methods to uncover the convergence are dynamical in nature and, near a branch of coherent structures and *away* from bifurcation points, can be summarized as follows: one decomposes the solution into a finite dimensional evolution on the manifold of coherent structures and a correction

$$u(t) = T(\omega(t)t)[\phi_{\omega(t)} + u_d(t)] \tag{19}$$

where the parameter $\omega(t)$ is to be chosen later. Then the equation for the correction becomes:

$$\frac{du_d}{dt} = JL_{\omega(t)}u_d + G(\omega(t), u_d(t)) \tag{20}$$

where L_ω is the linearization (with respect to u) of $D_u\mathscr{E}(u) - \omega D_u Q(u)$ at ϕ_ω and G contains only quadratic and higher order terms in u_d. Most of the times equation (20) is be analyzed via a Duhamel formula using the propagator of a *fixed* linearization i.e., $W(t)u_0$ solves:

$$\frac{du}{dt} = JL_{\omega_0}u, \quad u(0) = u_0$$

and

$$u_d(t) = W(t)u_d(0) + \int_0^t W(t-s)[JL_{\omega(s)} - JL_{\omega_0}]u_d(s)ds + \int_0^t W(t-s)G(\omega(s), u_d(s))ds. \tag{21}$$

In this case $\omega(t)$ is chosen such that $u_d(t)$ is always in the invariant subspace of JL_{ω_0} that complements the null space. In the absence of other eigenvalues the invariant space corresponds to the continuous spectrum, and the advantage is that, on this subspace, estimates of type:

$$\|W(t)\|_{Z^* \mapsto Z} \sim |t|^{-\alpha}, \quad \alpha > 0 \tag{22}$$

where Z is a Banach space with $X \hookrightarrow Z$ densely, were already available. The disadvantage is the presence of the linear term in the second integral. In fact, for the particular case of NLS, the linear term lead to restrictions on the nonlinearity to supercritical regimes $p > 4/n$ in (6), see [9–11, 13, 44, 48–50], or, when Stricharz type estimates were used, to critical and supercritical regimes $p \geq 4/n$, see [21, 39, 40]. The results in [29, 30, 32, 33] show that, for small solitary waves in NLS, from estimates (22) one can obtain estimates of the same type for the propagator of the *time dependent* linear operator in (20). Hence one can use:

$$u_d(t) = \tilde{W}(t,0)u_d(0) + \int_0^t \tilde{W}(t,s)G(\omega(s), u_d(s))ds. \tag{23}$$

where $\tilde{W}(t,s)u_0$ solves the non-autonomous equation

$$\frac{du}{dt} = JL_{\omega(t)}u, \qquad u(s) = u_0 \tag{24}$$

and now the $\omega(t)$ is chosen such that $u_d(t)$ is always in the invariant subspace of $JL_{\omega(t)}$ that complements the null space, hence, in the absence of other eigenvalues, on this subspace we have:

$$\|\tilde{W}(t,s)\|_{Z^* \mapsto Z} \sim |t-s|^{-\alpha}, \qquad \alpha > 0 \tag{25}$$

As a result the restriction to critical and supercritical nonlinearities has been lifted.

Essential in this approach is to obtain estimates (25) from (22). While technical in nature this step can be generalized to other equations, see [7], because it relies only on $V(t) = JL_{\omega(t)} - JL_{\omega_0}$ being a small, localized in space (but time dependent and maybe complex valued) scatterer (potential) and on strong dispersion of the linearized equation, i.e.

$$\alpha \geq 1. \tag{26}$$

Smallness is not necessary since orbital stability implies that large deviation in $V(t)$ are only along the orbits of the coherent structures which can be mod out, see [31]. Space localization will always be present when dealing with solitary waves i.e., the scatterer is a power of the solitary wave. The only hypothesis that cannot yet be relaxed is (26), in particular the method is inapplicable in one dimensional NLS.

A much more delicate dynamics occurs near an intersection of stable and unstable manifolds of coherent structures. Section 3.2 shows that existence of such bifurcation points is generally the rule rather than the exception, hence understanding the dynamics around bifurcations is a necessary step in studying asymptotic completeness. Note that finite time behavior of small solitary waves near the bifurcation point discovered first in [27] has been studied for example in [28] via an approximation with a finite dimensional dynamical system.

Unfortunately, the recent progress in asymptotic stability near an orbitally stable coherent state cannot yet determine the full dynamical picture near a bifurcation point. Current results, see [4, 14, 57], rely on a spectral assumption for the linearization JL_ω which fails at the bifurcation point. More precisely, as one approaches the bifurcation point ($\omega \to \omega_*$) from a stable branch, two, purely imaginary and complex conjugate eigenvalues (which can be non-simple) of JL_ω approach zero (this corresponds to one eigenvalue, maybe non-simple, of $L_\omega = D^2\mathscr{E}(\phi_\omega) - \omega D^2Q(\phi_\omega)$ approaching zero, see the discussion in Sect. 3.2). Past the bifurcation point they move from zero back up on the imaginary axis (along on the stable branch) or into a positive and negative eigenvalue (along the unstable branch). The above cited results may be applicable for some $\omega \neq \omega_*$ along the stable branch but not to all. This is because the two eigenvalue which approach zero as $\omega \to \omega_*$ have multiples close to any other eigenvalue, therefore violating the discrete spectrum non-resonance condition required by current results. G. Zhou has done yet unpublished work that may remove the non-resonance condition. Even if this work is vetted the results only say that for each $\omega \neq \omega_*$ along the *stable branch* there is a small ball in the Hilbert space X centered at ϕ_ω such that initial data from the ball asymptotically converge (in the weaker norm) to a coherent state (not necessarily ϕ_ω). However, the radius of this ball is related to the distance d between the smallest eigenvalue (in absolute value) and zero, and goes to zero as ω approaches the bifurcation point because of the $1/d$ numbers of change of variable necessary to bring the system in a normal form that uncovers the *radiation damping mechanism* which leads to the decay of the projection of the solution onto the invariant subspace of this smallest eigenvalue via a resonant interaction with the radiative part corresponding to the projection onto the continuous spectrum. In conclusion, asymptotic stability can be shown only in a conical neighborhood of the stable branch with vertex at the bifurcation point (and the vertex excluded).

What happens with initial data in a ball centered at the bifurcation point which, of course, has relatively large regions not contained in the conical neighborhoods described above? A first step would be to determine the stable invariant manifolds along the unstable branch. This is obtained via implicit function theorem type results from the invariant subspace of the linearization that complements the one corresponding to the positive eigenvalue. Initial data on this manifold will converge to an unstable coherent state, see [53] for a related result. Outside this manifold one can use a spectral decomposition of the dynamics with respect to the linearization at the bifurcation point JL_{ω_*}. Note that the invariant subspace corresponding to the zero eigenvalue contains the kernel of $D^2\mathscr{E}(\phi_\omega) - \omega D^2Q(\phi_\omega)$ which caused the bifurcation to occur in the first place, moreover the projection on this kernel parameterizes the branches near the bifurcation point via the standard Lyapunov-Schmidt decomposition. Since the initial data is away from the stable invariant manifold corresponding to the unstable branch one expects a short time exponential growth of the projection onto the direction of the stable branch. Once this projection becomes dominant a change of variables can be employed in order to use the linearization and associated spectral decomposition at the (time dependent) stable coherent state given (parameterized) by the values of this dominant projection. In this coordinates the techniques discussed in the above paragraph are expected to lead to asymptotic convergence towards a stable coherent state. This is work in progress.

5 Conclusions

While variational methods are inappropriate to study *all coherent states* of a given nonlinear wave equation (1) recent progress in NLS equation shows promise for bifurcation methods. We learned from the example described in Sect. 3.2 that one can start from any known solution of (3) and expect that it is part of a smooth manifold of solutions that can be extended to the boundary of the domain (subset of $X \times \mathbb{R}$)) inside which the linearization of (3), $D^2E(u) - \omega D^2Q(u)$, is Fredholm (a finite dimensional kernel suffices as this is a self-adjoint operator). Such results are the main theorems of the current global bifurcation theory as developed by Rabinowitz, Dancer, Toland, Buffoni and others. However, this is not enough to discover bifurcation points along the manifold or new branches of solutions. But if one has spectral information about the linearization at the starting point and at the end point of the manifold (both on the boundary) then existence of bifurcations can be proven and the branches emerging from them can be studied. Ideally one would want to find all limit points of manifolds of coherent states at the boundary of the Fredholm domain. Now, one can start from any such limit point and use the fact that the manifold approaching it can be continued (via global bifurcation theory) until it reaches another limit point on the same boundary (or the same point in case a loop forms). Loops can sometimes be ruled out via bifurcation in cones type arguments, see [8], and the symmetry or spectral properties of the initial limit point can severely reduce the choices of the end point such that a numerical investigation or a rigorous theorem can determine it as in the example. Once the two end points of the manifold are determined the change in number of negative eigenvalues of the linearized operator at the two ends can tell us the number of eigenvalues crossing zero hence the number and type of bifurcations along the manifold. These bifurcations can now be analyzed via Lyapunov-Schmidt decompositions and normal forms, see for example the singularity theory in [16]. The emerging branches of solutions have also limits at the boundary of the Fredholm domain. The process repeats until all branches of solutions are found.

There are four essential steps in the method summarized above:

(R1) identify the domain in the $X \times \mathbb{R}$ space where the linearized operator of (3) is Fredholm;

(R2) inside this domain show relative compactness of either the set of solutions of (3) or of a map for which the set of fixed points coincides with the set of solutions of (3);

(R3) identify all limit points of solution branches on the boundary of the Fredholm domain;

(R4) find the rules by which the limit points connect via manifolds of solutions inside the domain and characterize the bifurcation points along these manifolds.

(R1) is already done for most wave equations as linearization is used in any analysis or numerical simulation for nonlinear problems. While this method focuses on finding coherent states inside the domain where the linearization is Fredholm, note that outside it one can sometimes show non-existence of coherent states with certain localization properties based on smoothness of the spectral measure of the linearized operator, see [6].

(R2) is a technical step but an essential assumption in global bifurcation theory. It also helps with the (R3) step. For problems with real analytic energy, relative compactness of the solution set of (3) suffices and implies that the only obstacle for unique continuation of any manifold of coherent states (via implicit function theorem) is that it either reached the boundary of the Fredholm domain or a bifurcation (singularity) point. In the latter case, the manifold reemerges on the other side of the bifurcation point because of the structure of zeroes of analytical maps. Compactness and structure of zeroes for analytical maps combine again to prevent the existence of infinitely many singularities in bounded domains. Hence the manifold either reaches the boundary of the Fredholm domain or forms a loop, see [8]. If the energy is non-analytical a stronger form of compactness is required. Equation (3) is transformed into a fixed point problem, for example $\phi = (-\Delta + 1)^{-1}\psi$, $\psi = (1 - E)\phi - V(x)\phi - \gamma|\phi|^p\phi \overset{def}{=} \tilde{F}(\psi, E)$ in the NLS case (7), and the map $\tilde{F} : X^* \mapsto X^*$ (or from Y to Y where $X \hookrightarrow Y \hookrightarrow X^*$) is required to be relatively compact, see [45].

Note that the stronger type of compactness comes for free when $X \hookrightarrow Y$ is compact, which is the case for waves on bounded domains, or when the range of $\tilde{F}$ is made of functions with a prescribed decay at infinity, which is the case for problems with confining potentials, $\lim_{|x|\to\infty} V(x) = \infty$, or with repelling nonlinearities see [24, 25]. For NLS with attractive nonlinearities, a delicate argument involving concentration compactness and non-existence of bifurcations from profiles concentrated at infinity is used in [35] to obtain relative compactness of the set of solutions. It may be adapted for general wave problems since X, Y are usually Sobolev spaces.

(R3) can be split into two parts: first obtain rigorous estimates for coherent structures approaching the boundary of the Fredholm domain and use compactness arguments to identify possible limit points, then use local bifurcation theory from these limit points to identify all nearby branches of solutions. For the first part one can use the energy and charge Q together with the identity $\frac{dE}{d\omega}(\phi_\omega) = \omega\frac{dQ}{d\omega}(\phi_\omega)$ and the equation (3) both valid for coherent states. The goal is to obtain closed differential inequalities for the terms in the energy which can lead estimates in certain limits i.e., as the branch approaches different parts of the boundary of the Fredholm domain. In the NLS example with attractive nonlinearity, the quadratic terms in the energy (4) grow much faster than the superquadratic term corresponding to the nonlinearity provided the H^1 norm blows up and ω remains finite. Equation (7) becomes linear in this limit and it turns out that the limiting equation has no solution, hence the absence of coherent states near this boundary. At the $\omega \to -\infty$ boundary the kinetic and nonlinear terms grow faster than the potential term which leads to the limiting equation (18) via the re-scaling (17). The limiting points are then among the solutions of the limiting equation. In the example we were looking at positive coherent structures (ground states) and the limiting equation has exactly one such solution modulo translations, hence the limiting points described in Fig. 2. In general, one can focus on examples for which the limiting equations are well studied. However if the procedure leads to poorly understood equations this theory is a motivating factor in studying them.

The second part i.e., identify the nearby branches from the limit points, may be solved via standard local bifurcation theory when the linearization at the limiting points (which are solutions of the limiting equation) is Fredholm. However, this is not the case when coherent states approach the part of the boundary where the linearization has zero at an edge of the essential spectrum. This happens for repelling nonlinearities, see [24, 25], for attractive nonlinearities when the linearization of (3) at $u = 0$ has no discrete spectrum and especially for Dirac equation, regardless of nonlinearity, since the linear Dirac operator has essential spectrum everywhere except a bounded interval. In all these cases local bifurcations from the edge of the essential spectrum must be understood in order to find all limit points on this boundary and complete the bifurcation diagram. These are notoriously difficult problems but promising results in this direction are described in [56].

(R4) amounts to grouping the limiting points based on which closed subspace of X they belong to, usually based on symmetry properties such as the subspace of even functions in our example. Since global bifurcation theory applies in any Banach space each limit points must connect to another one in the same group. If there are more than two in a subgroup then not only numerical simulations can help but also rigorous arguments combining the mismatch in the negative eigenvalues of the linearization at the two endpoints with the type of bifurcations supported by the eigenspaces corresponding to the eigenvalues that cross zero as one moves from one limiting point to the other. Note that problems invariant under finite and continuous groups of Euclidian symmetries have non-simple eigenvalues in the spectrum of the linearized operator. If they cross zero a classification of bifurcations induced by them is necessary before we can proceed to identify all coherent states.

The bifurcation method relies and provides information on the spectrum of the linearized operator along the manifolds of coherent states. There is already a rich theory that uses the spectral information to determine the stability of coherent states and the long time dynamics of nearby solutions. While a resolution of the *Asymptotic Completeness Conjecture*, see Sect. 1, still seems far away, it appears that a systematic study of all coherent states supported by nonlinear wave equations, their bifurcation points, their stability and the nearby dynamics is within reach.

References

1. R. A. Adams, *Sobolev Spaces.* Academic Press, New York, 1975.
2. A. Ambrosetti, M. Badiale, S. Cingolani, “Semiclassical states of nonlinear Schrödinger equations with bounded potentials”, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl. 7 (1996), no. 3, 155–160.
3. W.H. Aschbacher, J. Fröhlich, G.M. Graf, K. Schnee, and M. Troyer, “Symmetry breaking regime in the nonlinear hartree equation”, J. Math. Phys. **43**, 3879–3891 (2002).
4. D. Bambusi, S. Cuccagna, “On dispersion of small energy solutions to the nonlinear Klein Gordon equation with a potential”, Amer. J. Math. **133** (2011), no. 5, 1421–1468.
5. V. Benci, D. Fortunato, *Variational methods in nonlinear field equations. Solitary waves, hylomorphic solitons and vortices.* Springer Monographs in Mathematics. Springer Cham Heidelberg New York Dordrecht London, 2014.
6. H. Berestycki, P.-L. Lion, “Nonlinear scalar field equations”, Arch. Ration. Mech. Anal. **82** (1983) 313–375.
7. N. Boussaid, E. Kirr, “Asymptotic stability of ground states in Dirac equation”, in preparation.
8. B. Buffoni, J. Toland, *Analytic theory of global bifurcation. An introduction.* Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2003.
9. V. S. Buslaev, G. S. Perelman, “Scattering for the nonlinear Schrödinger equation: states that are close to a soliton”. St. Petersburg Math. J. 4 (1993), no. 6, 1111–1142.
10. V. S. Buslaev, G. S. Perelman, “On the stability of solitary waves for nonlinear Schrödinger equations”. Nonlinear evolution equations, 75–98, Amer. Math. Soc. Transl. Ser. 2, 164, Amer. Math. Soc., Providence, RI, 1995.
11. V. S. Buslaev, C. Sulem, “On asymptotic stability of solitary waves for nonlinear Schrödinger equations”. Ann. Inst. H. Poincaré Anal. Non Linéaire 20 (2003), no. 3, 419–475.
12. T. Cazenave, *Semilinear Schrödinger equations*, volume 10 of *Courant Lecture Notes in Mathematics* (New York University Courant Institute of Mathematical Sciences, New York, 2003).
13. S. Cuccagna, “Stabilization of solutions to nonlinear Schrödinger equations”, Comm. Pure Appl. Math. **54** (2001), 1110–1145.
14. S. Cuccagna, T. Mizumachi, “On asymptotic stability in energy space of ground states for nonlinear Schrödinger equations”, Comm. Math. Phys. **284** (2008), no. 1, 51–77.
15. A. Floer and A. Weinstein, “Nonspreading wave packets for the cubic Schrödinger equation with a bounded potential”, J. Funct. Anal. **69**, 397–408 (1986).
16. M. Golubitsky, I. Stewart, D. G. Schaeffer, *Singularities and groups in bifurcation theory,* Vol. II. Applied Mathematical Sciences, 69, Springer-Verlag, New York, 1988.
17. M. Grillakis, “Linearized instability for nonlinear Schrödinger and Klein–Gordon equations”, Comm. Pure Appl. Math. **41**, 747–774 (1988).
18. M. Grillakis, J. Shatah, and W. Strauss, “Stability theory of solitary waves in the presence of symmetry. I,”, J. Funct. Anal. **74** (1987), no. 1, 160–197.
19. M. Grillakis, J. Shatah, W. Strauss, “Stability theory of solitary waves in the presence of symmetry. II”, J. Funct. Anal. **94** (1990), no. 2, 308–348.
20. Y. Guo, R. Seiringer, “On the mass concentration for Bose-Einstein condensates with attractive interactions.” Lett. Math. Phys. 104 (2014), no. 2, 141–156.
21. S. Gustafson, K. Nakanishi, T.-P. Tsai, “Asymptotic stability and completeness in the energy space for nonlinear Schrödinger equations with small solitary waves”. Int. Math. Res. Not. 2004, no. 66, 3559–3584.
22. R. K. Jackson, communication at SIAM Conference on Nonlinear Waves, Philadelphia, Aug. 2010.
23. R. K. Jackson, M. I. Weinstein, “Geometric analysis of bifurcation and symmetry breaking in a Gross-Pitaevskii equation.” J. Statist. Phys. 116 (2004), no. 1–4, 881–905.
24. H. Jeanjean, M. Lucia and C. Stuart, “Branches of solutions to semilinear elliptic equations on $\mathbb{R}^N$”, Math. Z. **230**, 79–105 (1999).

25. H. Jeanjean, M. Lucia and C. Stuart, " The branche of positive solutions to a semilinear elliptic equation on $\mathbb{R}^N$", Rend. Sem. Mat. Univ. Padova, **101**, 229–262 (1999).
26. P.G. Kevrekidis, R. Carretero-González, D.J. Frantzeskakis, "Vortices in Bose-Einstein Condensates: (Super)fluids with a twist", SIAM Dynamical Systems Magazine, October, 2011.
27. E.W. Kirr, P.G. Kevrekidis, E. Shlizerman, and M.I. Weinstein, "Symmetry-breaking bifurcation in nonlinear Schrödinger/Gross–Pitaevskii equations", SIAM J. Math. Anal. **40**, 56–604 (2008).
28. E. Kirr, P.G. Kevrekidis, D. Pelinovsky, "Symmetry-breaking bifurcation in the nonlinear Schrödinger equation with symmetric potentials", Commun. Math. Phys. 308 (2011), 795–844
29. E. Kirr, A. Zarnescu, *On the asymptotic stability of bound states in 2D cubic Schrödinger equation* Comm. Math. Phys. **272** (2007), no. 2, 443–468.
30. E. Kirr, A. Zarnescu, *Asymptotic stability of ground states in 2D nonlinear Schrödinger equation including subcritical cases,* J. Differential Equations **247** (2009), no. 3, 710–735.
31. E. Kirr, A. Zarnescu, *Asymptotic stability of large ground states in nonlinear Schrödinger equation*, in preparation.
32. E. Kirr and Ö. Mızrak, "Asymptotic stability of ground states in 3d nonlinear Schrödinger equation including subcritical cases", J. Funct. Anal. **257**, 3691–3747 (2009).
33. E. Kirr and Ö Mızrak, " On the stability of ground states in 4D and 5D nonlinear Schrödinger equation including subcritical cases" submitted to Int. Math. Res. Not. available online at: http://arxiv.org/abs/0906.3732
34. E. Kirr, P.G. Keverekidis and V. Natarajan, "Bifurcations of large ground states in one dimensional nonlinear Schrödinger equation", in preparation.
35. E. Kirr and V. Natarajan, "The global bifurcation picture for coherent states in nonlinear Schrödinger equation", in preparation.
36. P.-L. Lions, "The concentration-compactness principle in the calculus of Variations. The locally compact case. I." Ann. Inst. H. Poincare Anal. Non Lineaire **1** (1984), 109–145.
37. P.-L. Lions, "The concentration-compactness principle in the calculus of Variations. The locally compact case. II." Ann. Inst. H. Poincare Anal. Non Lineaire **1** (1984), 223–283.
38. J.L. Marzuola and M.I. Weinstein, "Long time dynamics near the symmetry breaking bifurcation for nonlinear Schrödinger/Gross–Pitaevskii equations", DCDS-A, to be published (2010).
39. T. Mizumachi, "Asymptotic stability of small solitary waves to 1D nonlinear Schrödinger equations with potential", J. Math. Kyoto Univ. **48** (2008), 471–497.
40. T. Mizumachi, "Asymptotic stability of small solitons for 2D Nonlinear Schrödinger equations with potential", J. Math. Kyoto Univ. **47** (2007), no. 3, 599–620.
41. L. Nirenberg, *Topics in nonlinear functional analysis*, Courant Lecture Notes **6** (New York, 2001).
42. E. S. Noussair, S. Yan, "On positive multipeak solutions of a nonlinear elliptic problem." J. London Math. Soc. (2) 62 (2000), no. 1, 213–227.
43. Y.-G. Oh, "On positive multi-lump bound states of nonlinear Schrödinger equations under multiple well potential", Comm. Math. Phys. 131 (1990), no. 2, 223–253.
44. C.A. Pillet, C.E. Wayne, "Invariant manifolds for a class of dispersive, Hamiltonian, partial differential equations", J. Diff. Eqs. **141**, 310–326 (1997).
45. P. H. Rabinowitz, "Some global results for nonlinear eigenvalue problems," J. Functional Anal. 7 (1971), 487–513.
46. M. Reed and B. Simon, *Methods of Modern Mathematical Physics. Analysis of Operators.* Volume IV. Academic Press San Diego New York Boston London Sydney Tokyo Toronto, 1972.
47. H.A. Rose and M.I. Weinstein, "On the bound states of the nonlinear Schrödinger equation with a linear potential, Physica D **30**, 207–218 (1988).
48. I. M. Sigal, G. Zhou, "Asymptotic stability of nonlinear Schrödinger equations with potential". Rev. Math. Phys. 17 (2005), no. 10, 1143–1207.
49. A. Soffer and M.I. Weinstein, "Multichannel nonlinear scattering for nonintegrable equations", Comm. Math. Phys. **133**, 119–146 (1990).

50. A. Soffer and M.I. Weinstein, “Multichannel nonlinear scattering for nonintegrable equations. II. The case of anisotropic potentials and data” J. Diff. Eqs. **98**, 376–390 (1992).
51. A. Soffer, M. I. Weinstein, *Selection of the ground state for nonlinear Schroedinger equations*, Rev. Math. Phys. 16 (2004), no. 8, 977–1071.
52. T.-P. Tsai, H.-T. Yau, Horng-Tzer *Relaxation of excited states in nonlinear Schrödinger equations*, Int. Math. Res. Not. **31** (2002), 1629–1673.
53. T.-P. Tsai, H.-T. Yau, *Stable directions for excited states of nonlinear Schrödinger equations.* Comm. Partial Differential Equations **27** (2002), no. 11–12, 2363–2402.
54. T.-P. Tsai, H.-T. Yau, *Classification of asymptotic profiles for nonlinear Schrödinger equations with small initial data.* Adv. Theor. Math. Phys. **6** (2002), no. 1, 107–139.
55. M.I. Weinstein, “Lyapunov stability of ground states of nonlinear dispersive evolution equations”, Comm. Pure Appl. Math. **39**, 51–68 (1986).
56. M.I. Weinstein, “Localized States and Dynamics in the Nonlinear Schrödinger/Gross-Pitaevskii Equation”, Frontiers of Applied Dynamical Systems: Reviews and Tutorials, vol. **3** (2015), 41–79.
57. G. Zhou, “Perturbation expansion and Nth order Fermi golden rule of the nonlinear Schrödinger equations”, J. Math. Phys. **48** (2007), no. 5, 053509–053532.

About Non Linear Stabilization for Scalar Hyperbolic Problems

Rémi Abgrall

Abstract This paper deals with the numerical approximation of linear and non linear hyperbolic problems. We are mostly interested in the development of parameter free methods that satisfy a local maximum principle. We focus on the scalar case, but extensions to systems are relatively straightforward when these techniques are combined with the ideas contained in Abgrall (J. Comput. Phys., 214(2):773–808, 2006). In a first step, we precise the context, give conditions that guaranty that, under standard stability assumptions, the scheme will converge to weak solutions. In a second step, we provide conditions that guaranty an arbitrary order of accuracy. Then we provide several examples of such schemes and discuss in some details two versions. Numerical results support correctly our initial requirements: the schemes are accurate and satisfy a local maximum principle, even in the case of non smooth solutions.

1 Introduction

In this paper, we are interested in the numerical solution of steady scalar hyperbolic equations. It is well known that the equations admit discontinuous solutions that are only bounded in L^∞, and belongs to L^1. We are particularly interested in the piecewise smooth solutions. Our focus is on methods that use unstructured conformal meshes with weak Dirichlet boundary conditions. These methods, as well as any of the methods that are devoted to the solution of these non linear problems must incorporate, for stability reasons, some dissipation mechanism, otherwise wild oscillations may develop. There are many classes of high order methods, and in this paper our focus is on the study of particular class called residual distribution schemes. These methods can be seen as some generalizations of classical finite element methods using continuous methods with stabilisation (such as the SUPG method [16]), but some variants allow to have a genuinely

R. Abgrall (✉)
Institute of Mathematics, University of Zürich, Winterthurerstrasse 190,
CH 8057 Zurich, Switzerland
e-mail: remi.abgrall@math.uzh.ch

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_4

non linear dissipation mechanism for which one can guaranty L^∞ stability bounds. Unfortunately, straightforward L^∞ stability procedure may lead to methods that admits spurious modes, in some circumstances. The main issue of this paper is two describe two ways of removing these spurious modes while keeping the L^∞ stability property, at least at the experimental level. One such technique is already known, see [1] for example, the second one is new. Having two methods for the same purpose is, in our opinion, a good thing because it can allow for additional flexibility.

In the following, we focus on steady problems, and to make things simpler, we focus on the scalar problem:

$$\text{div}\mathbf{f}(u) = 0 \tag{1a}$$

subject to

$$\min(\nabla_u f(u) \cdot \mathbf{n}(x), 0)(u - g) = 0 \text{ on } \partial\Omega \tag{1b}$$

In (1b), $\mathbf{n}(x)$ is the outward unit vector at $x \in \partial\Omega$ (thus we assume enough regularity for Ω). We will assume that Ω is bounded for technical reasons only. Extensions to the system case can be found in [5] for the pure hyperbolic case and [3, 4] for the scalar convection diffusion problem and the Navier Stokes equations.

Here the notations are standard: g is a regular enough function, we assume that Ω has a polyhedric boundary, and moreover $\Omega_h = \Omega$ for the chosen family of triangulations $\mathscr{T}_h$ in order to simplify. These assumptions are by no mean essential. We denote by $\mathscr{E}_h$ the set of edges/faces of $\mathscr{T}_h$ that are contained in $\partial\Omega$, and $\mathscr{K}$ stands either for an element K or a face/edge e.

In the finite element setting, there exists several variational formulations of this class of problems. The classical ones can be defined in three steps. We are given a family of meshes denoted by $(\mathscr{T}_h)_{h\in\mathscr{H}}$. These meshes are made of elements denoted generically by K. The parameter h, as usual, denotes the maximum of the diameters of K, $K \in \mathscr{T}_h$. The meshes can be geometrically conformal or not. Then we need to define the trial functions space, denoted by U_h and a test functions space V_h. The last step is to define a bi-linear form a on $U_h \times V_h$, as well as form ℓ defined on V_h. As usual, we assume that the spaces U_h and V_h encode some of the boundary conditions, while the others are encoded in ℓ. The problem is to find $u_h \in U_h$ such that a for any $v_h \in V_h$, we have

$$a(u_h, v_h) = \ell(v_h).$$

The ideal scheme would certainly be the Galerkin method, where the variational formulation is defined by: if $\hat{\mathbf{f}}$ is a consistent upwind numerical flux, we define a_{Gal} and ℓ for the variational formulation

$$\begin{aligned} a_{Gal}(u_h, v_h) &= -\int_\Omega \nabla v_h \cdot \mathbf{f}(u_h) + \sum_{e\in\mathscr{E}_h} \int_e v_h(\hat{\mathbf{f}}_\mathbf{n}(g, u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}) \\ \ell(v^h) &= \int_\Omega f v_h. \end{aligned} \tag{2}$$

They are defined for $u_h, v_h) \in U_h \times U_h$ where

$$U_h = U_h^G := \{u_h \in H^1(\Omega), \forall K \in \mathscr{T}_h, u_{h|K} \in \mathbb{P}^r(K)\} \cap C^0(\overline{\Omega}).$$

This method can be shown (on linear problems) to be formally accurate (i.e. of order $r+1$), but if the boundary conditions are not set in a very precise way (see [8]), it is also known to be widely unstable. In any case, the nonlinear case is not stable in the case of discontinuous solutions, as those we are expecting here. So the game has been since several decades to find ways to stabilize this operator while keeping its formal accuracy.

A first example is given by the streamline diffusion method [16, 17] for which there are two possible interpretations. In the first one, we consider a Petrov Galerkin formulation, .i.e we take $u_h \in U_h = U_h^G$ as for (2), but $v_h \in V_h$ where

$$V_h = V_h^S := \{v_h \in L^2(\Omega), \forall K \in \mathscr{T}_h, \exists w_h \in U_h, v_h = w_h + h_K\tau_K\nabla_u\mathbf{f}(u_h)\nabla w_h\}.$$

The formulation uses

$$\begin{aligned} a_{\text{SUPG1}}(u_h, v_h) &= -\int_\Omega v_h \cdot \text{div}\mathbf{f}(u_h) + \sum_{e\in\mathscr{E}_h} \int_e v_h(\hat{\mathbf{f}}_\mathbf{n}(g, u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}) \\ \ell(v_h) &= \int_\Omega f v_h. \end{aligned} \tag{3a}$$

The second interpretation is to take $V_h = U_h^G$ and use, instead of a_{SUPG1} the form a_{SUP2} defined by

$$\begin{aligned} a_{\text{SUPG2}}(u_h, v_h) &= -\int_\Omega \nabla v_h \cdot \mathbf{f}(u_h) + \sum_K h_K \int_K \big(\nabla_u f(u_h)\nabla v_h\big)\tau_K\big(\nabla_u f(u_h)\nabla u_h\big) \\ &\quad + \sum_{e\in\mathscr{E}_h} \int_e v_h(\hat{\mathbf{f}}_\mathbf{n}(g, u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}) \\ \ell(v_h) &= \int_\Omega f v_h. \end{aligned} \tag{3b}$$

This can be seen as a Galerkin approximation of a modified equation, namely

$$\text{div}\mathbf{f}(u) - \text{div}\Big(h\tau\,\text{div}\mathbf{f}(u)\Big) = 0 \tag{3c}$$

In (3), the parameters τ_K are positive functions (typically constant per element) and in (3c) the function τ is defined by its restrictions on each element, as well as h.

We can play further with the trial and test spaces. If one removes the continuity assumption, then we have a discontinuous Galerkin formulation, i.e. $U_h = V_h$ with

$$U_h = U_h^{DG} := \{u_h \in L^2(\Omega), \forall K \in \mathscr{T}_h, u_{h|K} \in \mathbb{P}^r(K)\}$$

and, for $(u_h, v_h) \in U_h^{DG} \times U_h^{DG}$,

$$\begin{aligned} a(u_h, v_h) &= \sum_{K\in\mathscr{T}_h} \Big(-\int_K \nabla v_h \cdot \mathbf{f}(u^h) + \int_{\partial K} v_h \hat{\mathbf{f}}_{\mathbf{n}}\big((u_h)_{|K}, (u_h)_{|K^-}\big)\Big) \\ \ell(v_h) &= \sum_{K\in\mathscr{T}_h} \int_K f v_h \end{aligned} \tag{4a}$$

where K^- denotes generically the element(s) that are on the other side of the faces of ∂K. Another formulation is, with the same ℓ,

$$\begin{aligned} a(u_h, v_h) = &\sum_{K\in\mathscr{T}_h} \Big(-\int_K \nabla v_h \cdot \mathbf{f}(u^h) + \int_{\partial K} v_h \hat{\mathbf{f}}_{\mathbf{n}}\big((u_h)_{|K}, (u_h)_{|K^-}\big)\Big) \\ &+ \sum_K h_K \int_K \big(\nabla_u f(u_h) \nabla v_h\big) \tau_K \big(\nabla_u f(u_h) \nabla u_h\big) \end{aligned} \tag{4b}$$

In (4), the Dirichlet boundary conditions are set weakly by imposing $u_h = g$ on the parts of ∂K which belongs to inflow part of $\partial\Omega$ as for (3).

Another example of stable method was initially described in [10]. The idea is to stabilize the Galerkin operator (2), not by a streamline operator as for the SUPG method (3), but by a jump operator on the internal edges/faces only: here $(u_h, v_h) \in U_h^G \times U_h^G$, and

$$a_{Burman}(u_h, v_h) = a_{Gal}(u_h, v_h) + \sum_{e\in \mathrm{e}_h} \Gamma_e h_e^2 \int_e [\nabla u_h][\nabla v_h]. \tag{5}$$

In (5), for any function φ which admit traces one each faces of K, $[\varphi] = \varphi_{K^+} - \varphi_{K^-}$ where K^+ and K^- are the two elements that share the face e (remember we assume that the mesh is conformal), h_e is the measure of e and Γ is a parameter that has the dimension of $\nabla_u \mathbf{f}(u)$.

The space U_h and V_h can be independently chosen, as well as a and ℓ, provided the variational problem is consistent with the problem (1), and of course the numerical method is stable. Formal accuracy is obtained via the choice the polynomial degree r, and effective accuracy is related to the stability of the scheme in suitable norm. Hence a natural question is: can we define U_h, V_h and the forms a and ℓ such that in addition with consistency and accuracy, we can also have non oscillatory properties. In the case of the streamline methods, this last property is obtained by modifying the formulation by adding a dissipation operator which is parameter dependent. In the case of the Discontinuous Galerkin method, this property is obtained via a proper choice of the arguments in $\hat{\mathbf{f}}_{\mathbf{n}}$, see [11, 12]. We note that only the averages in K are controlled. In both cases this stability property is obtained by introducing some genuine non linearity in the scheme, i.e. even if (1) is a linear problem, the scheme will be non linear.

In this paper, we show that, by introducing a solution-dependent operator χ from $U_h \cap C^0(\Omega)$ to $L^2(\Omega)$, the variational problem with a defined by

$$\begin{aligned} a(u_h, v_h) &= \sum_{K\in\mathscr{T}^h} \int_K \chi_u^h(v_h)\mathrm{div}\mathbf{f}(u^h) + \sum_{e\in\mathscr{E}} \int_e v_h(\hat{\mathbf{f}}_{\mathbf{n}}(g, u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}) \\ \ell(v^h) &= \sum_{K\in\mathscr{T}^h} \int_K \chi_u^h(v_h)f \end{aligned} \tag{6}$$

enables to get all the properties. The rest of this paper is organized as follow: inspired by a rewriting of (3), we introduce the residual distribution schemes. We provide a simple criteria which guaranties a Lax-Wendroff type theorem, provide a simple criteria that guaranties formal accuracy, show how the choice of norms guaranty the effective accuracy, and provide several examples of schemes. One of them is new.

2 Formulation of Residual Distribution Schemes

These schemes have original been introduced by P.L. Roe in [21] in one dimension, and [22] in the multidimensional case. As we see, there are many common points with the streamline method, the difference is that we try to combine ideas from the finite element community and from the finite volume one. The first scheme of this kind was probably designed by R. Ni [20] where he introduces a particular version of the Lax-Wendroff scheme.

2.1 Definition, Connection to Finite Element Methods

In what follows, $\mathscr{K}$ represents either an internal element or a face.

We make the standard remark that, for any internal degree of freedom σ, if φ_σ is the Lagrange basis function associated to σ, (3b) can be written as:

$$a_{\text{SUPG2}}(u_h,\varphi_\sigma)=\sum_K\bigg(-\int_K\nabla\varphi_\sigma\cdot\mathbf{f}(u_h)+h_K\int_K\big(\nabla_u\mathbf{f}(u_h)\nabla\varphi_\sigma\big)\tau_K\big(\nabla_u\mathbf{f}(u_h)\nabla u_h\big)\bigg)$$
$$+\sum_{e\in\mathscr{E}_h}\int_e\varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}).$$

Since the support of φ_σ is made of all the elements K that share σ, we have for any degree of freedom σ:

$$a_{SUPG2}(u_h,\varphi_\sigma)=\sum_{K\ni\sigma}\bigg(-\int_K\nabla\varphi_\sigma\cdot\mathbf{f}(u_h)+h_K\int_K\big(\nabla_u\mathbf{f}(u_h)\nabla\varphi_\sigma\big)\tau_K\big(\nabla_u\mathbf{f}(u_h)\nabla u_h\big)\bigg)$$
$$+\sum_{e\in\mathscr{E},\sigma\in e}\int_e\varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n})$$

and notice that

1. for any K,

$$\sum_{\sigma\in K}\bigg(\int_K\nabla\varphi_\sigma\cdot\mathbf{f}(u_h)+h_K\int_K\big(\nabla_u\mathbf{f}(u_h)\nabla\varphi_\sigma\big)\tau_K\big(\nabla_u\mathbf{f}(u_h)\nabla u_h\big)\bigg)=\int_{\partial K}\mathbf{f}(u_h)\cdot\mathbf{n},$$

2. for any $e\in\mathscr{E}_h$,

$$\sum_{\sigma\in e}\int_e\varphi_\sigma(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n})=\int_e(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}).$$

This is true because $\sum_{\sigma\in K}\varphi_\sigma(x)=1$ and thus $\sum_{\sigma\in K}\nabla\varphi_\sigma(x)=0$ for all $x\in\mathscr{K}$.

Let us notice that the discontinuous Galerkin schemes can also fit in a similar framework. Looking back at (4a), we see that we can introduce for the degree of freedom $\sigma\in K$ the residual

$$\Phi_\sigma^K(u_h)=-\int_K\nabla\varphi_\sigma\cdot\mathbf{f}(u^h)+\int_{\partial K}\varphi_\sigma\hat{\mathbf{f}}_{\mathbf{n}}\big((u_h)_{|K},(u_h)_{|K^-}\big).\tag{7a}$$

Then, (4a) is nothing more that

$$a(u_h,\varphi_\sigma)=\sum_{\sigma\in K}\Phi_\sigma^K(u_h).\tag{7b}$$

We also have

$$\sum_{\sigma\in K}\Phi_\sigma^K(u_h)=\int_{\partial K}\hat{\mathbf{f}}_{\mathbf{n}}\big((u_h)_{|K},(u_h)_{|K^-}\big)\tag{7c}$$

where, again, $\hat{\mathbf{f}}_{\mathbf{n}}$ is a consistent flux. This has been exploited in [2, 7].

This set of elementary remarks shows that most if not all known numerical schemes for solving (1) can be set in the Residual distribution setting: Given a tessellation of $\Omega = \cup_{K\in\mathscr{T}_h} K$, we consider the approximation spaces

$$U_h = \bigoplus_{K\in\mathscr{T}_h} \mathbb{P}^r(K)$$

or

$$U_h = \left[\bigoplus_{K\in\mathscr{T}_h} \mathbb{P}^r(K)\right] \cap C^0(\Omega) = U_h^G,$$

depending whether we are looking for a global continuous approximation or a piecewise continuous one.[1] The elements of $\mathbb{P}^r(K)$ are defined by a set of unisolvent degrees of freedom, and we denote by Σ the set of all degrees of freedom defining the elements of U_h. Throughout the paper, we consider Lagrange approximation, but more general approximation sets can be used, see [9] for example. *This means that $U_h = U_h^G = V_h$ throughout the paper.*

A residual distribution scheme is defined, considering any degree of freedom σ, by the sub-residuals that are "sent" to σ by the elements K (resp. a boundary edge e) that share this degree of freedom. We denote them by $\Phi_\sigma^K(u^h_{|K})$ (resp. $\Phi_\sigma^e(u^h_{|e})$). We look for $u_h \in U_h$ such that, for any internal degree of freedom σ,

$$\sum_{K\ni\sigma} \Phi_\sigma^K(u^h_{|K}) = 0, \tag{8a}$$

and for any degree of freedom on the boundary,

$$\sum_{K\ni\sigma} \Phi_\sigma^K(u^h_{|K}) + \sum_{e\ni\sigma} \Phi_\sigma^e(u^h_{|e}) = 0. \tag{8b}$$

We assume that the following structure condition holds true:

$$\sum_{\sigma\in K} \Phi_\sigma^K(u^h_{|K}) = \int_{\partial K} \hat{\mathbf{f}}_{\mathbf{n}}(u^h_K, u^h_{K^-}) \tag{9a}$$

$$\sum_{\sigma\in e} \Phi_\sigma^e(u^h_{|K}) = \int_e (\hat{\mathbf{f}}_{\mathbf{n}}\big(g, u_h) - \mathbf{f}(u_h)\cdot\mathbf{n}\big). \tag{9b}$$

We see that the SUPG method (3) and the Burman method [10] are particular cases of such scheme. There is a lot of freedom in defining the sub-residuals

[1] More complex situation can easily been imagined, such as global continous on Ω_1 and possibly discontinuous on Ω_2 with $\Omega_1 \cup \Omega_2 = \Omega$ and $\Omega_1 \cap \Omega_2$ of empty interior.

$\Phi_\sigma^K(u^h_{|K})$ and $\Phi_\sigma^e(u^h_e)$, we will show how we can take advantage of this freedom to achieve our goal. Note that in the definition of the sub-residual, we have implicitly assumed that only the degrees of freedom with K or e are necessary to define these quantities: the stencil of the method is the most possible compact which is a good point for the parallelization of the method.

Another example of sub-residual are the Galerkin residuals defined by: on the element K.[2]

$$\Phi_\sigma^{G,K}=\int_K \varphi_\sigma \mathrm{div}\mathbf{f}(u^h)=-\int_K \nabla\varphi_\sigma\cdot\mathbf{f}(u^h)+\int_{\partial K}\varphi_\sigma\hat{\mathbf{f}}_{\mathbf{n}}(u^h_K,u^h_{K^-}),\tag{10a}$$

and on the boundary face e:

$$\Phi_\sigma^{G,e}=\int_e \varphi_\sigma\big(\hat{\mathbf{f}}_{\mathbf{n}}(g,u_h)-\mathbf{f}(u_h)\cdot\mathbf{n}\big)\tag{10b}$$

We see that both $\{\Phi_\sigma^{G,K}\}_{\sigma\in K}$ and $\{\Phi_\sigma^{G,e}\}_{\sigma\in e}$ satisfy (9) with the same value of the total residual. Unfortunately, the scheme (8) with the Galerkin residual (10) is widely unstable in the case of continuous elements.

2.2 Structure Conditions

For any w^h (not necessarily a solution of (8) if it exists), and any test function v^h, we have (setting $v^h_\sigma=v^h(\sigma)$):

$$\begin{aligned}
\sum_{\sigma\notin\partial\Omega} v_\sigma^h\Big(\sum_{\mathcal{T}_h\ni K\ni\sigma}\Phi_\sigma^K(w^h_{|K})\Big)&+\sum_{\sigma\in\partial\Omega} v_\sigma^h\Big(\sum_{\mathcal{T}_h\ni K\ni\sigma}\Phi_\sigma^K(w^h_{|K})+\sum_{\mathcal{E}_h\ni e\ni\sigma}\Phi_\sigma^e(w^h_{|e})\Big)\\
&=\sum_{K\in\mathcal{T}_h}\Big(\sum_{\sigma\in K} v_\sigma^h\Phi_\sigma^K(w^h_{|K})\Big)+\sum_{e\in\mathcal{E}_h}\Big(\sum_{\sigma\in e} v_\sigma^h\Phi_\sigma^e(w^h_{|K})\Big)\\
&=\sum_{K\in\mathcal{T}_h}\Big(-\int_K\nabla v^h\cdot\mathbf{f}(u^h)+\int_{\partial K}v^h\hat{\mathbf{f}}_{\mathbf{n}}(u^h_K,u^h_{K^-})\Big)\\
&\quad+\sum_{K\in\mathcal{T}_h}\sum_{\sigma\in K}v_\sigma^h\big(\Phi_\sigma^K(w^h_{|K})-\Phi_\sigma^{G,K}(w^h_{|K})\big)\\
&\quad+\sum_{e\subset\partial\Omega,e\in\mathcal{E}_h}\sum_{\sigma\in e}v_\sigma^h\big(\Phi_\sigma^e(w^h_{|K})-\Phi_\sigma^{G,e}(w^h_{|K})\big)
\end{aligned}\tag{11}$$

[2] Of course, in the case of discontinuous approximation, this is nothing more that DG. Since we have a unified presentation, we need to introduce this.

thanks to (9).[3] In (11), we have used the following implicit convention: On the boundary edges, $u_{K^-} = g$ in order to weakly impose the boundary conditions. Then, since

$$\sum_{\sigma\in\mathscr{K}} \Big(\Phi_\sigma^K(w^h_{|K}) - \Phi_\sigma^{G,K}(w^h_{|K}) \Big) = 0,$$

(11) becomes, denoting by n_K and n_e the number of degree of freedom in K and e, with the convention that $w^h_{K-} = g$ on the boundary of Ω

$$\begin{aligned}\sum_{\sigma\in\Omega} v^h_\sigma \Big(\sum_{\mathscr{K}\ni\sigma} \Phi_\sigma^{\mathscr{K}}(w^h_{|K}) \Big) = \sum_{K\in\mathscr{T}_h} \Big(-\int_K \nabla v^h \cdot \mathbf{f}(u^h) + \int_{\partial K} v^h \hat{\mathbf{f}}_{\mathbf{n}}(u^h_K, u^h_{K}-) \\ + \sum_{K\in\mathscr{T}_h} \frac{1}{n_K} \sum_{\sigma,\sigma'\in K} (v^h_\sigma - v^h_{\sigma'}) \big(\Phi_\sigma^K(w^h_{|K}) - \Phi_\sigma^{G,K}(w^h_{|K}) \big) \\ + \sum_{e\subset\partial\Omega} \frac{1}{n_e} \sum_{\sigma,\sigma'\in e} (v^h_\sigma - v^h_{\sigma'}) \big(\Phi_\sigma^e(w^h_{|e}) - \Phi_\sigma^{G,e}(w^h_{|e}) \big)\end{aligned} \tag{12}$$

This relation is fundamental in our analysis.

2.2.1 Conservation

In [6], we prove the following result:

Theorem 1 *Assume the family of meshes $\mathscr{T} = (\mathscr{T}_h)_{h\in\mathscr{H}}$ is shape regular. We assume that the residuals $\{\Phi_\sigma^{\mathscr{K}}\}_{\sigma\in\mathscr{K}}$, for $\mathscr{K}$ an element or a boundary element of $\mathscr{T}_h$, satisfy:*

1. *For any $M \in \mathbb{R}^+$, there exists a constant C which depends only on the family of meshes $\mathscr{T}_h$ and M such that for any $u_h \in U_h$ with $||u^h||_\infty \leq C(M)$, then*

$$|\,|\Phi_\sigma^{\mathscr{K}}({u^h}_{|\mathscr{K}})\,|\,| \leq C(M) \sum_{\sigma,\sigma'\in\mathscr{K}} |u^h_\sigma - u^h_{\sigma'}|$$

2. *They satisfy the conservation property (9).*

Then if there exists a constant C_{max} such that the solutions of the scheme (8) satisfy $||u^h||_\infty \leq C_{max}$ and a function $v \in L^2(\Omega)$ such that $(u^h)_h$ (or at least a subsequence) converges to v in $L^2(\Omega)$, then v is a weak solution of (1)

Proof The proof can be found in [6], it uses (12) and some adaptation of the ideas of [19].

[3] $\mathscr{K}$ represents either an internal element or a face.

We can also state similar conditions for entropy inequalities:

Proposition 1 *Let* $(U, \mathbf{G})$ *be an couple entropy-flux for (1) and* $\hat{\mathbf{G}}_{\mathbf{n}}$ *an upwind numerical entropy flux consistent with* $\mathbf{G}\cdot\mathbf{n}$. *Assume that the residuals satisfy: for any element K,*

$$\sum_{\sigma\in K} U(u_\sigma)\cdot \Phi_\sigma^K \leq \int_{\partial K} \mathbf{G}(u^h_{|\mathscr{K}})\cdot\mathbf{n} \tag{13a}$$

and for any boundary edge e,

$$\sum_{\sigma\in e} U(u_\sigma)\cdot \Phi_\sigma^e \leq \int_e \big(\hat{\mathbf{G}}_{\mathbf{n}}(u^h_{|e}, g) - \mathbf{G}(u^h_{|K})\cdot\mathbf{n}\big). \tag{13b}$$

Then, under the assumptions of the theorem 1, the limit weak solution also satisfies the following entropy inequality: for any $\varphi\in C^1(\Omega)$, $\varphi\geq 0$,

$$-\int_\Omega \nabla\varphi\cdot\mathbf{G}(u) + \int_{\partial\Omega}\varphi\; \hat{\mathbf{G}}_{\mathbf{n}}(u, g)\leq 0.$$

Proof The proof is similar to that of theorem 1.

2.2.2 Accuracy

In most cases, assuming a smooth solution of (1), the formal accuracy analysis is done by checking how large is the error made when plugging the exact solution into the scheme. This is carried out using Taylor expansions, and the geometry of the computational stencil plays an important role. When the mesh has no particular symmetry, this leads to nowhere. Instead of looking to how far the numerical scheme departs from the strong form of the PDE, it is much more flexible to look at how far it departs its weak form, i.e. instead of checking $\text{div}\mathbf{f}(u) = 0$, it is better to test, for any φ smooth enough, $\int_\Omega \varphi \text{div}\mathbf{f}(u) = 0$, of course after using the Green formula.

In practice, we define the truncation error

$$\mathscr{E}(u^h, v^h) = \sum_{\sigma\in\Omega} v^h_\sigma \Bigg(\sum_{\mathscr{K}\ni\sigma} \Phi_\sigma^K(w^h_{|K})\Bigg),$$

and consider

$$\mathscr{E}(u^h) = \max_{v^h\in U_h^G, ||v^h||_{W^{1,\infty}}=1} \mathscr{E}(u^h, v^h). \tag{14}$$

We can then extend the classical definition of accuracy:

Definition 1 (Accuracy) We say that the scheme (8) is $r+1$-th order accurate if, for any smooth solution $u_{ex} \in C^{r+1}(\overline{\Omega})$ of (1), $\mathscr{E}(u_{ex}^h) \leq C\ h^{r+1}$. The constant C only depend on the family $\mathscr{T}$, the regularity of $\mathbf{f}$, on the $r+1$ derivative of u, and the boundary conditions.

Remark 1 This definition enables to get bounds on the error $u^h - u_{ex}$ if a coercivity-like inequality holds true in some adequate norm. This is well known for the SUPG/streamline diffusion method, see [17] for example, but we do not have any general result yet.

Since $u_{ex} \in C^{r+1}(\overline{\Omega})$, there are no jump across elements. Using (12), we see that, for any v^h:

$$\mathscr{E}(u_{ex}^h, v^h) = -\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}^h) + \int_{\partial\Omega} v^h \hat{\mathbf{f}}_{\mathbf{n}}(u_{ex}^h, g) \tag{15}$$

$$+ \sum_{\mathscr{K}} \frac{1}{n_{\mathscr{K}}} \sum_{\sigma,\sigma' \in \mathscr{K}} \big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^{\mathscr{K}}((u_{ex}^h)_{|\mathscr{K}}) - \Phi_\sigma^{G,\mathscr{K}}((u_{ex}^h)_{|\mathscr{K}})\big) \tag{16}$$

For the *steady* problem (1), we have the following result:

Lemma 1 *Let us recall that $\Omega \subset \mathbb{R}^d$ and is bounded.*
If the solution u_{ex} of the steady problem (1) is C^{r+1}, then

1. $\Phi_\sigma^{G,K}((u_{ex}^h)_{|K}) = O(h^{r+d})$,
2. $\Phi_\sigma^{G,e}((u_{ex}^h)_{|e}) = O(h^{r+d-1})$
3. *if the numerical flux $\hat{\mathbf{f}}$ is Lipschitz,* $-\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}^h) + \int_{\partial\Omega} v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u_{ex}^h) = O(h^{r+1})$,

Proof We start by showing the first result. The proof of the second one is similar and is omitted.

Since $u_{ex} \in C^{r+1}$, we have $\operatorname{div}\mathbf{f}(u_{ex}) = 0$ in a strong sense, thus for any $K \in \mathscr{T}_h$ and any σ,

$$\int_K \varphi_\sigma \operatorname{div}\mathbf{f}(u_{ex}) = -\int_K \nabla\varphi_\sigma \cdot \mathbf{f}(u_{ex}) + \int_{\partial K} \varphi_\sigma \mathbf{f}(u_{ex}) \cdot \mathbf{n} = 0.$$

We can subtract this relation to $\Phi_\sigma^{G,K}(u_{ex}^h)$ and get:

$$\Phi_\sigma^{G,K}(u_{ex}^h) = -\int_K \nabla\varphi_\sigma \cdot \bigg(\mathbf{f}(u_{ex}^h) - \mathbf{f}(u_{ex})\bigg) + \int_{\partial K} \varphi_\sigma \bigg(\hat{\mathbf{f}}_{\mathbf{n}}(u_{ex,|K}^h, u_{ex,K-}^h) - \mathbf{f}(u_{ex}) \cdot \mathbf{n}\bigg).$$

Since the mesh is regular, we have:

$$|K| = O(h^d), \qquad \nabla\varphi_\sigma = O(h^{-1}), \qquad |\partial K| = O(h^{d-1})$$

and since the flux $\mathbf{f}$ is C^1, we have

$$\mathbf{f}(u^h_{ex}) - \mathbf{f}(u_e) = O(h^{k+1}).$$

Last, the numerical flux is consistent so that,

$$\hat{\mathbf{f}}_{\mathbf{n}}(u^h_{ex,|K}, u^h_{ex,K^-}) - \mathbf{f}(u_{ex}) \cdot \mathbf{n} = O(h^{k+1}).$$

Gathering the pieces together, we get:

$$\left|\Phi^{G,K}_\sigma(u^h_{ex})\right| \leq C\Big(h^d \times h^{-1} \times h^{k+1} + h^{d-1} \times 1 \times h^{k+1}\Big) = O(h^{k+d}).$$

The third inequality is obtained in a similar manner: From (1), we have for any v^h, setting $\Gamma^- = \{x \in \partial\Omega, \nabla_u \mathbf{f}(u) \cdot \mathbf{n} < 0\}$,

$$-\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}) + \int_{\Gamma^-} v^h \mathbf{f}(u_{ex}) \cdot \mathbf{n} = 0.$$

Since the numerical flux $\hat{\mathbf{f}}$ is upwind, we can rewrite this as:

$$-\int_\Omega \nabla v^h \cdot \mathbf{f}(u_{ex}) + \int_{\partial\Omega} v^h \hat{\mathbf{f}}(g, u_{ex}) \cdot \mathbf{n} = 0.$$

so that

$$\begin{aligned}
-\int_\Omega \nabla v^h \cdot \mathbf{f}(u^h_{ex}) &+ \int_{\partial\Omega} v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u^h_{ex}) \\
&= -\int_\Omega \nabla v^h \cdot \big(\mathbf{f}(u^h_{ex}) - \mathbf{f}(u_{ex})\big) + \int_{\partial\Omega} v^h \Big(\hat{\mathbf{f}}_{\mathbf{n}}(g, u^h_{ex}) - \mathbf{f}(u^h_{ex}) \cdot \mathbf{n}\Big) \\
&= (I) + (II)
\end{aligned}$$

Using again the same arguments, since the numerical flux is Lipschitz continuous, we see that both (I) and (II) are of the order of $O(h^{k+1}) \times ||v^h||_{W^{1,\infty}(\Omega)}$.

Then, we have:

Proposition 2 *Under the assumptions of Lemma 1 and assuming that the family of meshes $\mathscr{F}$ is regular, the residuals satisfy:*

$$\textit{for all } \sigma \textit{ and all } \mathscr{K} = K \text{ or } e, \Phi^{\mathscr{K}}_\sigma((u_{ex})_{|\mathscr{K}}) = O(h^{r+D}) \qquad (17)$$

where $D = d$ for elements K and $D = d-1$ for $e \in \mathscr{E}$. The scheme is formally $r+1$ accurate.

Proof $\mathscr{E}(u^h_{ex})$ is the sum of

$$-\int_\Omega \nabla v^h \cdot \mathbf{f}(u^h_{ex}) + \int_\Omega v^h \hat{\mathbf{f}}_{\mathbf{n}}(g, u^h_{ex})$$

which is $O(h^{r+1})$ by lemma 1 and

$$\sum_K \frac{1}{n_K} \sum_{\sigma,\sigma'\in K} \big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^K(w^h_{|K}) - \Phi_\sigma^{G,K}(w^h_{|K})\big) \\ + \sum_{e\subset\Omega} \frac{1}{n_e} \sum_{\sigma,\sigma'\in e} \big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^e(w^h_{|K}) - \Phi_\sigma^{G,e}(w^h_{|K})\big)$$

Since the mesh is regular, the number of elements in the mesh is $O(h^{-d})$ and the number of boundary elements is $O(h^{d-1})$. Since $v \in W^{1,\infty}$, its Lagrange interpolant satisfy

$$\big|v_\sigma^h - v_{\sigma'}^h\big| \leq h||v^h||_{W^{1,\infty}}$$

and $\sup_h ||v^h||_{W^{1,\infty}}$ is bounded by a constant that depends on $\mathscr{T}$ and $||v||_{1,\infty}$. Then we see that

$$\begin{aligned}\Big|\sum_K \frac{1}{n_K} \sum_{\sigma,\sigma'\in K} &\big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^K(w^h_{|K}) - \Phi_\sigma^{G,K}(w^h_{|K})\big) \\ &+ \sum_{e\subset\partial\Omega} \frac{1}{n_e} \sum_{\sigma,\sigma'\in e} \big(v_\sigma^h - v_{\sigma'}^h\big)\big(\Phi_\sigma^e(w^h_{|K}) - \Phi_\sigma^{G,e}(w^h_{|K})\big)\Big| \\ &\leq C\big(h^{-d}\times h\times h^{d+r} + h^{-d+1}\times h\times h^{r+d-1}\big) \\ &\leq Ch^{r+1}.\end{aligned}$$

We can estimate the boundary terms in a similar way. This ends the proof.

3 Construction of Monotonicity Preserving Arbitrary Accurate Schemes

This section aims at showing how one can combine formal accuracy and non oscillatory properties of the solution. This relies on the use of a discrete local maximum principle. By this we mean the following. Considering a scheme which update the degrees of freedom $\{u_\sigma^m\}$ that describe the solution at time t_m, $m \in \mathbb{N}$. We assume the structure: for any σ,

$$u_\sigma^{n+1} = \Theta(u_\sigma^n, \{u_{\sigma'}^n, \sigma' \in \mathscr{N}_\sigma\}, \Lambda_\sigma)$$

where $\mathcal{N}_\sigma$ is the set of neighbors of σ and Λ_σ a set of discretisation parameters. The precise definition of $\mathcal{N}_\sigma$ depends on the operator Θ. Doing so we have in mind a graph connecting together the degrees of freedom, and the notion of neighbors has to be understood as the degrees of freedom that are connected for this graph to σ. In the RD schemes, this set of neighbors are the degrees of freedom that belong to all the element that share σ. Here Λ describes the geometry of the mesh and takes into account the time increment Δt. On the set of all possible sets Λ, we also assume there is a total order relation "<".

By local maximum principle, we mean that there exists Λ_0 such that for any σ and Λ_σ such that $\Lambda_\sigma < \Lambda_0$, and for any n,

$$|u_\sigma^{n+1}| \leq \max_{\sigma' \in \mathcal{N}_\sigma \cup \{\sigma\}} |u_{\sigma'}^n|.$$

In the following, the relation "<" will be made precise for the particular example we are dealing with.

3.1 A Preliminary Remark

We start by a basic remark that goes at least back to A. Harten [15], and we rephrase it in the Residual Distribution framework.

Lemma 2 *Assume that the residuals (for element and edges) write, for any degree of freedom,*

$$\Phi_\sigma^{\mathcal{K}}(u_h) = \sum_{\sigma' \ni \mathcal{K}} c_{\sigma\sigma'}^K (u_\sigma - u_{\sigma'}), \tag{18}$$

then the iterative scheme

$$u_\sigma^{n+1} = u_\sigma^n - \omega_\sigma \left(\sum_{K \ni \sigma} \Phi_\sigma^K + \sum_{e \ni \sigma} \Phi_\sigma^e \right)$$

admits a local maximum principle if

- *for any σ, σ', $c_{\sigma\sigma'}^K \geq 0$,*
- $\omega_\sigma \left(\sum_{K \ni \sigma} \sum_{\sigma' \in K} c_{\sigma\sigma'}^K + \sum_{\sigma' \in K} c_{\sigma\sigma'} \right) \leq 1$

Here, $\mathcal{N}_\sigma$ is the set of degrees of freedom that belong to any element sharing σ, and $\Lambda_\sigma = \{\omega_\sigma\}$. We say that $\Lambda = \{\omega\} < \Lambda' = \{\omega'\}$ if $\omega \leq \omega'$.

Proof It is clear that:

$$\begin{aligned}\sum_{K\ni\sigma}\Phi_\sigma^K+\sum_{e\ni\sigma}\Phi_\sigma^e&=\Big(\sum_{K\ni\sigma}\sum_{\sigma'\in K}c_{\sigma\sigma'}^K+\sum_{\sigma'\in K}c_{\sigma\sigma'}^K\Big)u_\sigma\\&\quad-\sum_{\sigma'}\Big(\sum_{\sigma,\sigma'\in K}c_{\sigma\sigma'}^K\Big)u_{\sigma'}\\&=d_\sigma u_\sigma-\sum_{\sigma'\in\mathscr{N}_\sigma}d_{\sigma'}u_{\sigma'}\end{aligned}$$

Here, in order to simplify the notations, we have set $c_{\sigma,\sigma'}^{\mathscr{K}}=0$ when $\sigma\notin\mathscr{K}$ or $\sigma'\notin\mathscr{K}$.

The results holds true because $c_{\sigma\sigma'}^{\mathscr{K}}\geq 0$, and

$$\sum_{\sigma'\in\mathscr{N}_\sigma}c_{\sigma\sigma'}^{\mathscr{K}}\geq 0$$

$$d_{\sigma'}=\sum_{\sigma'}\Big(\sum_{\sigma,\sigma'\in K}c_{\sigma\sigma'}^K\Big)\geq 0$$

and

$$d_\sigma=1-\omega_\sigma\sum_{\sigma'\in\mathscr{N}_\sigma}c_{\sigma\sigma'}^{\mathscr{K}}\geq 0$$

if and only if the second condition of lemma 2 holds true.

The idea is to construct schemes that satisfy the requirement $c_{\sigma,\sigma'}^{\mathscr{K}}\geq 0$. It is known since Godunov that one cannot have a scheme that is simultaneously monotonicity preserving, high order accurate and linear (for linear problems). Hence some sort of non linearity must be introduced. Before showing how we can meet the requirements, let us introduce our reference monotone scheme. It is a multidimensional extension of the Rusanov scheme, namely, for any $\mathscr{K}$ and σ,

$$\Phi_\sigma^{\mathscr{K}}=\frac{1}{n_{\mathscr{K}}}\Phi^{\mathscr{K}}+\alpha_k\big(u_\sigma-\overline{u}_{\mathscr{K}}\big),\qquad \overline{u}_{\mathscr{K}}=\frac{1}{n_{\mathscr{K}}}\sum_{\sigma\in\mathscr{K}}u_\sigma \tag{19}$$

In the case of continuous elements, this scheme has the form (18). It is monotone if $\alpha_K\geq\max_{\mathbf{x}\in\mathscr{K}}||\nabla_u\mathbf{f}(u^h(\mathbf{x})||$. In the discontinuous case, a simple variant can be found, see [2].

Other examples can be constructed, starting from any classical monotone finite volume scheme. However, the interesting ones are the residuals for which the condition $\Phi^{\mathscr{K}}(u_{ex}^h)=O(h_{\mathscr{K}}^{k+D})$ holds true because of proposition 2.

3.2 Explicit Construction

The construction is local to an element (or boundary edge) $\mathscr{K}$, so we drop the dependency with respect to the element. We start from a monotone first order scheme, such as the Rusanov or the N scheme, denote the first order residuals in the element as $\{\Phi_\sigma^M\}_{\sigma\in K}$ and the high order residuals (to be constructed) by $\{\Phi_\sigma^H\}_\sigma$. We then make the following formal observation:

$$\text{for all } \sigma\in\mathscr{K}, \Phi_\sigma^H = \frac{\Phi_\sigma^H}{\Phi_\sigma^M}\Phi_\sigma^M,$$

so that if $\Phi_\sigma^M = \sum\limits_{\sigma'\in\mathscr{K}} c_{\sigma\sigma'}^M(u_{\sigma'}-u_\sigma)$, we have

$$\begin{aligned}\phi_\sigma^H &= \frac{\Phi_\sigma^H}{\Phi_\sigma^M}\Big(\sum_{\sigma'\in\mathscr{K}} c_{\sigma\sigma'}^M(u_{\sigma'}-u_\sigma)\Big)\\ &= \sum_{\sigma'\in\mathscr{K}}\left(\frac{\Phi_\sigma^H}{\Phi_\sigma^M}c_{\sigma'\sigma}^M\right)(u_{\sigma'}-u_\sigma)\Big)\\ &= \sum_{\sigma'\in\mathscr{K}} c_{\sigma'\sigma}^H(u_{\sigma'}-u_\sigma)\Big)\end{aligned}$$

with $c_{\sigma'\sigma}^H := \frac{\Phi_\sigma^H}{\Phi_\sigma^M}c_{\sigma'\sigma}^M$. Hence, to have $c_{\sigma'\sigma}^H \geq 0$, it is enough that

$$\Phi_\sigma^H\;\Phi_\sigma^M \geq 0$$

Introducing the parameters $\beta_\sigma^M = \frac{\Phi_\sigma^M}{\Phi}$ and $\beta_\sigma^H = \frac{\Phi_\sigma^H}{\Phi}$ where Φ is the total residual on the element $\mathscr{K}$, we see that:

- $\Phi_\sigma^H\;\Phi_\sigma^M \geq 0$ for any $\sigma\in\mathscr{K}$ is equivalent to $\beta_\sigma^M\;\beta_\sigma^H \geq 0$ for any $\sigma\in\mathscr{K}$,
- the conservation relations translates into:

$$\sum_{\sigma\in\mathscr{K}}\beta_\sigma^M = \sum_{\sigma\in\mathscr{K}}\beta_\sigma^H = 1. \tag{20}$$

- In order to guaranty the condition (17), a sufficient condition is that : for any C, and u^h such that $||u^h||_\infty \leq C$, there exists C' such that $|\beta_\sigma^H| \leq C'(C)$, uniformly for all meshes $\mathscr{T}_h$.

These constraints can easily be interpreted geometrically. Consider an (abstract) simplex $\mathscr{S} = (\mathbf{a}_1,\dots\mathbf{a}_{N_\mathscr{K}})$ of dimension $n_\mathscr{K}-1$ points, i.e. a triangle when $n_\mathscr{K} = 3$, a tetrahedron for $n_\mathscr{K} = 4$ and so on. These points have nothing to do with the mesh, they are only used to represent easily the constraint (20): it is well known that

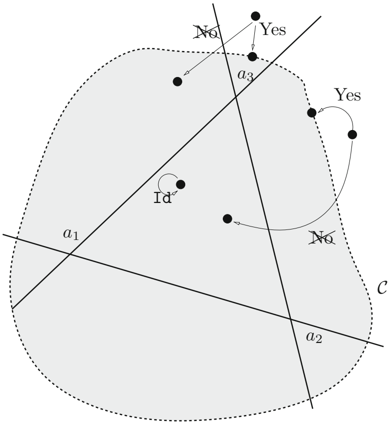


Fig. 1 Geometrical representation of the monotonicity conditions. The invariant domain is materialized by the domain inside of $\mathscr{C}$

any point $\mathbf{M}$ of an affine space of dimension $n_{\mathscr{K}} - 1$ can be uniquely described in term of its barycentric coordinates with respect to $\mathscr{S}$:

$$M = \sum_{i=1}^{n_{\mathscr{K}}-1} \lambda_i \mathbf{a}_i, \sum_{i=1}^{n_{\mathscr{K}}-1} \lambda_i = 1$$

so thus this suggests to interpret the parameters β_σ^M and β_σ^H as barycentric coordinates with respect to the simplex $\mathscr{S}$: we interpret a scheme as a point in this abstract affine space, and finding the mapping $(\beta_\sigma^M)_{\sigma \in \mathscr{K}} \mapsto (\beta_\sigma^H)_{\sigma \in \mathscr{K}}$ can be interpreted to find a mapping from this affine space onto itself. Then, to make the discussion more visual, we switch to $n_{\mathscr{K}} = 3$, see Fig. 1. The conditions $\beta_\sigma^H \beta_\sigma^L \geq 0$ are interpreted as saying that β_i^H and β_i^L must be on the same side of the line $\lambda_i = 0$. The condition $|\beta_\sigma| \leq C$ is materialized, on Fig. 1, by the domain inside curve $\mathscr{C}$. Inside the invariant domain bounded by $\mathscr{C}$, the mapping is the identity, outside of $\mathscr{C}$ project the point $L = \sum_\sigma \beta_\sigma^L \mathbf{a}_\sigma$ on $\mathscr{C}$ without crossing the lines $\lambda_{\sigma_i} = 0$. Once the β_σ^H are defined, we set simply $\Phi_\sigma^H = \beta_\sigma^H \Phi$.

The simplest invariant domain is certainly the simplex $(\mathbf{a}_1, \ldots, \mathbf{a}_{n_{\mathscr{K}}})$ for which $0 \leq \lambda_\sigma \leq 1$. In that case, the most common formula is [6, 23]:

$$\beta_\sigma^H = \frac{\max(\beta_\sigma^M, 0)}{\sum\limits_{\sigma \in \mathscr{K}} \max(\beta_\sigma^M, 0)}. \tag{21}$$

Note that $\sum\limits_{\sigma \in \mathscr{K}} \max(\beta_\sigma^M, 0) \geq 1$ because

$$1 = \sum_{\sigma \in \mathscr{K}} \beta_\sigma^M = \sum_{\sigma \in \mathscr{K}} \max(\beta_\sigma^M, 0) + \sum_{\sigma \in \mathscr{K}} \min(\beta_\sigma^M, 0) \leq \sum_{\sigma \in \mathscr{K}} \max(\beta_\sigma^M, 0).$$

When $\Phi = 0$, we simply set $\Phi_\sigma^H = 0$

3.3 Filtering

In practice, this method is excellent for computing discontinuous solutions. When computing smoother solutions, we can see “wiggles” appearing, see Sect. 4. They are not a manifestation of any instability since the scheme is perfectly L^∞ stable, but it is too over compressive, i.e. not dissipative enough.

It is quite easy to understand what is going on. We first, let us consider the problem on $[0,1]^2$:

$$\frac{\partial u}{\partial x} = 0 \tag{22}$$

with the boundary condition $u = g$ on $\{0\} \times [0,1]$. The grid is made of quadrangles, with vertices (x_i, y_j), $x_i = \frac{i}{N}$, $y_j = \frac{j}{N}$, $0 \leq i, j \leq N$. The function g is piecewise linear, and $g(0, y_j) = (-1)^j$. The exact solution is independent of x.

The scheme is defined by

$$u_{ij}^{n+1} = u_{ij}^n - \omega_{ij} \sum_{K \ni (x_i, y_j)} \Phi_{i,j}^{H,K}(u_h^n)$$

with u_{ij}^0 given, and $u_{0j}^n = g(0, y_j)$. There are many ways of initializing, we consider two initializations:

- Initialization with the exact solution: $u_{ij}^0 = g(0, y_j) = (-1)^j$
- Check-board mode: $u_{ij}^0 = (-1)^{i+j}$

The solution at the n-th iteration is reconstructed with the $\mathbb{Q}^1$ interpolation. It is easy to see that for both initialization, we have, for any K,

$$\Phi^K = \int_{\partial K} u^h \mathbf{n}_x = 0$$

so that in both cases, for any i, j, n, $u_{ij}^n = u_{ij}^0$! The method as such is not well posed, and there are spurious modes.

To remedy to this serious drawback, there are several possibilities. Here we discuss a solution already described in see [1], and a new one that is inspired by Burman’s variational formulation.

3.3.1 Streamline Filtering

The one discussed in [1] is inspired by the streamline diffusion method. Namely starting from an unfiltered family of residuals $\{\Phi_\sigma^{H,K}\}$ constructed as in Sect. 3.2, we add a streamline diffusion term:

$$\Phi_\sigma^{H,K,\star} = \Phi_\sigma^{H,K} + \theta_K h_K \int_K \left(\nabla_u \mathbf{f}(u^h) \cdot \nabla \varphi_\sigma\right) N \left(\nabla_u \mathbf{f}(u^h) \cdot \nabla u^h\right) \tag{23}$$

where N is defined by

$$N = \left(\sum_{\sigma\in K} \max\left(\overline{\nabla_{\mathbf{u}}\mathbf{f}}, 0\right) + \varepsilon\right)^{-1}$$

with the gradient $\overline{\nabla_{\mathbf{u}}\mathbf{f}}$ evaluated at the centroid and ε is a small number to avoid singularity. The choice of where is evaluated the average gradient does not seem to be fundamental. The parameter should be $\theta_K \approx 0$ in discontinuities and $\theta_K \approx 1$ away from discontinuities. When we apply this correction (with $\theta = 1$) to (22) this corrects the problem. By construction, we see that the accuracy requirement of lemma 1 are met if they are met for the unfiltered scheme

To see what is the rational behind (23), let us first switch to the one dimensional problem:

$$\begin{aligned}\frac{\partial f(u)}{\partial x} &= 0 \quad x\in[0,1]\\ u(0) &= u_0\\ u(1) &= u_1.\end{aligned}\tag{24}$$

The boundary conditions are imposed weakly, and to make things simple, assume $f'(u_0) > 0$ and $f'(u_1) < 0$ so that the solution is $u = u_0$. The interval $[0, 1]$ is discretized with the mesh which elements are $[x_i, x_{i+1}]$, $0 = x_0 < x_1 < \ldots < x_{n-1} < x_n = 1$. Whatever the order, the total residual is for $K_{i+1/2} = [x_i, x_{i+1}]$

$$\Phi^{K_{i+1/2}} = f(u_{i+1}) - f(u_i)$$

so that the high order residuals are simply, for any degree of freedom $\sigma \in K$, $\Phi_\sigma^K = \beta_\sigma^K\big(f(u_{i+1} - f(u_i)\big)$. In particular, the internal degrees of freedom play no role. Assume now that $k = 1$, there is no internal degree of freedom, and let us evaluate the entropy balance for the entropy $U(u) = \frac{1}{2}u^2$: using the notation $\gamma_j^{K_{i+1/2}} = \beta_j^{K_{i+1/2}} - \frac{1}{2}$, we have

$$\begin{aligned}\mathscr{E} &= \sum_{i=0}^{N-1} u_i\Big(\beta_i^{K_{i-1/2}}\big(f(u_i) - f(u_{i-1})\big) + \beta_i^{K_{i+1/2}}\big(f(u_{i+1}) - f(u_i)\big)\Big)\\ &= \int_0^1 u^h \frac{\partial f}{\partial x}(u^h) + \sum_{i=0}^{N-1}\Big(\gamma_i^{K_{i+1/2}} u_i + \gamma_{i+1}^{K_{i+1/2}} u_{i+1/2}\Big)\big(f(u_{i+1}) - f(u_i)\big)\\ &= \int_0^1 u^h \frac{\partial f}{\partial x}(u^h) + \sum_{i=0}^{N-1}\gamma_{i+1}^{K_{i+1/2}}(f(u_{i+1}) - f(u_i))(u_{i+1} - u_i).\end{aligned}$$

with the convention $u_{-1} = u_0$ and $u_{N+1} = u_N$ to take into account the boundary conditions. For the scheme to be dissipative, a sufficient condition is that for all i, $\gamma_{i+1}^{K_{i+1/2}}(f(u_{i+1}) - f(u_i))(u_{i+1} - u_i) \geq 0$, i.e.

$$\gamma_{i+1}^{K_{i+1/2}}\frac{f(u_{i+1}) - f(u_i)}{u_{i+1} - u_i} \geq 0$$

with a strict inequality for at least one interval.

The evaluation of $\beta_\sigma^{K_{i+1/2}}$ is done with the only aim of having an L^∞ stable scheme, so that this inequality might not be true.[4] Adding the streamline term, i.e. in this case,

$$\theta(u_{i+1} - u_i)\int_{x_i}^{x_{i+1}} N\big(\frac{\partial f}{\partial u}\big)^2\frac{\partial \varphi_\sigma}{\partial x} = (u_{i+1} - u_i)\big|\frac{\partial f}{\partial u}\big|\big(\varphi_\sigma(x_{i+1}) - \varphi_\sigma(x_i)\big)$$

will modify the entropy balance into

$$\mathscr{E} = \int_0^1 u^h\frac{\partial f}{\partial x}(u^h) + \sum_{i=0}^{N-1}\left(\gamma_{i+1}^{K_{i+1/2}}\frac{f(u_{i+1}) - f(u_i)}{u_{i+1} - u_i} + \theta\big|\frac{\partial f}{\partial u}\big|\right)(u_{i+1} - u_i)^2$$

and $\mathscr{E} \leq \int_0^1 u^h\frac{\partial f}{\partial x}(u^h)$ provided that $\theta \geq 1$.

3.3.2 Jump Filtering

The idea is to add to the unfiltered residuals $\{\Phi_\sigma^{H,K}\}$ constructed as in Sect. 3.2, we add a jump term inspired by Burman's construction, namely:

$$\Phi_\sigma^{H,K,\star} = \Phi_\sigma^{H,K} + \sum_{e\in \mathrm{e}_h, e\subset K}\Gamma_e h_e^2\int_e [\nabla u_h][\nabla\varphi_\sigma]. \tag{25}$$

We first check that the conditions of lemma 1 are met if the unfiltered scheme satisfies them too. For this, we only need to check that if the exact solution is $C^{r+1}(\Omega)$ and if we are using polynomials of degree at most r, then

$$\sum_{e\in \mathrm{e}_h, e\subset K}\Gamma_e h_e^2\int_e [\nabla u_h][\nabla\varphi_\sigma] = O(h^{r+d}).$$

[4] However, in 1D it is very simple to show that the sign condition is true, let us ignore this fact however.

Since $[\nabla u_{ex}] = 0$, we have

$$\int_e [\nabla u_h][\nabla \varphi_\sigma] = \int_e [\nabla (u_h - u_{ex})][\nabla \varphi_\sigma] = O(h^{d-1}) \times O(h^r) \times O(h^{-1}) = O(h^{d+r-2})$$

and thus the conditions are met.

We notice that if for any internal face e, $\int_e [\nabla u]^2 = 0$, then u is globally a polynomial of degree r. We first show if v is a polynomial of degree q on each element K such that for any face e, $\int_e [v]^2 = 0$ then v is a polynomial of degree q defined on the whole domain Ω. The second step is to apply this to $v = \nabla u$.

Let $v \in \bigoplus\limits_{K \in \mathcal{T}_h} \mathbb{P}^q(K)$, we define the operator π that maps v The operator π is defined as follow: for any σ,

$$\pi(v)(\sigma) = \frac{1}{\#\{K, \sigma \in K\}} \sum_{K, \sigma \in K} v_{|K}(\sigma)$$

and then

$$\pi(v) = \sum_{\sigma \in \Sigma_h} \pi(v)(\sigma)\varphi_\sigma.$$

This definition assumes that the we are using Lagrange interpolation, we have done this for simplicity but this is not essential.

Let us have a look at $v - \pi(v)$ on any K. Since $v - \pi(v) = \sum\limits_{\sigma \in K} (v_{|K}(\sigma) - \pi(v)(\sigma))\varphi_\sigma$, we look at the difference $v_{|K}(\sigma) - \pi(v)(\sigma)$. We have

$$v_{|K}(\sigma) - \pi(v)(\sigma) = \frac{1}{\#\{K', \sigma \in K'\}} \sum_{\sigma \in K'} (v_K(\sigma) - v_{K'}(\sigma)).$$

If σ is internal to K, $v_{|K}(\sigma) - \pi(v)(\sigma) = 0$, so the difference is possibly $\neq 0$ only for degrees of freedom on the edges. If the mesh is regular, we can easily see that

$$\begin{aligned}\int_K \Big(\big(\sum_{\sigma \in K'} (v_K(\sigma) - v_{K'}(\sigma)) \big) \varphi_\sigma \Big)^2 &\leq C|K| \sum_{\sigma \in \partial K} [v(\sigma)]^2 \\ &\leq C' \sum_{e \subset \partial K} h_e \int_e [v]^2\end{aligned}$$

where C and C' are constants that depends on the mesh regularity, so that

$$\begin{aligned}\int_\Omega |v-\pi(v)|^2 &= \sum_K \int_K |v-\pi(v)|^2 \\ &= \sum_K \int_K \Big(\frac{1}{\#\{K', \sigma \in K'\}} \sum_{\sigma \in K' \cap K} (v_K(\sigma) - v_{K'}(\sigma))\varphi_\sigma \Big)^2 \\ &\leq C \sum_e h_e \int_e [v]^2\end{aligned}$$

From this we see that if for any internal face e, $\int_e [v]^2 = 0$, then $v = \pi(v)$. If we apply this result to ∇u, we see that $\pi(\frac{\partial u}{\partial x_i}) = \frac{\partial u}{\partial x_i}$ for any component, so by integration, u is a global polynomial. This a very particular case of much general results, see [13].

This remark explains the potential role of the jump term: if the solution is smooth, the setting $\Gamma_e > 0$ will constraint the continuity of the solution across faces, and hopefully will bound $||\nabla u||$. If $\Gamma = 0$, this constraint is relaxed. So the idea is, again, to take $\Gamma > 0$ where the solution is expected to be smooth, and $\Gamma = 0$ where it is expected to be discontinuous. For now, the main justification of these choices is purely heuristic and motivated by numerical experiments.

4 Numerical Examples

In this section, we illustrate the behavior of the method on two examples: a linear transport problem and a non linear one. In $\Omega = [0,1]^2$, we consider

$$\boldsymbol{\lambda} = (y, -x)^T \quad \text{and} \quad u(x,y) = \varphi_0(x) \mathrm{if} y = 0 \tag{26}$$

with the boundary conditions

$$\varphi_0(x) = \begin{cases} \cos^2(2\pi x) & \text{if } x \in [\frac{1}{4}, \frac{3}{4}] \\ 0 & \text{else} \end{cases}$$

The isolines of the exact solution are circles of center $(0,0)$. The form of the Burgers equation is the following:

$$\begin{aligned}&\frac{\partial u}{\partial y} + \tfrac{1}{2}\frac{\partial u^2}{\partial x} = 0 \quad \text{if} \quad x \in [0,1]^2 \\ &u(x,y) = 1.5 - 2x \text{ on the inflow boundary.}\end{aligned} \tag{27a}$$

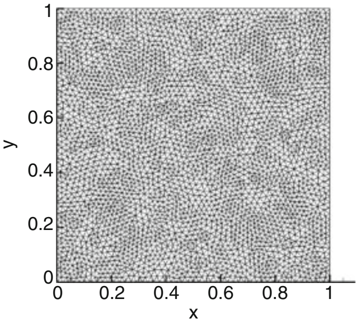


Fig. 2 Mesh for the numerical experiments

The exact solution consists in a fan that merges into a shock which foot is located at $(x,y)=(3/4,1/2)$. More precisely, the exact solution is

$$u(x,y)=\begin{cases} \text{if} \quad y\geq 0.5 & \begin{cases} -0.5 \text{ if} -2(x-3/4)+(y-1/2)\leq 0 \\ 1.5 \qquad\qquad \text{else}\end{cases} \\ \quad\text{else} & \max\left(-0.5, \min\left(1.5, \dfrac{x-3/4}{y-1/2}\right)\right)\end{cases} \tag{27b}$$

The mesh displayed on Fig. 2 is used to obtain the solutions shown on Figs. 3 and 4. All the meshes used in this paper have been generated by GMSH [14]. We see, on Fig. 3a that without the streamline term in (23), the solution looks very wiggly. Again, it is not an instability, only a manifestation of spurious modes that are completely eliminated using (23) or (25). If one makes a convergence study on this problem using $\mathbb{P}^1$, $\mathbb{P}^2$ and $\mathbb{P}^3$ elements, we recover the expected order of convergence, see Table 1.

Figure 4 give the results for the problem (27). The solution is composed of a compressive fan and a discontinuity. The exact solution is plotted as well as what is obtained for the SUPG scheme (3b), the Galerkin scheme with jump stabilisation (5), the original non linear RD scheme using (21) to evaluate β_σ, and the schemes when this RD scheme is combined with streamline (23) and jump filtering (25). As expected the SUPG and Galerkin+jump methods are oscillatory (and the latter one proves to be extremely oscillatory; we have chosen $\Gamma=0.1$ here). The non linear methods behave very well. For the streamline filtering, we have taken $\theta=1$, and for the jump filtering $\Gamma=0.1$ in the smooth part, 0 elsewhere. We need to improve this, this work is in progress. The jump filtering seems to be less dissipative than the stream line stabilisation. All these results use quadratic reconstruction, the last two figures use linear reconstruction. The same conclusions hold.

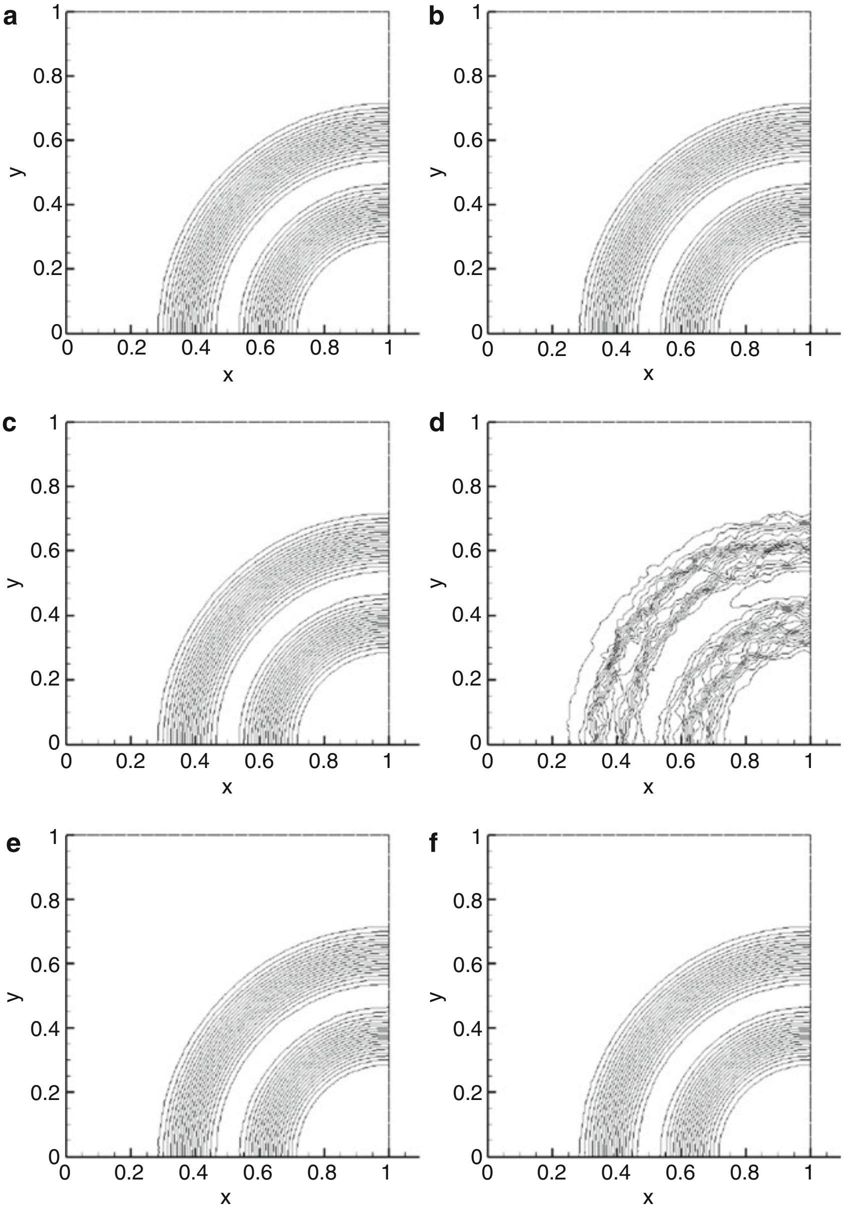


Fig. 3 Solution of (26) with (21), (23) and (25), $\mathbb{P}^2$ elements. In each figure, 19 isolines form 0.05 to 0.95 are plotted. (**a**) exact; (**b**) Supg; (**c**) Galerkin+jump term; (**d**) without streamline term in (23); (**e**) with the streamline term (23); (**f**) with the jump term (25)

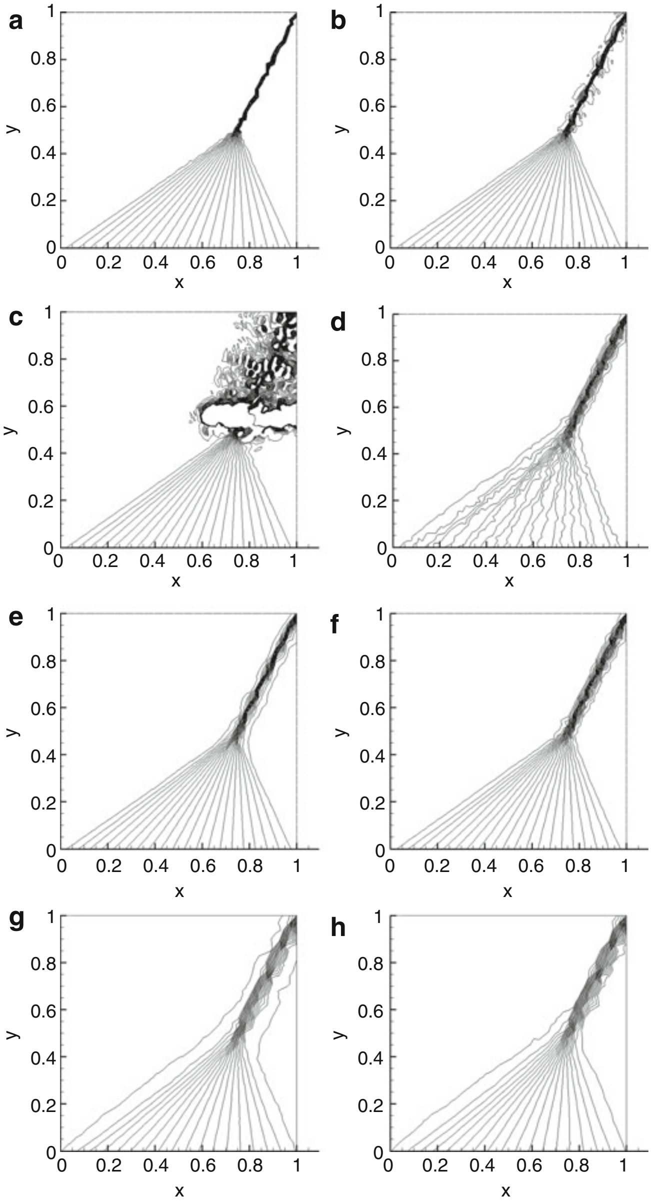


Fig. 4 Solution of (27). The solutions of (h) and (i) are obtained by $\mathbb{P}^1$ elements while the other are obtained with $\mathbb{P}^2$ elements. The number of degrees of freedom is the same for each plot. We use 20 isolines form -0.6 to 1.6 in all sub-figures. (**a**) Exact sln; (**b**) Supg: scheme (3b); (**c**) Galerkin+Jump: scheme (5); (**d**) Psi without filtering (RDS using only: scheme (21)); (**e**) Psi+stream: scheme (23); (**f**) Psi+Jump: scheme (25); (**g**) Psi+stream (P1); (**h**) Psi+Jump(P1)

Table 1 Order of accuracy on refined mesh constructed from the mesh of Fig. 2, L^2 norm. The slopes are obtained by least square

h	$\epsilon_{L^2}(\mathbb{P}^1)$	$\epsilon_{L^2}(\mathbb{P}^2)$	$\epsilon_{L^2}(\mathbb{P}^3)$
1/25	0.50493E-02	0.32612E-04	0.12071E-05
1/50	0.14684E-02	0.48741E-05	0.90642E-07
1/75	0.74684E-03	0.13334E-05	0.16245E-07
1/100	0.41019E-03	0.66019E-06	0.53860E-08
	$\mathcal{O}^{\text{ls}}_{L^2}$ =1.790	$\mathcal{O}^{\text{ls}}_{L^2}$ =2.848	$\mathcal{O}^{\text{ls}}_{L^2}$ =3.920

Strictly speaking, the streamline term in (23) or the jump term in (25) destroy the maximum preserving nature of the scheme: the operators defined by (23) or (25) are not, a priori, of the type (18) with positive coefficients. We have not been able, so far, to analyze in full detail the schemes from this point of view, but all the numerical experiments that we have done, including with system case (for (23), the second solution has not yet been tested for systems), indicate that the streamline term (23) or the jump term (25) act as a filter, and do not spoil the monotonicity preserving properties that we are seeking for. Actually, this property is violated, but the over- and undershoot are negligible, as what occurs for the ENO and WENO schemes.

5 Conclusions

We have shown a systematic way of constructing high order finite element like methods for scalar hyperbolic problems that preserve, in practice, a local maximum principle. The problems can be linear or not, and the solutions regular or not. We have shown that the accuracy can actually be reached. This paper present two classes of methods, one of them has already been extended to systems [5] and even to the Navier Stokes equations [3, 4]; the second one has to be extended to systems, and this should be straightforward.

Many things remain to be done. The methods are intended to be parameter free. One part of the numerical operator is proved to be maximum principle preserving, without any parameter to tune. Unfortunately, by using only this operator, we can see that the solution may develop spurious modes (while keeping the maximum principle), and we have shown how to cure this. Unfortunately, we had to introduce one tunable parameter. When we filter out by using a streamline filter, we have proposed solutions [5] to monitor the filter, but because of the very writing of the scheme it is difficult to make decisions by looking at the local structure of the solution without violating the structure of the numerical stencil. In the second case, the stabilisation is done via an integral term involving the jump of the first derivative of the solution. This opens new perspectives for a better design of the filtering parameter. An extension to unsteady problems is also in progress. The extension to 3D system is straightforward and has already be done with an extension of the streamline filtering, see [18] for the Euler equations and [3, 4] for the Navier-Stokes ones.

Acknowledgements This work has been partially funded by the SNF grant # 200021_153604 of the Swiss National Foundation. Stimulating discussions with Dr. A. Burbeau (CEA, France), are also warmly acknowledged.

References

1. R. Abgrall. Essentially non-oscillatory residual distribution schemes for hyperbolic problems. *J. Comput. Phys.*, 214(2):773–808, 2006.
2. R. Abgrall. A residual method using discontinuous elements for the computation of possibly non smooth flows. *Adv. Appl. Math. Mech*, 2010.
3. R. Abgrall and D. de Santis. High-order preserving residual distribution schemes for advection-diffusion scalar problems on arbitrary grids. *SIAM I. Sci. Comput.*, 36(3):A955–A983, 2014. also http://hal.inria.fr/docs/00/76/11/59/PDF/8157.pdf.
4. R. Abgrall and D. de Santis. Linear and non-linear high order accurate residual distribution schemes for the discretization of the steady compressible navier-stokes equations. *J. Comput. Phys.*, 283:326–359, 2015.
5. R. Abgrall, A. Larat, and M. Ricchiuto. Construction of very high order residual distribution schemes for steady inviscid flow problems on hybrid unstructured meshes. *J. Comput. Phys.*, 230(11):4103–4136, 2011.
6. R. Abgrall and P. L. Roe. High-order fluctuation schemes on triangular meshes. *J. Sci. Comput.*, 19(1–3):3–36, 2003.
7. R. Abgrall and C.W. Shu. Development of residual distribution schemes for discontinuous galerkin methods. *Commun. Comput. Phys.*, 5:376–390, 2009.
8. R. Abgrall, S. Tokereva, and J. Nordström. Entropy stable fem methods. in preparation.
9. R. Abgrall and J. Trefilick. An example of high order residual distribution scheme using non-lagrange elements. *J. Sci. Comput.*, 45(1–3):3–25, 2010.
10. E. Burman and P. Hansbo. Edge stabilisation for galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193:1437–1453, 2004.
11. B. Cockburn, S. Hou, and C.-W. Shu. TVB Runge-Kutta local projection discontinuous finite element method for conservation laws IV: the multidimensional case. *Math. Comp.*, 54: 545–581, 1990.
12. B. Cockburn and C.-W. Shu. TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws II: General framework. *Math. Comp.*, 52:411–435, 1989.
13. A. Ern and J.L. Guermond. Finite element quasi-interpolation and best approximation. arXiv:1505.06931, May 2015.
14. C. Geuzaine and J.-F. Remacle. A three-dimensional finite element mesh generator with built-in pre- and post-processing facilities. http://gmsh.info/.
15. A. Harten. On a class of high resolution total-variational-stable finite-difference schemes (with appendix by Peter D. Lax). *SIAM J. Numer. Anal.*, 21:1–23, 1984.
16. T.J.R. Hughes, L.P. Franca, and M. Mallet. A new finite element formulation for CFD: I. symmetric forms of the compressible Euler and Navier-Stokes equations and the second law of thermodynamics. *Comp. Meth. Appl. Mech. Engrg.*, 54:223–234, 1986.
17. C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic problems. *Computer methods in applied mechanics and engineering*, 45:285–312, 1985.
18. N. Kroll, H. Bieler, H. Deconinck, V. Couaillier, H. van der Ven, and K. Sorensen, editors. *ADIGMA- A European Initiative on the Development of Adaptive Higher-Order Variational Methods for Aerospace Applications*. Notes on Numerical Fluid Mechanics and Multidisciplinary Design. Springer, 2010. Results of a Collaborative Research Project Funded by the European Union, 2006–2009.
19. D. Kröner, M. Rokyta, and M. Wierse. A Lax-Wendroff type theorem for upwind finite volume schemes in 2-d. *East-West J. Numer. math.*, 4(4):279–292, 1996.

20. R.-H. Ni. A multiple grid scheme for solving the Euler equations. In *5th Computational Fluid Dynamics Conference*, pages 257–264, 1981.
21. P.L. Roe. Approximate Riemann solvers, parameter vectors, and difference schemes. *J. Comput. Phys.*, 43:357–372, 1981.
22. P.L. Roe. Characteristic-based schemes for the Euler equations. Annu. Rev. Fluid Mech. 18, 337–365 (1986)., 1986.
23. R. Struijs, H. Deconinck, and P.L. Roe. Fluctuation splitting schemes for the 2D Euler equations. VKI-LS 1991-01, 1991. Computational Fluid Dynamics.

Part IV
Group-Theoretical Approaches to Conservation Laws and Their Applications

Generalization of Noether's Theorem in Modern Form to Non-variational Partial Differential Equations

Stephen C. Anco

Abstract A general method using multipliers for finding the conserved integrals admitted by any given partial differential equation (PDE) or system of partial differential equations is reviewed and further developed in several ways. Multipliers are expressions whose (summed) product with a PDE (system) yields a local divergence identity which has the physical meaning of a continuity equation involving a conserved density and a spatial flux for solutions of the PDE (system). On spatial domains, the integral form of a continuity equation yields a conserved integral. When a PDE (system) is variational, multipliers are known to correspond to infinitesimal symmetries of the variational principle, and the local divergence identity relating a multiplier to a conserved integral is the same as the variational identity used in Noether's theorem for connecting conserved integrals to invariance of a variational principle. From this viewpoint, the general multiplier method is shown to constitute a modern form of Noether's theorem in which the variational principle is not directly used. When a PDE (system) is non-variational, multipliers are shown to be an adjoint counterpart to infinitesimal symmetries, and the local divergence identity that relates a multiplier to a conserved integral is shown to be an adjoint generalization of the variational identity that underlies Noether's theorem. Two main results are established for a general class of PDE systems having a solved-form for leading derivatives, which encompasses all typical PDE systems of physical interest. First, all non-trivial conserved integrals are shown to arise from non-trivial multipliers in a one-to-one manner, taking into account certain equivalence freedoms. Second, a simple scaling formula based on dimensional analysis is derived to obtain the conserved density and the spatial flux in any conserved integral, just using the corresponding multiplier and the given PDE (system). Also, a general class of multipliers that captures physically important conserved integrals such as mass, momentum, energy, angular momentum is identified. The derivations use a few basic tools from variational calculus, for which a concrete self-contained formulation is provided.

S.C. Anco (✉)
Department of Mathematics and Statistics, Brock University, St. Catharines,
ON, Canada L2S 3A1
e-mail: sanco@brocku.ca

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute
Communications 79, DOI 10.1007/978-1-4939-6969-2_5

1 Introduction and Overview

In the study of partial differential equations (PDEs), conserved integrals and local continuity equations have many important uses. They yield fundamental conserved quantities and constants of motion, which along with symmetries are an intrinsic coordinate-free aspect of the structure of a PDE system. They also yield potentials and nonlocally-related systems. They provide conserved norms and estimates, which are central to the analysis of solutions. They detect if a PDE system admits an invertible transformation into a target class of PDE systems (e.g., nonlinear to linear, or linear variable coefficient to constant coefficient). They typically indicate if a PDE system has integrability structure. They allow checking the accuracy of numerical solution methods and also give rise to good discretizations (e.g., conserving energy or momentum).

For a dynamical PDE system in one spatial dimension, a local continuity equation is a total divergence expression

$$D_tT + D_xX = 0 \tag{1}$$

vanishing on the solution space of the system, where T is a conserved density and X is a spatial flux. (Here D_t and D_x are total derivatives with respect to time and space coordinates.) Every local continuity equation physically represents a conservation law for the quantity T. The conservation law can be formulated by integrating the local continuity equation over any spatial domain $\Omega \subseteq \mathbb{R}$, yielding

$$\frac{d}{dt}\int_{\Omega} Tdx = -X\Big|_{\partial\Omega}. \tag{2}$$

This shows that the rate of change of the integral of the conserved density T on the domain Ω is balanced by the net outward flux through the domain endpoints $\partial\Omega$.

In two and three spatial dimensions, local continuity equations have the more general total divergence form

$$D_tT + \mathrm{Div}\,\mathbf{X} = 0. \tag{3}$$

The corresponding physical conservation law is given by

$$\frac{d}{dt}\int_{\Omega} TdV = -\oint_{\partial\Omega} \mathbf{X}\cdot\boldsymbol{\nu}dA \tag{4}$$

where Ω is a spatial domain and $\boldsymbol{\nu}$ is the outward unit normal of the domain boundary. This conservation law shows that the net outward flux of $\mathbf{X}$ integrated over $\partial\Omega$ balances the rate of change of the integral of the conserved density T on Ω.

Another type of conservation law in two and three spatial dimensions can be formulated on the boundary of a spatial domain Ω,

$$\frac{d}{dt}\oint_{\partial\Omega} \mathbf{T}\cdot\boldsymbol{\nu}dA = 0 \tag{5}$$

holding on the solution space of a PDE system. This boundary conservation law corresponds to a local continuity equation (3) in which the conserved density is a total spatial divergence, $T = \mathrm{Div}\,\mathbf{T}$, and the flux is a total spatial curl, $\mathbf{X} = \mathrm{Div}\,\boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}$ is an antisymmetric tensor. Its physical meaning is that the net flux of $\mathbf{T}$ over $\partial\Omega$ is a constant of the motion for the PDE system.

When hydrodynamical PDE systems for fluid/gas flow are considered, a more physically useful formulation of conservation laws is given by considering moving spatial domains $\Omega(t)$, or moving spatial boundaries $\partial\Omega(t)$, that are transported by the flow of the fluid/gas.

For a moving domain, a physical conservation law has the form

$$\frac{d}{dt}\int_{\Omega(t)} TdV = -\oint_{\partial\Omega(t)} (\mathbf{X} - T\mathbf{u})\cdot\boldsymbol{\nu}dA \tag{6}$$

where $\mathbf{u}$ is the fluid/gas velocity, and $\mathbf{X} - T\mathbf{u} = \mathscr{X}$ is the moving flux. The local continuity equation (3) is then equivalent to a transport equation

$$(D_t + \mathbf{u}\cdot D_x)T = -(\nabla\cdot\mathbf{u})T - \mathrm{Div}\,\mathscr{X} \tag{7}$$

for the conserved density T, with $D_t+\mathbf{u}\cdot D_x$ being the material (advective) derivative, and $\nabla\cdot\mathbf{u}$ being the expansion or contraction factor of an infinitesimal moving volume of the fluid/gas. If the net moving flux over the domain boundary $\partial\Omega(t)$ vanishes, then the integral of the conserved density T on the moving domain $\Omega(t)$ is a constant of motion.

For a moving boundary, the physical form of a conservation law is given by

$$\frac{d}{dt}\oint_{\partial\Omega(t)} \mathbf{T}\cdot\boldsymbol{\nu}dA = 0 \tag{8}$$

which shows the net flux of $\mathbf{T}$ integrated over $\partial\Omega(t)$ is a constant of motion. In the corresponding transport equation (7), the conserved density is a total spatial divergence, $T = \mathrm{Div}\,\mathbf{T}$, and the moving flux is a total spatial curl, $\mathscr{X} = \mathrm{Div}\,\boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}$ is an antisymmetric tensor.

A related type of conservation law in two and three spatial dimensions arises from the total spatial divergence of a flux vector that is not a total spatial curl,

$$\mathrm{Div}\,\mathbf{X} = 0, \quad \mathbf{X} \neq \mathrm{Div}\,\boldsymbol{\Gamma} \tag{9}$$

holding on the solution space of a PDE system. This yields a physical conservation law on any spatial domain Ω enclosed by an inner boundary $\partial_-\Omega$ and an outer boundary $\partial_+\Omega$. The conservation law shows that the net outward flux across each boundary is the same,

$$\oint_{\partial_- \Omega} \mathbf{X} \cdot \boldsymbol{\nu}_- dA = \oint_{\partial_+ \Omega} \mathbf{X} \cdot \boldsymbol{\nu}_+ dA \qquad (10)$$

where $\boldsymbol{\nu}_\mp$ is the outward unit normal of the respective boundaries.

The most well-known method [1–3] for finding conservation laws is Noether's theorem, which is applicable only to PDE systems that possess a variational principle. Noether's theorem shows that the infinitesimal symmetries of the variational principle yield conserved integrals (2), (4), (10) of the PDE system, including conserved boundary integrals (5) when the PDE system satisfies a differential identity. In the case of PDE systems that possess a generalized Cauchy-Kovalevskaya form [1, 4], all conserved integrals (2), (4), (10) arise from Noether's theorem. This direct connection between conserved integrals and symmetries is especially useful because, typically, the symmetries of a given PDE system have a direct physical meaning related to basic properties of the system, while, computationally, all infinitesimal symmetries of a given PDE system can be found in a systematic way by solving a linear system of determining equations.

Over the past few decades, a modern formulation of Noether's theorem has been developed in which the components of a variational symmetry are expressed as the components of a multiplier whose summed product with a given variational PDE system yields a total divergence that reduces on the space of solutions of the PDE system to a local continuity equation. The main advantage of this reformulation is that multipliers can be sought for any given PDE system, regardless of whether it possesses a variational principle or not. In general, multipliers are simply the natural PDE counterpart of integration factors for ordinary differential equations [2], and for any given PDE system, a linear system of determining equations can be formulated [1, 3] to yield all multipliers. As a consequence, local continuity equations can be derived without any restriction required on the nature of the PDE system. Moreover, for PDE systems that possess a generalized Cauchy-Kovalevskaya form, all conserved integrals (2), (4), (10) arise from multipliers [1, 4]. A review of the history of Noether's theorem and of the multiplier method for finding conservation laws can be found in Ref. [5].

In recent years, the multiplier method has been cast into the form of a generalization of Noether's theorem which is applicable to PDE systems without a variational principle. The generalization [6–9] is based on the structure of the determining system for multipliers, which turns out to be an augmented, adjoint version of the determining equations for infinitesimal symmetries. In particular, multipliers can be viewed as an adjoint generalization of variational symmetries, and most significantly, the determining system for multipliers can be solved by the use of the same standard procedure that is used for solving the determining equations for symmetries [1–3]. Moreover, the physical conservation law determined by a multiplier can be constructed directly from the multiplier and the given PDE system by various integration methods [1, 3, 7–10].

In this modern generalization, the problem of finding all conservation laws for a given PDE system thereby becomes a kind of adjoint of the problem of finding all infinitesimal symmetries. As a consequence, for any PDE system, there is no need to use special methods or ansatzes (e.g., [11–15]) for determining its conservation laws, just as there is no necessity to use special methods or ansatzes for finding its symmetries.

The present work is intended to review and extend these recent developments, with an emphasis on applications to PDE systems arising in physical models. The most natural mathematical framework for understanding the methods and the results is variational calculus in jet spaces [1]. This framework will be given a concrete formulation, which is useful both for formulating general statements and for doing calculations for specific PDE systems.

As a starting point, in Sect. 2, a wide range of examples of local conservation laws and conserved integrals are presented, covering dynamical systems that model convection, diffusion, wave propagation, fluid flow, gas dynamics and plasma dynamics, as well as non-dynamical (static equilibrium) systems.

In Sect. 3, for general PDE systems, the standard formulations of local conservation laws, conserved integrals, and symmetries, as well as some other preliminaries, are stated. Additionally, local versus global aspects of conservation laws are discussed and are related to the distinction between trivial and non-trivial conservation laws and their physical meaning. This discussion clears up some confusion in the existing literature.

In Sect. 4, some basic modern tools from variational calculus are reviewed using a concrete self-contained approach. These tools are employed in Sect. 5 to derive the determining equations for multipliers and symmetries, based on a characteristic form for conservation laws and symmetry generators. An important technical step in this derivation is the introduction of a coordinatization for the solution space of PDE systems in jet space, which involves expressing a given PDE system in a solved form for a set of leading derivatives after the system is closed by appending all integrability conditions (if any). This coordinatization is applicable to all PDE systems of physical interest, including systems that possess differential identities. It is used to show that the characteristic form for trivial conservation laws is given by trivial multipliers which vanish on the solution space of a PDE system when the system has no differential identities. This directly leads to an explicit one-to-one correspondence between non-trivial conservation laws and non-trivial multipliers, taking into account the natural equivalence freedoms in conservation laws and multipliers. An explicit generalization of this correspondence is established in the case when a PDE system possesses a differential identity (or set of identities). The generalization involves considering gauge multipliers [16] that arise from a conservation law connected with the differential identity.

These new results significantly extend the explicit correspondence between non-trivial conservation laws and non-trivial multipliers previously obtained [1, 4, 7, 8] only by requiring PDE systems to have a generalized Cauchy-Kovalevskaya form (which restricts a system from possessing any differential identities).

Furthermore, as another result, a large class of multipliers that captures physically important conserved integrals such as mass, momentum, energy, angular momentum is identified for general PDE systems by examining the numerous examples of conservation laws presented earlier.

In Sect. 6, the variational calculus tools are used to state Noether's theorem in a modern form for variational PDE systems, along with the determining equations for variational symmetries. The generalization of Noether's theorem in modern form to non-variational PDE systems is explained in Sect. 7. First, the determining equations for multipliers are shown to be an augmented, adjoint counterpart of the determining equations for symmetries. More precisely, the multiplier determining system has a natural division into two subsystems [6–8]. One subsystem is the adjoint of the symmetry determining system, whose solutions can be viewed as adjoint-symmetries (also known as cosymmetries). The remaining subsystem comprises equations that are necessary and sufficient for an adjoint-symmetry to be a multiplier, analogously to the conditions required for an infinitesimal symmetry to be a variational symmetry in the case of a variational PDE system. Next, the role of a Lagrangian in constructing a conserved integral from a symmetry of a variational principle is replaced for non-variational PDE systems by several different constructions: an explicit integral formula, an explicit algebraic scaling formula, and a system of determining equations, all of which use only a multiplier and the PDE system itself. The scaling formula is based on dimensional analysis and generalizes a formula previously derived only for PDE systems that admit a scaling symmetry [9].

These main results cover both the case of PDE systems without differential identities and the case of PDE systems with differential identities. It is emphasized that this general method for explicitly deriving the conservation laws of PDE systems reproduces the content of Noether's theorem whenever a PDE system has a variational principle. (For comparison, an abstract, cohomological approach to determining conservation laws of PDE systems can be found in Ref. [17–19].)

Some concluding remarks, including discussion of the geometrical meaning of adjoint-symmetries and multipliers, are provided in Sect. 8.

Several running examples will be used to illustrate the main ideas and the main results in every section.

2 Examples

The following seven examples illustrate some basic conserved densities and fluxes (2) arising in physical PDE systems in one spatial dimension.

Ex 1 transport equation

$$u_t = (c(x,u)u)_x \tag{11}$$

$$T = u \text{ is mass density,} \quad X = -c(x,u)u \text{ is mass flux i.e., momentum.} \tag{12}$$

Ex 2 diffusion/heat conduction equation

$$u_t = (k(x,u)u_x)_x \tag{13}$$

$$T = u \text{ is heat density (temperature)}, \quad X = -k(x,u)u_x \text{ is heat flux.} \tag{14}$$

Ex 3 telegraph equation

$$u_{tt} + a(t)u_t - (c(x)^2 u_x)_x = 0 \tag{15}$$

$$\begin{aligned} &T = \tfrac{1}{2}\exp(2\textstyle\int a(t)dt)({u_t}^2 + c(x)^2 {u_x}^2) \text{ is energy density,} \\ &X = -c(x)^2 \exp(2\textstyle\int a(t)dt)u_x u_t \text{ is energy flux.} \end{aligned} \tag{16}$$

Ex 4 nonlinear dispersive wave equation

$$u_t + f(u)u_x + u_{xxx} = 0, \quad f(u) \neq \text{const.} \tag{17}$$

$$T = u \text{ is mass density}, \quad X = \textstyle\int f(u)du + u_{xx} \text{ is mass flux i.e., momentum;} \tag{18a}$$

$$\begin{aligned} &T = u^2 \text{ is elastic energy density,} \\ &X = 2\textstyle\int uf(u)du + 2uu_{xx} - {u_x}^2 \text{ is elastic energy flux;} \end{aligned} \tag{18b}$$

$$\begin{aligned} &T = \textstyle\int g(u)du - \tfrac{1}{2}{u_x}^2 \text{ is gradient energy density,} \\ &X = \tfrac{1}{2}(g(u) + u_{xx})^2 + u_x u_t \text{ is gradient energy flux,} \\ &\quad g(u) = \textstyle\int f(u)du. \end{aligned} \tag{18c}$$

Ex 5 compressible viscous fluid equations

$$\begin{aligned} &\rho_t + (u\rho)_x = 0 \\ &\rho(u_t + uu_x) = -p_x + \mu u_{xx} \end{aligned} \tag{19}$$

$$T = \rho \text{ is mass density}, \quad X = u\rho \text{ is mass flux;} \tag{20a}$$

$$T = \rho u \text{ is momentum density}, \quad X = p - \mu u_x + Tu \text{ is momentum flux;} \tag{20b}$$

$$\begin{aligned} &T = \rho(tu - x) \text{ is Galilean momentum density,} \\ &X = t(p - \mu u_x) + Tu \text{ is Galilean momentum flux.} \end{aligned} \tag{20c}$$

The next two examples are integrable PDE systems that possess an infinite hierarchy of higher-order conservation laws.

Ex 6 barotropic gas flow/compressible inviscid fluid equations

$$\begin{aligned}&\rho_t + (u\rho)_x = 0\\ &u_t + uu_x = -p_x/\rho\\ &p = p(\rho) \text{ (barotropic equation of state)}\\ &e = \textstyle\int p/\rho^2 d\rho \text{ (thermodynamic energy)}\end{aligned} \tag{21}$$

$$T = \rho(\tfrac{1}{2}u^2 + e) \text{ is energy density,} \quad X = (p + T)u \text{ is energy flux;} \tag{22a}$$

$$\begin{aligned}&T = \rho_x/(u_x^2 - p'\rho_x^2/\rho^2) \text{ is higher-derivative quantity,}\\ &X = \rho u_x/(u_x^2 - p'\rho_x^2/\rho^2) \text{ is higher-derivative flux.}\end{aligned} \tag{22b}$$

Ex 7 breaking wave (Camassa-Holm) equation

$$m_t + 2u_x m + um_x = 0, \quad m = u - u_{xx} \tag{23}$$

$$\begin{aligned}&T = m \text{ is momentum density,}\\ &X = \tfrac{1}{2}(u^2 - u_x^2) + um \text{ is momentum flux;}\end{aligned} \tag{24a}$$

$$T = \tfrac{1}{2}(u^2 + u_x^2) \text{ is energy density,} \quad X = u(um - u_{tx}) \text{ is energy flux;} \tag{24b}$$

$$\begin{aligned}&T = \tfrac{1}{2}u(u^2 + u_x^2), \text{ is energy-momentum density,}\\ &X = \tfrac{1}{2}(u_{tx} - u(um + \tfrac{1}{2}u) + \tfrac{1}{2}u_x^2)^2 - u_t(uu_x + \tfrac{1}{2}u_t) \text{ is energy-momentum flux;}\end{aligned} \tag{24c}$$

$$T = m^{1/2} \text{ is Hamiltonian Casimir,} \quad X = 2um^{1/2} \text{ is Casimir flux;} \tag{24d}$$

$$\begin{aligned}&T = m^{-5/2}m_x^2 + 4m^{-1/2} \text{ is higher-derivative energy density,}\\ &X = -m^{-5/2}(2m_t + um_x)m_x - 4m^{-3/2}u_x m_x - 4m^{-1/2}u - 8m^{1/2}\\ &\qquad \text{is higher-derivative flux.}\end{aligned} \tag{24e}$$

The following three examples illustrate some intrinsically multi-dimensional conservation laws (4) that arise in physical PDE systems in two or more dimensions.

Ex 8 porous media equation

$$u_t = \nabla \cdot (k(u)\nabla u) \tag{25}$$

$$\begin{aligned}&T = \alpha(x)u \text{ is a general mass-density moment,}\\ &\mathbf{X} = \textstyle\int k(u)du\nabla\alpha(x) - \alpha(x)\nabla\int k(u)du \text{ is flux moment of mass-density,}\\ &\Delta\alpha = 0 \quad \text{(arbitrary solution of Laplace equation).}\end{aligned} \tag{26}$$

Ex 9 non-dispersive wave equation

$$u_{tt} - c^2 \Delta u = f(u) \tag{27}$$

$$\begin{aligned}
&T = u_t(\mathbf{a}\cdot\mathbf{x})\cdot\nabla u \text{ is angular momentum density,}\\
&\mathbf{X} = (\tfrac{1}{2}c^2|\nabla u|^2 - \tfrac{1}{2}u_t^2 - \textstyle\int f(u)du)(\mathbf{a}\cdot\mathbf{x}) - c^2((\mathbf{a}\cdot\mathbf{x})\cdot\nabla u)\nabla u\\
&\qquad\qquad \text{is angular momentum flux,}\\
&\quad \mathbf{a}\cdot\mathbf{x} \quad \text{(arbitrary constant antisymmetric tensor } \mathbf{a}\text{).}
\end{aligned} \tag{28a}$$

$$\begin{aligned}
&T = \mathbf{b}\cdot\mathbf{x}(\tfrac{1}{2}u_t^2 + \tfrac{1}{2}c^2|\nabla u|^2 - \textstyle\int f(u)du) + c^2tu_t\mathbf{b}\cdot\nabla u \text{ is boost momentum density,}\\
&\mathbf{X} = c^2t(\tfrac{1}{2}c^2|\nabla u|^2 - \tfrac{1}{2}u_t^2 - \textstyle\int f(u)du)\mathbf{b} - c^2(\mathbf{b}\cdot\mathbf{x}u_t + c^2t\mathbf{b}\cdot\nabla u)\nabla u\\
&\qquad\quad \text{is boost momentum flux,}\\
&\quad \text{(arbitrary constant antisymmetric vector } \mathbf{b}\text{).}
\end{aligned} \tag{28b}$$

Ex 10 inviscid (compressible/incompressible) fluid equation

$$\begin{aligned}
&\mathbf{u}_t + \mathbf{u}\cdot\nabla\mathbf{u} = -(1/\rho)\nabla p\\
&e = \textstyle\int p/\rho^2 d\rho \text{ (thermodynamic energy)}
\end{aligned} \tag{29}$$

$$\text{in three dimensions} \begin{cases} T = \mathbf{u}\cdot(\nabla\times\mathbf{u}) \text{ is local helicity,}\\ \mathscr{X} = \boldsymbol{X} - \boldsymbol{Tu} = \tfrac{1}{2}(|\mathbf{u}|^2 + (p/\rho) + e) \text{ is moving helicity flux;} \end{cases} \tag{30a}$$

$$\text{in two dimensions} \begin{cases} T = \rho f((\mathrm{curl}\,\mathbf{u})/\rho) \text{ is local enstrophy,}\\ \mathscr{X} = \boldsymbol{X} - \boldsymbol{Tu} = 0 \text{ is moving enstrophy flux,}\\ \quad \text{(arbitrary function} f\text{).} \end{cases} \tag{30b}$$

The last four examples illustrate spatial boundary conservation laws (5) and spatial flux conservation laws (10) for physical PDE systems in three dimensions.

Ex 11 electric (displacement) field equation inside matter

$$\begin{aligned}
&\mathbf{D}_t = c\nabla\times\mathbf{H}\\
&\nabla\cdot\mathbf{D} = 4\pi\rho\\
&\mathbf{J} = 0 \text{ (no currents)}\\
&\rho_t = 0 \text{ (static charges)}
\end{aligned} \tag{31}$$

$$\mathbf{T} = \mathbf{D} \text{ is flux density of electric field lines.} \tag{32}$$

Ex 12 magnetohydrodynamics (infinite conductivity) equations

$$\begin{aligned}
&\mathbf{u}_t + \mathbf{u}\cdot\nabla\mathbf{u} = (1/\rho)(\mathbf{J}\times\mathbf{B} - \nabla p)\\
&\mathbf{B}_t = \nabla\times(\mathbf{u}\times\mathbf{B})\\
&\nabla\times\mathbf{B} = 4\pi\mathbf{J}\\
&\nabla\cdot\mathbf{B} = 0
\end{aligned} \tag{33}$$

$$\mathbf{T} = \mathbf{X} = \mathbf{B} \text{ is flux density of magnetic field lines.} \tag{34}$$

Ex 13 fluid incompressibility equation

$$\nabla\cdot\mathbf{u} = 0 \tag{35}$$

$$\mathbf{X} = \mathbf{u} \text{ is flux density of streamlines.} \tag{36}$$

Ex 14 charge source equation (in empty space)

$$\nabla\cdot\mathbf{E} = 0 \tag{37}$$

$$\mathbf{X} = \mathbf{E} \text{ is flux density of electric field lines.} \tag{38}$$

3 Conserved Integrals, Conservation Laws, and Symmetries

Throughout, the following notation will be used. Let t, $x = (x^1, \ldots, x^n)$ be independent variables, $n \geq 1$, and let $u = (u^1, \ldots, u^m)$ be dependent variables, $m \geq 1$. Partial derivatives of u with respect to t, x are denoted $\partial u = (u_t, u_{x^1}, \ldots, u_{x^n})$, and kth-order partial derivatives are denoted $\partial^k u$, $k \geq 2$. The coordinate space $J = (t, x, u, \partial u, \partial^2 u, \ldots)$ is called the *jet space* associated with the variables t, x, u. Partial derivatives with respect to these variables are given by $\partial/\partial t$, $\partial/\partial x = (\partial/\partial x^1, \ldots, \partial/\partial x^n)^{\mathrm{t}}$, $\partial/\partial u = (\partial/\partial u^1, \ldots, \partial/\partial u^m)^{\mathrm{t}}$, with a superscript "t" denoting the transpose, and similarly for partial derivatives with respect to the derivative variables in J. Total derivatives with respect to t, x, acting by the chain rule, are denoted $D = (D_t, D_{x^1}, \ldots, D_{x^n})$. In particular, $Du = \partial u$, $D\partial u = \partial^2 u$, and so on. D^k denotes all of the kth order total derivatives with respect to t, x. Spatial divergences are denoted Div $= D_x\cdot$, with a dot denoting the vector dot product.

Consider an Nth-order system of $M \geq 1$ PDEs

$$G = (G^1(t, x, u, \partial u, \ldots, \partial^N u), \ldots, G^M(t, x, u, \partial u, \ldots, \partial^N u)) = 0. \tag{39}$$

The space of all locally smooth solutions $u(t,x)$ of the system will be denoted $\mathscr{E}$. This space has an embedding as a subspace in J, since $u(t,x) \in \mathscr{E}$ determines $(t,x,u(t,x),\partial u(t,x),\partial^2 u(t,x),\ldots) \in J$. (In the applied mathematics and physics literature, $\mathscr{E}$ is commonly identified with the set of equations $G = 0, DG = 0, D^2G = 0,\ldots$ in J, which assumes these equations are locally solvable [1].)

A *local conservation law* of a given PDE system (39) is a local continuity equation

$$(D_tT + D_x \cdot X)|_{\mathscr{E}} = 0 \tag{40}$$

which holds on the whole solution space $\mathscr{E}$ of the system, where $T(t,x,u,\partial u,\ldots,\partial^r u)$ is the *conserved density* and $X = (X^1(t,x,u,\partial u,\ldots,\partial^r u),\ldots,X^n(t,x,u,\partial u,\ldots,\partial^r u))$ is the *spatial flux*. The pair

$$(T,X) = \Phi \tag{41}$$

is called a *conserved current*.

Every conservation law (40) can be integrated over any given spatial domain $\Omega \subseteq \mathbb{R}^n$ to get

$$\frac{d}{dt}\int_{\Omega} T|_{\mathscr{E}} dV = -\oint_{\partial\Omega} X|_{\mathscr{E}} \cdot \nu dA \tag{42}$$

by the divergence theorem, where $\partial\Omega$ is the boundary of the domain and ν denotes the outward pointing unit normal vector. This shows that the rate of change of the quantity

$$\mathscr{C}[u] = \int_{\Omega} T|_{\mathscr{E}} dV \tag{43}$$

in the domain is balanced by the net flux escaping through the domain boundary. The quantity (43) is called a *conserved integral*, and the relation (42) is called a *global conservation law* or *global balance equation*.

Two conservation laws are *locally equivalent* if they give the same *global balance equation* (42) up to boundary terms. This occurs iff their conserved densities differ by a total spatial divergence $D_x \cdot \Theta$ on the solution space $\mathscr{E}$, and correspondingly, their fluxes differ by a total time derivative $-D_t\Theta$ modulo a divergence-free vector. A conservation law is thereby called *locally trivial* if

$$T_{\text{triv}}|_{\mathscr{E}} = D_x \cdot \Theta|_{\mathscr{E}}, \quad X_{\text{triv}}|_{\mathscr{E}} = -D_t\Theta|_{\mathscr{E}} + D_x \cdot \Gamma|_{\mathscr{E}} \tag{44}$$

holds for some vector function $\Theta(t,x,u,\partial u,\ldots,\partial^{r-1}u)$ and some antisymmetric tensor function $\Gamma(t,x,u,\partial u,\ldots,\partial^{r-1}u)$. The *differential order of a conservation law* is defined to be the smallest differential order among all locally equivalent conserved currents. (It is common in the mathematics literature to define a local conservation law itself as the equivalence class of locally equivalent conserved currents.)

The global form of a locally trivial conservation law is given by

$$\frac{d}{dt}\oint_{\partial\Omega} \Theta|_{\mathscr{E}} \cdot \nu dA = \oint_{\partial\Omega} D_t\Theta|_{\mathscr{E}} \cdot \nu dA \tag{45}$$

since $\oint_{\partial\Omega}(D_x \cdot M)|_{\mathscr{E}} \cdot \nu dA = 0$ by Stokes' theorem. This integral equation (45) is just an identity, with no physical content, unless the spartial flux of $D_t\Theta|_{\mathscr{E}}$ vanishes. From the divergence theorem, this integral will vanish for all domains Ω iff $D_x \cdot D_t\Theta|_{\mathscr{E}} = 0$ holds. In such cases, the boundary integral

$$\int_{\Omega} T|_{\mathscr{E}} dV = \oint_{\partial\Omega} \Theta|_{\mathscr{E}} \cdot \nu dA \tag{46}$$

will be a constant of motion for solutions of the given PDE system. This type of boundary conservation law arises for PDE systems typically when the PDEs in the system are related by obeying a differential identity, as will be discussed further in Sect. 5. In all cases when both $D_x \cdot \Theta|_{\mathscr{E}}$ and $D_x \cdot D_t\Theta|_{\mathscr{E}}$ do not vanish identically, a locally trivial conservation law has no physical content.

For a given PDE system (39), the set of all non-trivial conservation laws (up to local equivalence) forms a vector space on which the symmetries of the system have a natural action [1, 3].

An *infinitesimal symmetry* [1–3] of a given PDE system (39) is a generator

$$\mathbf{X} = \tau\partial/\partial t + \xi\partial/\partial x + \eta\partial/\partial u \tag{47}$$

whose prolongation leaves invariant the PDE system,

$$\mathrm{pr}\mathbf{X}(G)|_{\mathscr{E}} = 0 \tag{48}$$

which holds on the whole solution space $\mathscr{E}$ of the system. Here $\tau(t, x, u, \partial u, \ldots, \partial^r u)$, $\xi = (\xi^1(t, x, u, \partial u, \ldots, \partial^r u), \ldots, \xi^n(t, x, u, \partial u, \ldots, \partial^r u))$, and $\eta = (\eta^1(t, x, u, \partial u, \ldots, \partial^r u), \ldots, \eta^m(t, x, u, \partial u, \ldots, \partial^r u))$ are called the *characteristic functions* in the symmetry generator. When acting on the solution space $\mathscr{E}$, an infinitesimal symmetry generator can be formally exponentiated to produce a one-parameter group of transformations $\exp(\epsilon\mathrm{pr}\mathbf{X})$, with parameter ϵ, where the infinitesimal transformation is given by

$$\begin{aligned} u(t,x) \to u(t,x) + \epsilon\big(&\eta(t, x, u(t,x), \partial u(t,x), \ldots, \partial^r u(t,x)) \\ &- u_t(t,x)\tau(t, x, u(t,x), \partial u(t,x), \ldots, \partial^r u(t,x)) \\ &- u_x(t,x) \cdot \xi(t, x, u(t,x), \partial u(t,x), \ldots, \partial^r u(t,x))\big) + O\big(\epsilon^2\big) \end{aligned} \tag{49}$$

for all solutions $u(t, x)$ of the PDE system.

Two infinitesimal symmetries are equivalent if they have the same action on the solution space $\mathscr{E}$ of a given PDE system. An infinitesimal symmetry is thereby called *trivial* if it leaves all solutions $u(t, x)$ unchanged. This occurs iff its characteristic functions satisfy the relation

$$\eta|_{\mathscr{E}} = (u_t\tau + u_x \cdot \xi)|_{\mathscr{E}}. \tag{50}$$

The corresponding generator (47) of a trivial symmetry on the solution space $\mathscr{E}$ is thus given by

$$\mathbf{X}_{\text{triv}} = \tau\partial/\partial t + \xi \cdot \partial/\partial x + (u_t\tau + u_x \cdot \xi)\partial/\partial u \tag{51}$$

which has the prolongation $\text{pr}\mathbf{X}_{\text{triv}} = \tau D_t + \xi \cdot D_x$. Conversely, any generator of this form (51) represents a trivial symmetry. The *differential order of an infinitesimal symmetry* is defined to be the smallest differential order among all equivalent generators.

In jet space J, a group of transformations $\exp(\epsilon\text{pr}\mathbf{X})$ in general will not act in a closed form on t, x, u, and derivatives $\partial^k u$ up to a finite order, except [1, 3] for point transformations acting on (t, x, u), and contact transformations acting on (t, x, u, u_t, u_x). Moreover, a contact transformation is a prolonged point transformation when the number of dependent variables is $m = 1$ [1, 3]. A *point symmetry* is defined as a symmetry transformation group on (t, x, u), whose generator is given by characteristic functions of the form

$$\mathbf{X} = \tau(t, x, u)\partial/\partial t + \xi(t, x, u)\partial/\partial x + \eta(t, x, u)\partial/\partial u \tag{52}$$

corresponding to the infinitesimal point transformation

$$\begin{aligned} t &\to t + \epsilon\tau(t, x, u) + O(\epsilon^2), \\ x &\to x + \epsilon\xi(t, x, u) + O(\epsilon^2), \\ u &\to u + \epsilon\eta(t, x, u) + O(\epsilon^2). \end{aligned} \tag{53}$$

Likewise, a *contact symmetry* is defined as a symmetry transformation group on (t, x, u, u_t, u_x) whose generator corresponds to an infinitesimal transformation that preserves the contact relations $u_t = \partial_t u$, $u_x = \partial_x u$. The set of all admitted point symmetries and contact symmetries for a given PDE system comprises its group of *Lie symmetries.*

Common examples of point symmetries admitted by PDE systems arising in physical applications are time translations, space translations, and scalings. Higher-order symmetries are typically admitted only by integrable PDE systems. However, it is worth emphasizing that any admitted symmetry can be used to obtain a mapping of a given solution $u = f(t, x)$ of a PDE system into a one-parameter family of solutions $u = \tilde{f}(t, x, \epsilon) = \big(\exp(\epsilon\text{pr}\hat{\mathbf{X}})u\big)|_{u=f(t,x)} = \big(u + \epsilon\hat{\eta} + \frac{1}{2}\epsilon^2 pr\hat{\boldsymbol{X}}\hat{\eta} + \cdots\big)|_{u=f(t,x)}$, where $\hat{X} = X - X_{\text{triv}} = \hat{\eta}\partial/\partial u$; and also to find symmetry-invariant solutions $u = f(t,x)$ of a PDE system by considering the invariance condition $(\hat{X}u)|_{u=f(t,x)} = \hat{\eta}|_{u=f(t,x)} = 0$. Thus, for these two main purposes, symmetries of any differential order are equally useful.

Similar remarks can be made for conservation laws. In physical applications, the most common examples of conserved densities admitted by PDE systems are mass, momentum, and energy. These densities are always of low differential order, whereas higher-order densities are typically admitted only by integrable PDE systems. Nevertheless, for the many purposes outlined in Sect. 1, any admitted conservation law of a given PDE system can be useful.

3.1 Regular PDE Systems and Computation of Symmetry Generators, Conserved Densities and Fluxes

To determine if a current (41) is conserved for a given PDE system, and if a generator (47) is an infinitesimal symmetry of a given PDE system, it is necessary to coordinatize the solution space $\mathscr{E}$ of the system in jet space J. This can be accomplished in a general way by the following steps. First, for any PDE system (39), introduce an index notation for the components of x and u: x^i, $i = 1,\ldots,n$; and u^α, $\alpha = 1,\ldots,m$. Next, suppose each PDE $G^a = 0$, $a = 1,\ldots,M$, in the given system can be expressed in a solved form

$$G^a = \partial_{(\ell_a)}u^{\alpha_a} - g^a \tag{54}$$

for some derivative of a single dependent variable u^{α_a}, after a point transformation (change of variables) if necessary, such that all other terms in the system contain neither this derivative nor its differential consequences, namely

$$\begin{aligned} \partial_{(\ell_a)}u^{\alpha_a} \neq \partial^k\partial_{(\ell_b)}u^{\alpha_b}, \quad a,b = 1,\ldots,M, \quad k \geq 1, \\ \frac{\partial g^a}{\partial(\partial^k\partial_{(\ell_b)}u^{\alpha_b})} = 0, \quad a,b = 1,\ldots,M, \quad k \geq 0. \end{aligned} \tag{55}$$

Such derivatives $\{\partial_{(\ell_a)}u^{\alpha_a}\}_{a=1,\ldots,M}$ are called a set of *leading derivatives* for the PDE system. Last, suppose the given PDE system is closed in the sense that it has no integrability conditions and all of its differential consequences produce PDEs that have a solved form in terms of differential consequences of the leading derivatives. Note that if a PDE system is not closed then it can always be enlarged to get a closed system by appending any integrability conditions and differential consequences that involve the introduction of more leading derivatives. Then, coordinates for the solution space $\mathscr{E}$ of the closed PDE system in J are provided by the independent variables t, x^i, the dependent variables u^α, and all of the non-leading derivatives of u^α. A closed PDE system (39) admitting such a solved form (54)–(55) will be called *regular*.

A more restrictive class of PDE systems is given by Cauchy-Kovalevskaya systems and their generalizations. Recall, a PDE system (39) is of Cauchy-Kovalevskaya form [1, 20] if the leading derivatives in the solved form of the system consist of pure derivatives of u with respect to a single independent variable, namely $\partial_{(\ell_a)} u^{\alpha_a} = \partial_z^{k_a} u^{\alpha_a}$, $a = 1, \ldots, M$, $z \in \{t, x^i\}$, and if their differential order k_a is equal to the differential order N of the system, namely $k_a = N$, $a = 1, \ldots, M$. Cauchy-Kovalevskaya systems, and their generalizations [4] in which k_a differs from N, have the feature that they do not possess any differential identities and that none of their differential consequences possess differential identities. Such PDE systems are usually called *normal*. Note that, in contrast to normal systems, the leading derivatives in a regular PDE system can be, for instance, a mixed derivative of all the dependent variables u^α or a different derivative of each of the dependent variables u^α.

Running Ex. (1) Generalized Korteweg-de Vries (gKdV) equation

$$u_t + u^p u_x + u_{xxx} = 0, \quad p > 0. \tag{56}$$

This is a regular PDE since it has the leading derivative $u_t = -u^p u_x - u_{xxx}$. It also has a third-order leading derivative $u_{xxx} = -u_t - u^p u_x$. Both of these solved forms are of generalized Cauchy-Kovalevskaya type.

Running Ex. (2) Breaking wave equation [21]

$$m_t + bu_x m + um_x = 0, \quad m = u - u_{xx}, \quad b \neq -1. \tag{57}$$

This is a regular PDE system since it has the leading derivatives $m_t = -bu_x m - um_x$, $u_{xx} = u - m$. Equivalently, if m is eliminated through the second PDE, this yields a scalar equation $u_t - u_{txx} + (b+1)uu_x = bu_x u_{xx} + uu_{xxx}$ which is a regular PDE with respect to the leading derivative

$$u_{txx} = u_t + (b+1)uu_x - bu_x u_{xx} - uu_{xxx}. \tag{58}$$

Neither of these solved forms are of generalized Cauchy-Kovalevskaya type. However, the alternative solved forms $u_{xxx} = u_x + u^{-1}(bu_x(u - u_{xx}) - u_t + u_{txx})$ and $m_x = -u^{-1}(m_t + bu_x m)$, $u_{xx} = u - m$ are of generalized Cauchy-Kovalevskaya type.

Running Ex. (3) Euler equations for constant density, inviscid fluids in two dimensions

$$\begin{aligned} \nabla \cdot \mathbf{u} = 0, \quad \rho = \text{const.}, \\ \mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} = -(1/\rho)\nabla p, \\ \Delta p = -\rho(\nabla \mathbf{u}) \cdot (\nabla \mathbf{u})^{\mathrm{t}}. \end{aligned} \tag{59}$$

In this system, the independent variables are t and (x, y), and the dependent variables consist of p and $\mathbf{u} = (u^1, u^2)$, in Cartesian components. Leading derivatives are given by writing the PDEs in the solved form

$$\begin{aligned}
u^1_x &= -u^2_y,\\
u^1_t &= -(u^1 u^1_x + u^2 u^1_y + (1/\rho)p_x)\big|_{u^1_x=-u^2_y} = -(u^2 u^1_y - u^1 u^2_y + +(1/\rho)p_x),\\
u^2_t &= -(u^1 u^2_x + u^2 u^2_y + (1/\rho)p_y),\\
p_{xx} &= -p_{yy} - \rho((u^1_x)^2 + (u^2_y)^2 + 2u^1_y u^2_x)\big|_{u^1_x=-u^2_y} = -p_{yy} - 2\rho((u^2_y)^2 + u^1_y u^2_x).
\end{aligned}$$

Thus, this system is a regular PDE system, but it does not have a generalized Cauchy-Kovalevskaya form. A related feature is that the PDEs in the system obey a differential identity

$$\mathrm{Div}\,(\mathbf{u}_t + \mathbf{u}\cdot\nabla\mathbf{u} + (1/\rho)\nabla p) - (D_t + \mathbf{u}\cdot\nabla)(\nabla\cdot\mathbf{u}) = (1/\rho)\mathit{\Delta} p + (\nabla\mathbf{u})\cdot(\nabla\mathbf{u})^{\mathrm{t}}. \tag{60}$$

Note that the pressure equation is often not explicitly considered in writing down the Euler equations. However, without including the pressure equation, the system would not be closed, since the differential identity (60) shows that the pressure equation arises as an integrability condition of the other equations. Correspondingly, the pressure equation does not have a solved form in terms of the set of derivatives $\{u^1_x, u^1_t, u^2_t\}$.

Running Ex. (4) Magnetohydrodynamics equations for a compressible, infinite conductivity plasma in three dimensions

$$\begin{aligned}
p = P(\rho), \quad \nabla\times\mathbf{B} = 4\pi\mathbf{J}, \quad \nabla\cdot\mathbf{B} &= 0,\\
\rho_t + \nabla\cdot(\rho\mathbf{u}) &= 0,\\
\mathbf{u}_t + \mathbf{u}\cdot\nabla\mathbf{u} &= (1/\rho)(\mathbf{J}\times\mathbf{B} - \nabla p),\\
\mathbf{B}_t &= \nabla\times(\mathbf{u}\times\mathbf{B}).
\end{aligned} \tag{61}$$

The independent variables in this system are t and (x, y, z), and the dependent variables consist of ρ, $\mathbf{u} = (u^1, u^2, u^3)$ and $\mathbf{B} = (B^1, B^2, B^3)$, in Cartesian components. This is a regular PDE system, where a set of leading derivatives is given by writing the PDEs in the solved form

$$\begin{aligned}
B^1_x &= -B^2_y - B^3_z,\\
\rho_t &= \rho(u^1_x + u^2_y + u^3_z) + \rho_x u^1 + \rho_y u^2 + \rho_z u^3,\\
u^1_t &= -(u^1 u^1_x + u^2 u^1_y + u^3 u^1_z + (1/\rho)(P'(\rho)\rho_x + B^2 J^3 - B^3 J^2)),\\
u^2_t &= -(u^1 u^2_x + u^2 u^2_y + u^3 u^2_z + (1/\rho)(P'(\rho)\rho_y + B^3 J^1 - B^1 J^3)),\\
u^3_t &= -(u^1 u^3_x + u^2 u^3_y + u^3 u^3_z + (1/\rho)(P'(\rho)\rho_z + B^1 J^2 - B^2 J^1)),
\end{aligned}$$

$$\begin{aligned}
B^1_t &= u^1B^2_y - u^2B^1_y + u^1B^3_z - u^3B^1_z + u^1_yB^2 - u^2_yB^1 + u^1_zB^3 - u^3_zB^1,\\
B^2_t &= (u^2B^3_z - u^3B^2_z + u^2B^1_x - u^1B^2_x + u^2_zB^3 - u^3_zxB^2 + u^2_xB^1 - u^1_xB^2)\big|_{B^1_x=-B^2_y-B^3_z}\\
&= -u^2B^2_y - u^3B^2_z - u^1B^2_x + u^2_zB^3 - u^3_zB^2 + u^2_xB^1 - u^1_xB^2,\\
B^3_t &= (u^3B^1_x - u^1B^3_x + u^3B^2_y - u^2B^3_y + u^3_xB^1 - u^1_xB^3 + u^3_yB^2 - u^2_yB^3)\big|_{B^1_x=-B^2_y-B^3_z}\\
&= -u^3B^3_z - u^1B^3_x - u^2B^3_y + u^3_xB^1 - u^1_xB^3 + u^3_yB^2 - u^2_yB^3,
\end{aligned}$$

with

$$4\pi J^1 = B^3_y - B^2_z, \quad 4\pi J^2 = B^1_z - B^3_x, \quad 4\pi J^3 = B^2_x - B^1_y.$$

These PDEs lack a generalized Cauchy-Kovalevskaya form, which is related to the feature that they obey a differential identity

$$\mathrm{Div}\,(\mathbf{B}_t - \nabla \times (\mathbf{u} \times \mathbf{B})) = D_t(\nabla \cdot \mathbf{B}). \tag{62}$$

As seen from the examples here and in Sect. 2, all typical PDE systems arising in physical applications belong to the class of regular systems.

For any given regular PDE system, the standard approach [22–25] to look for symmetries consists of solving the invariance condition $\mathrm{pr}\mathbf{X}(G)|_{\mathscr{E}} = 0$ to find the characteristic functions η, τ, ξ in the generator $\mathbf{X}$. The computations in this approach are reasonable for finding point symmetries, but become much more complicated for finding contact symmetries and higher-order symmetries.

Running Ex. (1) Consider the gKdV equation (56). Since this is a scalar PDE, its Lie symmetries are generated by point transformations and contact transformations, with the general infinitesimal form

$$\mathbf{X} = \tau(t,x,u,u_t,u_x)\partial/\partial t + \xi(t,x,u,u_t,u_x)\partial/\partial x + \eta(t,x,u,u_t,u_x)\partial/\partial u.$$

Substitution of this generator into the determining condition $\mathrm{pr}\mathbf{X}(u_t + u^p u_x + u_{xxx})|_{\mathscr{E}} = 0$ requires prolonging $\mathbf{X}$ to first-order with respect to t and third-order with respect to x:

$$\mathrm{pr}\mathbf{X} = \mathbf{X} + \eta^{(t)}\partial/\partial u_t + \eta^{(x)}\partial/\partial u_x + D^2_x\eta^{(xx)}\partial/\partial u_{xx} + D^3_x\eta^{(xxx)}\partial/\partial u_{xxx}$$

where

$$\begin{aligned}
\eta^{(t)} &= D_t\eta - u_tD_t\tau - u_xD_t\xi,\\
\eta^{(x)} &= D_x\eta - u_tD_x\tau - u_xD_x\xi,\\
\eta^{(xx)} &= D_x\eta^{(x)} - u_{tx}D_x\tau - u_{xx}D_x\xi,\\
\eta^{(xxx)} &= D_x\eta^{(xx)} - u_{txx}D_x\tau - u_{xxx}D_x\xi.
\end{aligned}$$

This yields

$$\big(u_x\eta + D_t\eta - u_tD_t\tau - u_xD_t\xi + uD_x\eta - (uu_t + 3u_{txx})D_x\tau - (uu_x + 3u_{xxx})D_x\xi \\ - 3u_{tx}D_x^2\tau - 3u_{xx}D_x^2\xi + D_x^3\eta - u_tD_x^3\tau - u_xD_x^3\xi\big)|_{\mathscr{E}} = 0.$$

There are two steps in solving this determining condition. First, since the condition is formulated on the gKdV solution space $\mathscr{E}$, a leading derivative of u (and all of its differential consequences) needs to be eliminated. The most convenient choice is $u_{xxx} = -u_t - u^pu_x$ rather than $u_t = -u^pu_x - u_{xxx}$, since τ, ξ, η depend on u_t. Next, after the total derivatives of τ, ξ, η are expanded out, the resulting equation needs to be split with respect to the jet variables $u_{tt}, u_{tx}, u_{xx}, u_{txx}, u_{xxx}, u_{txxx}, u_{xxxx}$ which do not appear in τ, ξ, η. Finally, the split equations need to be simplified, as some are differential consequences of others. After these lengthy computations and simplifications, a linear system of 6 determining equations is obtained for τ, ξ, η:

$$\begin{gathered}
2\tau_{u_t} + u_t\tau_{u_tu_t} + u_x\xi_{u_tu_t} - \eta_{u_tu_t} = 0,\\
2\xi_{u_x} + u_t\tau_{u_xu_x} + u_x\xi_{u_xu_x} - \eta_{u_xu_x} = 0,\\
\tau_{u_x} + \xi_{u_t} + u_t\tau_{u_tu_x} + u_x\xi_{u_tu_x} - \eta_{u_tu_x} = 0,\\
3pu_t\eta + 2u(u_t\tau_t + u_x\xi_t - \eta_t) + 3p\big(u_t^2(u_t\tau_{u_t} + u_x\tau_{u_x})\\
+ u_tu_x(u_t\xi_{u_t} + u_x\xi_{u_x}) - u_t(u_t\eta_{u_t} + u_x\eta_{u_x})\big) = 0,\\
pu_x\eta + 2u(u_t\tau_x + u_x\xi_x - \eta_x) + p\big(u_tu_x(u_t\tau_{u_t} + u_x\tau_{u_x})\\
+ u_x^2(u_t\xi_{u_t} + u_x\xi_{u_x}) - u_x(u_t\eta_{u_t} + u_x\eta_{u_x})\big) = 0,\\
\eta + u(u_t\tau_u + u_x\xi_u - \eta_u) + u_t(u_t\tau_{u_t} + u_x\tau_{u_x})\\
+ u_x(u_t\xi_{u_t} + u_x\xi_{u_x}) - (u_t\eta_{u_t} + u_x\eta_{u_x}) = 0.
\end{gathered}$$

This system can be solved, with p treated as an unknown, to get

$$\tau = \tilde{\tau}(t,x,u,u_t,u_x), \quad \xi = \tilde{\xi}(t,x,u,u_t,u_x),\\
\eta = u_t(\tilde{\tau} - c_1 - 3c_3) + u_x(\tilde{\xi} - c_2 - c_3x - c_4t) - \tfrac{2}{p}c_3u + c_4, \quad c_4 = 0 \text{ if } p \neq 1,$$

which is a linear combination of a time translation (c_1), a space translation (c_2), a scaling (c_3), and a Galilean boost (c_4), plus a trivial symmetry involving two arbitrary functions $\tilde{\tau}(t,x,u,u_t,u_x)$, $\tilde{\xi}(t,x,u,u_t,u_x)$.

Clearly, for finding higher-order symmetries, or for dealing with PDE systems that have a high differential order or that involve more spatial dimensions, the previous standard approach becomes increasingly complicated, as the general solution of the symmetry determining condition will always contain a trivial symmetry involving arbitrary differential functions. In particular, the resulting linear system of determining equations for finding τ, ξ, η becomes less over-determined and hence more computationally difficult to solve when going to higher orders.

The situation for finding conservation laws is quite similar. For any given regular PDE system, it is possible to look for conservation laws by solving the local continuity equation $(D_tT+D_x\cdot X)|_{\mathscr{E}}=0$ to find T and X. This approach is workable when the conserved densities T and fluxes X being sought have a low differential order and when the number of spatial dimensions is low.

Running Ex. (1) Consider again the gKdV equation (56). This is a time evolution PDE of third order in spatial derivatives, while the conserved currents in lowest order form for mass, energy, and L^2 norm are of first order in derivatives for the densities and of second order in derivatives for the fluxes. Substitution of functions

$$T(t,x,u,u_t,u_x),\quad X(t,x,u,u_t,u_x,u_{tt},u_{tx},u_{xx})$$

into the determining condition $(D_tT+D_xX)|_{\mathscr{E}}=0$ yields

$$\begin{aligned}&\big(T_t+u_tT_u+u_{tx}(T_{u_x}+X_{u_t})+u_{tt}T_{u_t}+X_x+u_xX_u+u_{xx}X_{u_x}\\&\quad+u_{txx}X_{u_{tx}}+u_{ttx}X_{u_{tt}}+u_{xxx}X_{u_{xx}}\big)|_{\mathscr{E}}=0.\end{aligned}$$

The steps in solving this determining condition are similar to those used in solving the symmetry determining equation. First, a leading derivative of u (and all of its differential consequences) needs to be eliminated. The most convenient choice is $u_{xxx}=-u_t-u^pu_x$ rather than $u_t=-u^pu_x-u_{xxx}$, since T and X depend on u_t. Next, the resulting equation needs to be split with respect to the jet variables u_{ttx}, u_{txx}, which do not appear in T, X. This splitting immediately leads to a further splitting with respect to u_{tx}, u_{tt}, giving a linear system of 5 PDEs for T, X:

$$\begin{gathered}T_{u_t}=0,\quad X_{u_{tt}}=0,\quad X_{u_{tx}}=0,\quad T_{u_x}+X_{u_t}=0,\\T_t+u_tT_u+X_x+u_xX_u+u_{xx}X_{u_x}-(u_t+u^pu_x)X_{u_{xx}}=0.\end{gathered}$$

This system can be solved, treating p as an unknown, to obtain

$$\begin{aligned}T={}&c_1u^2+c_2u+c_3(\tfrac{1}{(p+1)(p+2)}u^{p+2}-\tfrac{1}{2}u_x{}^2)+c_4(xu-\tfrac{1}{2}tu^2)\\&+c_5(t(\tfrac{1}{2}u^2-3u_x{}^2)-xu^2)+D_x\Theta(t,x,u),\\X={}&c_1(\tfrac{2}{p+2}u^{p+2}+2uu_{xx}-u_x{}^2)+c_2(\tfrac{1}{p+1}u^{p+1}+u_{xx})\\&+c_3(\tfrac{1}{2}(\tfrac{1}{p+1}u^{p+1}+u_{xx})^2+u_xu_t)+c_4(x(\tfrac{1}{2}u^2+u_{xx})-t(\tfrac{1}{3}u^3+uu_{xx}-\tfrac{1}{2}u_x{}^2)-u_x)\\&+c_5(t(3(\tfrac{1}{3}u^3+u_{xx})^2+6u_tu_x)+x(u_x{}^2-2uu_{xx}-\tfrac{1}{2}u^3)+2uu_x)\\&-D_t\Theta(t,x,u),\\&c_4=0\text{ if }p\neq 1,\quad c_5=0\text{ if }p\neq 2\end{aligned}$$

which yields a linear combination of the densities and the fluxes representing conserved currents for the L^2-norm (c_1), mass (c_2), energy (c_3), Galilean momentum (c_4), and Galilean energy (c_5), plus a term involving an arbitrary function $\Theta(t, x, u)$ which represents a locally trivial conserved current.

However, when going to higher orders or to higher spatial dimensions, it becomes increasingly more difficult to solve the local continuity equation $(D_t T + D_x \cdot X)|_{\mathscr{E}} = 0$, as the general solution will contain a trivial density term $D_x \cdot \Theta$ in T and a trivial flux term $-D_t\Theta + D_x \cdot \Gamma$ in X involving a differential vector function Θ and a differential antisymmetric tensor function Γ, which are arbitrary. In particular, the resulting linear system of determining equations for finding T and X will be less over-determined and hence more computationally difficult to solve, compared to the low order case or the one dimensional case.

These difficulties motivate introducing a characteristic form (or canonical representation) for conserved currents so that all locally equivalent conserved currents have the same characteristic form, and likewise for symmetry generators. To derive this formulation, some tools from variational calculus will be needed.

4 Tools in Variational Calculus

For working with symmetries and conservation laws of PDE systems, the natural setting in which to apply variational calculus is the space of *differential functions* defined by locally smooth functions of finitely many variables in jet space $J = (t, x, u, \partial u, \partial^2 u, \ldots)$.

As examples, in the nonlinear dispersive wave equation Ex. 4, if the constitutive nonlinearity function $f(u)$ is smooth, then the conserved density and flux for mass and energy are smooth functions of u, u_x, u_{xx} in J, but if $f(u)$ blows up when $u = 0$ then these functions are singular at points in J such that $u = 0$; in the barotropic gas flow Ex. 6, the higher-derivative density and flux are singular functions of ρ, u, ρ_x, u_x at points in J where $u_x^2 = p(\rho)'/\rho$, but at all other points these functions are smooth.

The basic tools that will be needed from variational calculus are the Fréchet derivative and adjoint derivative, the Euler operator, a homotopy integral, a total null-divergence identity, and a scaling identity. Throughout, $f(t, x, u, \partial u, \ldots, \partial^k u)$ denotes a differential function of order $k \geq 0$, and $v = (v^1(t, x, u, \partial u, \partial^2 u, \ldots), \ldots, v^m(t, x, u, \partial u, \partial^2 u, \ldots))$, $w(t, x, u, \partial u, \partial^2 u, \ldots)$ denote differential functions of arbitrary finite order.

The *Fréchet derivative* of a differential function is the linearization of the function as defined by

$$\begin{aligned} \delta_v f &= \frac{\partial}{\partial \epsilon} f(t, x, u + \epsilon v, \partial(u + \epsilon v), \ldots, \partial^k (u + \epsilon v))\big|_{\epsilon=0} \\ &= v \frac{\partial f}{\partial u} + Dv \cdot \frac{\partial f}{\partial(\partial u)} + \cdots + D^k v \cdot \frac{\partial f}{\partial(\partial^k u)} \end{aligned} \tag{63}$$

which can be viewed as a local directional derivative in jet space, corresponding to the action of a generator $\hat{\mathbf{X}} = v\partial_u$ in characteristic form, $\hat{\mathbf{X}}(f) = \delta_v f$. It is useful also to view the Fréchet derivative as a linear differential operator acting on v. Then the relation

$$w\delta_v f - v\delta^*_w f = D \cdot \Psi(v, w; f) \tag{64}$$

as obtained using integration by parts defines the *Fréchet adjoint derivative*

$$\delta^*_w f = w\frac{\partial f}{\partial u} - D \cdot \left(w\frac{\partial f}{\partial(\partial u)}\right) + \cdots + (-D)^k \cdot \left(w\frac{\partial f}{\partial(\partial^k u)}\right) \tag{65}$$

which is a linear differential operator acting on w. The associated current $\Psi(v, w; f) = (\Psi^t, \Psi^x)$ is given by

$$\begin{aligned}\Psi(v, w; f) = {} & vw\frac{\partial f}{\partial(\partial u)} + (Dv) \cdot \left(w\frac{\partial f}{\partial(\partial^2 u)}\right) - vD \cdot \left(w\frac{\partial f}{\partial(\partial^2 u)}\right) + \cdots \\ & + \sum_{l=1}^{k}(D^{k-l}v) \cdot \left((-D)^{l-1} \cdot \left(w\frac{\partial f}{\partial(\partial^k u)}\right)\right).\end{aligned} \tag{66}$$

An alternative notation for the Fréchet derivative and its adjoint is $\delta_v f = f'(v)$ and $\delta^*_w f = f'^*(w)$, or sometimes $\delta_v f = D_v f$ and $\delta^*_w f = D^*_w f$.

The Fréchet derivative of a differential function f can be inverted to recover f by using a line integral along any curve C in J, where the endpoints ∂C are given by a general point $(t, x, u, \partial u, \ldots, \partial^k u) \in J$ and any chosen point $(t, x, u_0, \partial u_0, \ldots, \partial^k u_0) \in J$ at which f is non-singular. This yields

$$f\big|_{\partial C} = \int_C \frac{\partial f}{\partial u^{\mathrm{t}}} du^{\mathrm{t}} + \frac{\partial f}{\partial(\partial u^{\mathrm{t}})} \cdot d\partial u^{\mathrm{t}} + \cdots + \frac{\partial f}{\partial(\partial^k u^{\mathrm{t}})} \cdot d\partial^k u^{\mathrm{t}}. \tag{67}$$

If the curve C is chosen so that the contact relations hold, $d\partial u|_C = \partial du|_C$, $\ldots, d\partial^k u|_C = \partial^k du|_C$, then the line integral becomes a general homotopy integral

$$f = f\big|_{u=u_0} + \int_0^1 (\delta_v f)\Big|_{v=\partial_\lambda u_{(\lambda)}, u=u_{(\lambda)}} d\lambda, \quad u_{(1)} = u, \quad u_{(0)} = u_0 \tag{68}$$

where $u_{(\lambda)}(t, x)$ is a homotopy curve, given by a parametric family of functions. If f is non-singular when $u = 0$, then the homotopy curve can be chosen simply to be a homogeneous line, which yields a standard linear-homotopy integral [1]

$$f = f\big|_{u=0} + \int_0^1 (\delta_u f)\big|_{u=u_{(\lambda)}} \frac{d\lambda}{\lambda}, \quad u_{(\lambda)} = \lambda u. \tag{69}$$

The *Euler operator* E_u is defined in terms of the Fréchet derivative through the relation

$$\delta_v f = vE_u(f) + D\cdot \Upsilon_f(v) \tag{70}$$

obtained from integration by parts, which gives

$$E_u(f) = \frac{\partial f}{\partial u} - D\cdot\Big(\frac{\partial f}{\partial(\partial u)}\Big) + \cdots + (-D)^k\cdot\Big(\frac{\partial f}{\partial(\partial^k u)}\Big) \tag{71}$$

where

$$\begin{aligned}\Upsilon_f(v) = \Psi(v,1;f) = v\frac{\partial f}{\partial(\partial u)} + Dv\cdot\frac{\partial f}{\partial(\partial^2 u)} - vD\cdot\frac{\partial f}{\partial(\partial^2 u)} + \cdots \\ + \sum_{l=1}^{k}(D^{k-l}v)\cdot\Big((-D)^{l-1}\cdot\frac{\partial f}{\partial(\partial^k u)}\Big) = \sum_{l=0}^{k-1}(D^l v)\cdot E_{\partial_u^{l+1}}(f).\end{aligned} \tag{72}$$

The Euler-Lagrange relation (70) can be combined with the general homotopy integral (68) to obtain the following useful formula.

Lemma 4.1

$$f = \int_0^1 \partial_\lambda u_{(\lambda)} E_u(f)\big|_{u=u_{(\lambda)}}\, d\lambda + D\cdot F \tag{73}$$

is an identity, where

$$F = \int_0^1 \Upsilon_f(\partial_\lambda u_{(\lambda)})\big|_{u=u_{(\lambda)}}\, d\lambda + F_0 \tag{74}$$

with $F_0 = (F_0^t(t,x), F_0^x(t,x))$ *being any current such that* $D\cdot F_0 = f|_{u=u_0}$.

A useful relation is

$$\Upsilon_f(v) = vE_u^{(1)}(f) + D\cdot(vE_u^{(2)}(f)) + \cdots + D^{k-1}\cdot(vE_u^{(k)}(f)) \tag{75}$$

which arises through repeated integration by parts on the expression (72), where

$$\begin{aligned}E_u^{(l)}(f) = \frac{\partial f}{\partial(\partial^l u)} - \binom{l+1}{l} D\cdot\Big(\frac{\partial f}{\partial(\partial^{l+1}u)}\Big) + \cdots \\ + \binom{k}{l}(-D)^{k-l}\cdot\Big(\frac{\partial f}{\partial(\partial^k u)}\Big), \quad l = 1,\ldots,k\end{aligned} \tag{76}$$

define the *higher Euler operators*. Equations (70) and (75) then provide an alternative formula for the Fréchet derivative

$$\delta_v f = vE_u(f) + D \cdot (vE_u^{(1)}(f)) + \cdots + D^k \cdot (vE_u^{(k)}(f)) \tag{77}$$

which leads to a similar formula for the Fréchet adjoint derivative

$$\delta_w^* f = wE_u(f) - (Dw) \cdot E_u^{(1)}(f) + \cdots + (-D)^k w \cdot E_u^{(k)}(f) \tag{78}$$

after integration by parts. Explicit coordinate formulas for all of the Euler operators are stated in Ref.[1]; coordinate formulas for the Fréchet derivative and its adjoint, as well as the associated divergence, are shown in Ref.[27].

The Euler operators (71) and (76) have the following important properties.

Lemma 4.2 *(i) $E_u(fg) = \delta_g^* f + \delta_f^* g$ is a product rule. (ii) $E_u(f) = 0$ holds identically iff $f = D \cdot F$ for some differential current function $F = (F^t, F^x)$. (iii) $E_u^{(1)}(D \cdot F) = E_u(F^t, F^x) = (E_u(F^t), E_u(F^x))$ and $E_u^{(l+1)}(D \cdot F) = (E_u^{(l)}, E_u^{(l)}) \odot (F^t, F^x), l \geq 1$, are descent rules, where $\odot$ denotes the symmetric tensor product.*

The proof of (i) is an immediate consequence of the ordinary product rule applied to each partial derivative term in $E_u(fg)$. To prove the first part of (ii), if $f = D \cdot F$ then $\delta_v f = D \cdot \delta_v F$ combined with the Euler-Lagrange relation (70) yields $vE_u(f) = D \cdot (\delta_v F - \Upsilon_f(v))$. Since v is an arbitrary differential function, this implies $E_u(f) = 0$ (and $\Upsilon_f(v) = \delta_v F$ modulo a divergence-free term). Conversely, for the second part of (ii), if $E_u(f) = 0$ then the general homotopy integral (73) shows $f = D \cdot F$ holds, with F given by the formula (74). The proof of (iii) starts from the property $\delta_v(D \cdot F) = D \cdot \delta_v F$. Next, the Fréchet derivative relation (77) is applied separately to $f = D \cdot F$ and $f = F$. This yields $D \cdot (vE_u^{(1)}(D \cdot F)) = D \cdot (vE_u(F))$, $D^2 \cdot (vE_u^{(2)}(D \cdot F)) = D \cdot (vD \cdot E_u^{(1)}(F))$, and so on. The expressions for $E_u^{(1)}(D \cdot F)$, $E_u^{(2)}(D \cdot F)$, and so on are then obtained by recursively expanding out each Euler operator in components $E_u^{(1)} = E_u^{(t,x)} = (E_u^{(t)}, E_u^{(x)})$ and $E_u^{(l+1)} = E_u^{(l,t,x)} = (E^{(l,t)}, E_u^{(t,x)}), l \leq 1)$, followed by symmetrizing over these components together with the components of $F = (F^t, F^x)$. This completes the proof of Lemma 4.2. □

A *null-divergence* is a total divergence $D \cdot \Phi = 0$ vanishing identically in jet space, where $\Phi = (\Phi^t, \Phi^x)$ is a differential current function. Similarly to Poincaré's lemma, which shows that ordinary divergence-free vectors in $\mathbb{R}^n$ can be expressed as curls, null-divergences are total curls in jet space.

Lemma 4.3 *If a differential current function $\Phi = (\Phi^t(t, x, u, \partial u, \ldots, \partial^k u), \Phi^x(t, x, u, \partial u, \ldots, \partial^k u))$ has a null-divergence,*

$$D \cdot \Phi = D_t \Phi^t + D_x \cdot \Phi^x = 0 \text{ in } J, \tag{79}$$

then it is equal to a total curl

$$\Phi = D \cdot \Psi = (D_x \cdot \Theta, -D_t \Theta + D_x \cdot \Gamma) \text{ in } J \tag{80}$$

with

$$\Psi = \begin{pmatrix} 0 & \Theta \\ -\Theta & \Gamma \end{pmatrix} \tag{81}$$

holding for some differential vector function $\Theta(t,x,u,\partial u,\ldots,\partial^{k-1}u)$ and some differential antisymmetric tensor function $\Gamma(t,x,u,\partial u,\ldots,\partial^{k-1}u)$, both of which can be expressed in terms of Φ^t, Φ^x.

The proof begins by taking the Fréchet derivative of the null-divergence to get $D \cdot \delta_v \Phi = 0$. A descent argument will be used to solve this equation. Let the terms in $\delta_v \Phi = (\delta_v \Phi^t, \delta_v \Phi^x)$ containing highest derivatives $\partial^k v$ be denoted $(T^{(k)}\partial^k v, X^{(k)}\partial^k v)$, where the coefficients $T^{(k)}$ and $X^{(k)}$ of each term are given by a differential scalar function and a differential vector function in J. Then the highest derivative terms $\partial^{k+1}v$ in the equation $D\cdot\delta_v\Phi = 0$ consist of $T^{(k)}\partial_t\partial^k v + X^{(k)}\cdot\partial_x\partial^k v$. The coefficients of $\partial^{k+1}v$ in this expression must vanish, which can be shown to give $T^{(k)}\partial^k v = \theta^{(k-1)}\cdot\partial_x\partial^{k-1}v$ and $X^{(k)}\partial^k v = -\theta^{(k-1)}\partial_t\partial^{k-1}v + \gamma^{(k-1)}\cdot\partial_x\partial^{k-1}v$, where $\theta^{(k-1)}$ is some differential vector function, and $\gamma^{(k-1)}$ is some differential antisymmetric tensor function. Integration by parts on these expressions yields

$$T^{(k)}\partial^k v = D_x\cdot(\theta^{(k-1)}\partial^{k-1}v) + \text{ lower order terms},$$
$$X^{(k)}\partial^k v = -D_t(\theta^{(k-1)}\partial^{k-1}v) + D_x\cdot(\gamma^{(k-1)}\partial^{k-1}v) + \text{ lower order terms},$$

and hence

$$(T^{(k)}\partial^k v, X^{(k)}\partial^k v) = D\cdot\Psi^{(k-1)}(v) + \text{ lower order terms}$$

where

$$\Psi^{(k-1)}(v) = \begin{pmatrix} 0 & \Theta^{(k-1)}(v) \\ -\Theta^{(k-1)}(v) & \Gamma^{(k-1)}(v) \end{pmatrix}$$

with $\Theta^{(k-1)}(v) = \theta^{(k-1)}\partial^{k-1}v$ and $\Gamma^{(k-1)}(v) = \gamma^{(k-1)}\partial^{k-1}$. This shows that the highest derivative terms in $\delta_v\Phi$ have the form of a total curl, modulo lower order terms. Subtraction of this curl $D\cdot\Psi^{(k-1)}(v)$ from $\delta_v\Phi$ will now eliminate all terms containing $\partial^k v$, so that

$$\delta_v\Phi - D\cdot\Psi^{(k)}(v) = (T^{(k-1)}\partial^{k-1}v, X^{(k)}\partial^{k-1}v) + \text{ lower order terms}$$

where the coefficients $T^{(k-1)}$ and $X^{(k-1)}$ of the $\partial^{k-1}v$ terms are again a differential scalar function and a differential vector function in J. Since total curls have a vanishing total divergence, the highest derivative terms remaining in the null-divergence equation $0 = D\cdot\delta_v\Phi$ are given by $T^{(k-1)}\partial_t\partial^{k-1}v + X^{(k-1)}\cdot\partial_x\partial^{k-1}v$, which has the same form as the expression obtained at highest order. This completes the first step in the descent argument. Continuing to lower orders, the descent argument will terminate at the equation $T^{(0)}\partial_t v + X^{(0)} \cdot \partial_x v = 0$, which yields $T^{(0)} = 0$ and $X^{(0)} = 0$. As a result, the solution of the null-divergence equation $D \cdot \delta_v \Phi = 0$ is given by $\delta_v \Phi = \sum_{l=1}^{k} D \cdot \Psi^{(l-1)}(v)$.

The final step in the proof is simply to apply the general homotopy integral (68) to the Fréchet derivative $\delta_v \Phi = \sum_{l=1}^{k} D \cdot \Psi^{(l-1)}(v)$, which gives

$$\Phi - \Phi\big|_{u=u_0} = \int_0^1 \big(\sum_{l=1}^{k} D \cdot \Psi^{(l-1)}(\partial_\lambda u_{(\lambda)})\big)\big|_{u=u_{(\lambda)}}\, d\lambda.$$

This shows $\Phi = D \cdot \Psi$ is a total curl, where

$$\Psi = \Psi_0 + \int_0^1 \big(\sum_{l=1}^{k} \Psi^{(l-1)}(\partial_\lambda u_{(\lambda)})\big)\big|_{u=u_{(\lambda)}}\, d\lambda$$

has the form (81), with $D \cdot \Psi_0$ being an ordinary curl determined by Poincare's lemma applied to the vanishing divergence $D \cdot (\Phi|_{u=u_0}) = 0$. This completes the proof of Lemma 4.3. □

Scaling transformations are a one-parameter Lie group whose action is given by

$$t \to \lambda^a t, \quad x^i \to \lambda^{b_{(i)}} x^i, \quad u^\alpha \to \lambda^{c_{(\alpha)}} u^\alpha, \quad \lambda \neq 0 \tag{82}$$

prolonged to jet space, where the constants $a, b_{(i)}, c_{(\alpha)}$ are the scaling weights of t, x^i, u^α. Note the generator of these transformations is $\mathbf{X}_{\text{scal}} = \tau\partial_t + \xi\partial_x + \eta\partial_u$ where

$$\tau = at, \quad \xi = (b_{(1)}x^1, \ldots, b_{(n)}x^n), \quad \eta = (c_{(1)}u^1, \ldots, c_{(m)}u^m). \tag{83}$$

In characteristic form, the scaling generator is $\hat{\mathbf{X}}_{\text{scal}} = P_{\text{scal}}\partial_u$ with $P_{\text{scal}} = \eta - u_t\tau - u_x \cdot \xi$. Now consider a differential function f that is homogeneous under the action of the scaling transformation (82), such that $f \to \lambda^s f$. Then the infinitesimal action is given by $\hat{\mathbf{X}}_{\text{scal}}(f) = \delta_{P_{\text{scal}}} f = sf - \tau D_t f - \xi \cdot D_x f$. A useful identity comes from integrating this expression by parts and combining it with the Euler-Lagrange relation (70), yielding

$$\omega f = P_{\text{scal}} E_u(f) + D_t F^t + D_x \cdot F^x, \quad \omega = s + D_t\tau + D_x \cdot \xi = s + a + \sum_{i=1}^{n} b_{(i)} \tag{84}$$

where

$$F^t = f\tau + \Upsilon_f^t(P_{\text{scal}}), \quad F^x = f\xi + \Upsilon_f^x(P_{\text{scal}}) \tag{85}$$

with $\Upsilon_f = (\Upsilon_f^t, \Upsilon_f^x)$ given by expression (72). Note here ω is equal to the scaling weight of the integral quantity $\int_{t_0}^{t_1} \int_{\Omega} f\, dV\, dt$, as defined on any given spatial domain $\Omega \subseteq \mathbb{R}^n$ and any time interval $[t_0, t_1] \subseteq \mathbb{R}$.

Finally, for subsequent developments, the following technical result (which is a straightforward application of Hadamard's lemma [28] to the setting of jet space) will be useful.

Lemma 4.4 *If a differential function $f(t, x, u, \partial u, \ldots, \partial^k u)$ vanishes on the solution space $\mathscr{E}$ of a given regular PDE system* (39)*, then*

$$f = R_f(G) \tag{86}$$

holds identically, where

$$R_f = R_f^{(0)} + R_f^{(1)} \cdot D + \cdots + R_f^{(k-N)} \cdot D^{k-N} \tag{87}$$

is a linear differential operator, depending on f, with coefficients given by differential functions $R_f^{(0)}$, $R_f^{(1)}$, …, $R_f^{(k-N)}$ that are non-singular when evaluated on $\mathscr{E}$. The operator $R_f|_{\mathscr{E}}$ is uniquely determined by the function f if the PDE system has no differential identities. Otherwise, if the PDE system satisfies a differential identity

$$\mathscr{D}(G) = \mathscr{D}_1 G^1 + \cdots + \mathscr{D}_M G^M = 0 \tag{88}$$

with $\mathscr{D}_1, \ldots, \mathscr{D}_M$ being linear differential operators whose coefficients are non-singular differential functions when evaluated on $\mathscr{E}$, then the operator $R_f|_{\mathscr{E}}$ is determined by the function f only modulo $\chi\mathscr{D}$, where χ is an arbitrary differential function.

The proof relies heavily on the coordinatization property (54) that characterizes a PDE system being regular. For a regular PDE system $G = 0$ of order $N \geq 1$, consider its prolongation to order $k \geq 1$, $\mathrm{pr}G = (G, DG, \ldots, D^k G) = 0$, which has differential order $k+N$. Let $(\zeta^1 - g^1(Z), \zeta^2 - g^2(Z), \ldots)$ be the solved-form derivative expressions for the PDEs in $\mathrm{pr}G$, where $\zeta = (\zeta^1, \zeta^2, \ldots) \in J$ denotes the leading derivatives with respect to u^α chosen for the prolonged system, and $Z = (Z^1, Z^2, \ldots) \in J$ denotes the coordinates for the prolonged solution space $\mathscr{E} \subset J$ of the system. Note that $\mathrm{pr}G = 0$ represents $\mathscr{E}$ as a set of surfaces $\zeta^1 = g^1(Z)$, $\zeta^2 = g^2(Z)$,… in J. Then we have $f(\zeta, Z)|_{\mathscr{E}} = f(g(Z), Z) = 0$. We now use the standard line integral identity

$$f(\zeta, Z) = \int_{g(Z)}^{\zeta} \partial_y f(y, Z) \cdot dy = \int_0^1 (\zeta - g(Z)) \cdot \partial_\zeta f(s\zeta + (1-s)g(Z), Z)\, ds.$$

This shows that $f(\zeta, Z) = F(\zeta, Z) \cdot (\zeta - g(Z))$, with $F(\zeta, Z) = \int_0^1 \partial_\zeta f(s\zeta + (1 - s)g(Z), Z)\, ds$ being a vector function. Note $F(\zeta, Z)|_{\mathscr{E}} = F(g(Z), Z) = \partial_\zeta f(g(Z), Z)$ is non-singular since f is a differential function. Hence we obtain $F(\zeta, Z) \cdot (\zeta - g(Z)) = R_f(G)$ where R_f is a linear differential operator whose coefficients $F^1(\zeta, Z)$,

$F^2(\zeta, Z)$, ... are non-singular when evaluated on $\mathscr{E}$. Furthermore, the expression for $F(\zeta, Z)$ shows that it is canonically determined by f, unless the PDE system satisfies a differential identity, whereby $0 = \mathscr{D}(G) = h(Z) \cdot (\zeta - g(Z))$ holds identically for some vector function $h(Z)$. In this case, $R_f(G)$ is well-defined only modulo $\chi\mathscr{D}(G) = 0$, where χ is any differential function. This completes the proof of Lemma 4.4. □

5 Characteristic Forms and Determining Equations for Conservation Laws and Symmetries

Consider an infinitesimal symmetry (47) of a regular PDE system (39). When acting on the solution space $\mathscr{E}$ of the PDE system in jet space J, the symmetry generator is equivalent to a generator given by

$$\hat{\mathbf{X}} = \mathbf{X} - \mathbf{X}_{\text{triv}} = P\partial/\partial u, \quad P = \eta - u_t\tau - u_x \cdot \xi \tag{89}$$

under which u is infinitesimally transformed while t, x are invariant. This generator (89) defines the *characteristic form* (or canonical representation) for the infinitesimal symmetry. The symmetry invariance (48) of the PDE system can then be expressed by

$$\text{pr}\hat{\mathbf{X}}(G)|_{\mathscr{E}} = 0 \tag{90}$$

holding on the whole solution space $\mathscr{E}$ of the given system. Note that the action of $\text{pr}\hat{\mathbf{X}}$ is the same as a Fréchet derivative (63), and hence an equivalent, modern formulation [1, 3] of this invariance (90) is given by the *symmetry determining equation*

$$(\delta_P G)|_{\mathscr{E}} = 0. \tag{91}$$

This formulation of infinitesimal symmetries has several advantages compared to the standard formulation shown in Sect. 3. Firstly, a symmetry is trivial iff its characteristic function P vanishes on $\mathscr{E}$. Also, the differential order of a symmetry is simply given by the differential order of $P|_{\mathscr{E}}$. Secondly, the symmetry determining equation (91) can be set up without doing any prolongations of the generator (89), as only total differentiation is needed. Thirdly, when contact symmetries or higher-order symmetries are sought, the generator can be formulated simply as

$$\hat{\mathbf{X}} = P(t, x, u, \partial u, \ldots, \partial^r u)\partial_u \tag{92}$$

with the symmetry determining equation then being a linear PDE for the characteristic function P. This formulation (92) eliminates arbitrary functions depending on all of the variables $t, x, u, \partial u, \ldots, \partial^r u$ in the solution for P.

Now consider a conservation law (40) of a regular PDE system (39). The starting point to obtain an equivalent characteristic form of the conservation law is provided by equations (86) and (87) in Lemma 4.4. These equations show that the conservation law can be expressed as a divergence identity

$$D_t T + D_x \cdot X = R_\Phi(G) = R_\Phi^{(0)} G^{\mathrm{t}} + R_\Phi^{(1)} \cdot DG^{\mathrm{t}} + \cdots + R_\Phi^{(r+1-N)} \cdot D^{r+1-N} G^{\mathrm{t}} \quad (93)$$

which is obtained by moving off solutions of the PDE system, where $u(t,x)$ is an arbitrary (sufficiently smooth) function. Here r is the differential order of the conserved current $\Phi = (T, X)$, and N is the differential order of the PDE system. The next step is to integrate by parts on the righthand side in the divergence identity (93), yielding

$$D_t \tilde{T} + D_x \cdot \tilde{X} = GQ \quad (94)$$

with

$$\begin{aligned}(\tilde{T}, \tilde{X}) = (T, X) &+ R_\Phi^{(1)} G^{\mathrm{t}} + R_\Phi^{(2)} \cdot DG^{\mathrm{t}} - (D \cdot R_\Phi^{(2)}) G^{\mathrm{t}} \\ &+ \cdots + \sum_{l=0}^{r-N} \left((-D)^l \cdot R_\Phi^{(r+1-N)}\right) \cdot D^{r-N-l} G^{\mathrm{t}}\end{aligned} \quad (95)$$

and

$$Q^{\mathrm{t}} = (Q_1, \ldots, Q_M) = R_\Phi^{(0)} - D \cdot R_\Phi^{(1)} + \cdots + (-D)^{r+1-N} \cdot R_\Phi^{(r+1-N)}. \quad (96)$$

On the solution space $\mathscr{E}$, note that $(\tilde{T}, \tilde{X})|_{\mathscr{E}} = (T, X)|_{\mathscr{E}}$ reduces to the conserved density and the flux in the given conservation law $(D_t T + D_x \cdot X)|_{\mathscr{E}} = 0$, and hence

$$(D_t \tilde{T} + D_x \cdot \tilde{X})|_{\mathscr{E}} = 0 \quad (97)$$

is a locally equivalent conservation law. The identity (94) is called the *characteristic equation* for the conservation law, and the set of functions (96) is called the *multiplier*. Explicit coordinate formulas for the density $\tilde{T}$ and the flux $\tilde{X}$ in terms of T and X are shown in Ref.[27].

When a regular PDE system is expressed in a solved form (54)–(55) for a set of leading derivatives, note that these leading derivatives (and their differential consequences) can be eliminated from the expression for a conserved current $\Phi = (T, X)$ without loss of generality, since this only changes the conserved current by the addition of a locally trivial current. Then it is straightforward to derive explicit expressions for the coefficient functions in the operator R_Φ by applying the chain rule to $D_t T$ and $D_{x^i} X^i$ with the use of subleading derivatives defined by the relations $\partial_t \partial_{(\ell_a/t)} u^{\alpha_a} = \partial_{x^i} \partial_{(\ell_a/x^i)} u^{\alpha_a} = \partial_{(\ell_a)} u^{\alpha_a}$. This leads to an explicit Euler-Lagrange expression

$$Q^{\mathrm{t}} = \big(E_{\partial_{(\ell_1/t)} u^{\alpha_1}}(T) + \sum_{i=1}^{n} E_{\partial_{(\ell_1/x^i)} u^{\alpha_1}}(X^i), \ldots, E_{\partial_{(\ell_M/t)} u^{\alpha_M}}(T) + \sum_{i=1}^{n} E_{\partial_{(\ell_M/x^i)} u^{\alpha_M}}(X^i)\big) \tag{98}$$

for the components of the multiplier (96), where $\partial_{(\ell_a/t)} u^{\alpha_a}$ and $\partial_{(\ell_a/x^i)} u^{\alpha_a}$ denote the subleading derivatives. As a result, the multiplier components (98) can contain leading derivatives $\partial_{(\ell_a)} u^{\alpha_a}$ (and their differential consequences) at most polynomially.

Also note that, as asserted by Lemma 4.4, if a regular PDE system has no differential identities (88), then the operator $R_\Phi|_{\mathscr{E}}$ will be canonically determined by the expression for $\Phi = (T, X)$. This implies the relation

$$\begin{aligned} Q^{\mathrm{t}}|_{\mathscr{E}} = (Q_1, \ldots, Q_M)|_{\mathscr{E}} &= E_G(D_t \tilde{T} + D_x \cdot \tilde{X})|_{\mathscr{E}} \\ &= (E_{G^1}(D_t T + D_x \cdot X), \ldots, E_{G^M}(D_t T + D_x \cdot X))|_{\mathscr{E}} \end{aligned} \tag{99}$$

for the multiplier (96).

In general, for a given regular PDE system (39), a set of functions

$$Q = (Q_1(t, x, u, \partial u, \partial^2 u, \ldots \partial^r u), \ldots, Q_M(t, x, u, \partial u, \partial^2 u, \ldots \partial^r u))^{\mathrm{t}} \tag{100}$$

will be a multiplier iff each function is non-singular on the PDE solution space $\mathscr{E}$ and their summed product with the expressions $G = (G^1, \ldots, G^M)$ for the PDEs has the form of a total space-time divergence.

The characteristic equation (94) establishes that, up to local equivalence, all non-trivial conservation laws for any regular PDE system arise from multipliers. A determining condition to find all multipliers comes from Lemma 4.2 applied to the characteristic equation (94), yielding

$$0 = E_u(GQ) = \delta_Q^* G + \delta_G^* Q. \tag{101}$$

This condition, which is required to hold identically in jet space, is necessary and sufficient for Q to be a multiplier. For each solution Q, a corresponding conserved current that satisfies the characteristic equation (94) can be obtained from the expression $f = GQ$ by using Lemma 4.1. This yields

$$\tilde{\Phi} = \int_0^1 \Upsilon_{GQ}(\partial_\lambda u_{(\lambda)})\big|_{u=u_{(\lambda)}} \, d\lambda \tag{102}$$

whose multiplier (96) is Q. An explicit formula for this conserved current is stated next.

Lemma 5.5 *For a regular PDE system* (39)*, each multiplier* (100) *yields a conserved current* (94) *which is explicitly given by a homotopy integral*

$$\tilde{T} = \int_0^1 \Big(\sum_{l=0}^{k-1} \partial_\lambda \partial^l u_{(\lambda)} \cdot \Big(E_{\partial^l \partial_t u}(GQ)\Big)\Big|_{u=u_{(\lambda)}} \Big) d\lambda + D_x \cdot \Theta, \tag{103}$$

$$\tilde{X} = \int_0^1 \Big(\sum_{l=0}^{k-1} \partial_\lambda \partial^l u_{(\lambda)} \cdot (E_{\partial^l \partial_x u}(GQ))\Big|_{u=u_{(\lambda)}} \Big) d\lambda - D_t \cdot \Theta + D_x \cdot \Gamma \tag{104}$$

along a homoptopy curve $u_{(\lambda)}(t,x)$, with $u_{(1)} = u$ and $u_{(0)} = u_0$ such that $(GQ)|_{u=u_0}$ is non-singular. Here $k = \max(r, N)$.

Note the conserved current formula (103)–(104) can be simplified by evaluating it on the solution space $\mathscr{E}$ of the given regular PDE system. Modulo a locally trivial current, this yields

$$\tilde{\Phi}|_{\mathscr{E}} = \int_0^1 \sum_{j=1}^{k} \Big(\partial_\lambda \partial^{j-1} u_{(\lambda)} \sum_{l=j}^{k} (-D)^{l-j} \cdot \Big(\frac{\partial G}{\partial(\partial^l u)} Q \Big)\Big|_{u=u_{(\lambda)}} \Big) d\lambda \tag{105}$$

where the curve $u_{(\lambda)}(t,x)$ is now in the solution space $\mathscr{E}$.

5.1 Correspondence Between Conservation Laws and Multipliers

As shown by the following key result, multipliers provide a unique characteristic form (or canonical representation) for locally equivalent conservation laws, in analogy to the characteristic form (89) for symmetries, if a regular PDE system has no differential identities. A generalization holding for regular PDE systems with differential identities will be stated later.

Proposition 1 *For any regular PDE system* (39) *that has no differential identities, a conserved current is locally trivial* (44) *iff its corresponding multiplier* (96) *vanishes when evaluated on the solution space of the system.*

The proof has two parts. For the "only if part", suppose a conserved current is locally trivial (44). By Lemma 4.4, the conserved density and the flux will have the respective forms $T = D_x \cdot \Theta + \hat{T}(G)$ and $X = -D_t\Theta + D_x \cdot \Gamma + \hat{X}(G)$ for some linear differential operators $\hat{T}$ and $\hat{X}$ whose coefficients are differential functions that are non-singular when evaluated on $\mathscr{E}$. For this conserved current $\Phi = (T, X)$, consider the divergence identity (93), where $R_\Phi(G) = D_t\hat{T}(G) + D_x \cdot \hat{X}(G)$. As the PDE system is assumed to have no differential identities, then the homotopy integral formula for the operator R_Φ from the proof of Lemma 4.4 shows that integration by parts applied to $R_{\Phi}(G)$ yields $\tilde{T} = T - \hat{T}(G)$, $\tilde{X} = X - \hat{X}(G)$, in the characteristic equation (94)–(95), and hence $GQ = D_t\tilde{T} + D_x \cdot \tilde{X} = 0$.

It is now straightforward to determine Q from the equation $GQ = 0$. In the case when G comprises a single PDE (i.e., $M = 1$), then $Q = 0$ is immediate. In the case when G contains more than one PDE (i.e., $M > 1$), the equation $GQ = 0$ can be solved by linear algebra as follows.

First express each PDE $G^a = 0$, $a = 1, \ldots, M$, in the solved form (54)–(55) in terms of a leading derivative $\partial_{(\ell_a)}u^{\alpha_a}$. Then take the Fréchet derivative of $GQ = 0$, which yields

$$(\delta_v G)Q + G(\delta_v Q) = 0.$$

To solve this Fréchet derivative equation, consider the terms involving $\partial^k\partial_{(\ell_a)}v^{\alpha_a}$ and let $w = (\partial_{(\ell_1)}v^{\alpha_1}, \ldots, \partial_{(\ell_M)}v^{\alpha_M})$ for ease of notation. It is easy to see the expression $\delta_v G$ contains only one term of this form, which is simply given by w itself, as a consequence of the solved form of the PDEs $G = (G^1, \ldots, G^M)$. The expression $\delta_v Q$ contains a sum of terms involving derivatives of w, which will have the form $\sum_{k=0}^{r} Q^{(k)}\partial^k w^{\mathrm{t}}$, where r is the differential order of the highest derivatives of the variables $\partial_{(\ell_a)}u^{\alpha_a}$ in Q, and where the coefficients $Q^{(k)}$ are differential $M \times M$ matrix functions in J. Hence, all of the terms involving $\partial^k\partial_{(\ell_a)}v^{\alpha_a}$ in the Fréchet derivative equation consist of $wQ + \sum_{k=0}^{r} GQ^{(k)}\partial^k w^{\mathrm{t}} = 0$. Then the coefficients of each jet variable $\partial^k w$, $k = 0, 1, \ldots, r$, must vanish separately. This immediately yields $Q^{(k)} = 0$ for $k = 1, \ldots, r$. The remaining terms are given by $wQ + GQ^{(0)}w^{\mathrm{t}} = 0$. This is a linear homogeneous equation in w^{t}, after the transpose relation $wQ = (wQ)^t = Q^t w^t$ is used, which gives $(Q^{\mathrm{t}} + GQ^{(0)})w^t = 0$. The vanishing of the coefficient of w^t yields $Q^{\mathrm{t}} = -GQ^{(0)}$, and hence $Q|_{\mathscr{E}} = 0$.

For the "if part", suppose a multiplier satisfies $Q|_{\mathscr{E}} = 0$. Then, Lemma 4.4 can be applied to get $Q = \hat{Q}(G)$, where $\hat{Q}$ is some linear differential operator whose coefficients are differential functions that are non-singular when evaluated on $\mathscr{E}$. The characteristic equation (94) must now be solved to determine the corresponding conserved density $\tilde{T}$ and flux $\tilde{X}$. This will be done in two main steps.

For the first step, a descent argument will be given to solve the Fréchet derivative equation

$$D \cdot (\delta_v\tilde{\Phi}) = (\delta_v G)Q + G(\delta_v Q)$$

for $\delta_v\tilde{\Phi} = (\delta_v\tilde{T}, \delta_v\tilde{X})$, similarly to the proof of Lemma 4.3. Let $F(v) = (\delta_v G)Q + G(\delta_v Q)$, with $Q = \hat{Q}(G)$. The terms in $F(v)$ containing highest derivatives of v will be denoted $F^{(k)}\partial^k v$, where k is the larger of the differential orders of Q and G, and where the coefficients $F^{(k)}$ are differential functions in J such that $F^{(k)}|_{\mathscr{E}} = 0$ since $F(v)|_{\mathscr{E}} = ((\partial_v G)\hat{Q}(G))|_{\mathscr{E}} = (\partial_v G)|_{\mathscr{E}}\hat{Q}(0) = 0$. Note the differential order of $\delta_v\tilde{\Phi}$ then can be assumed to be $k-1$. Next, let $\Upsilon(v) = (\delta_v\tilde{T}, \delta_v\tilde{X})$, and denote the terms containing highest derivatives of v in $\Upsilon(v)$ as $\tilde{T}^{(k-1)}\partial^{k-1}v$ and $\tilde{X}^{(k-1)}\partial^{k-1}v$, respectively, where the coefficients $\tilde{T}^{(k-1)}$ and $\tilde{X}^{(k-1)}$ are given by a set of differential scalar functions and a set of differential vector functions in J. In this notation, the Fréchet derivative equation becomes

$$D \cdot \varUpsilon(v) = F(v).$$

Now the highest derivative terms $\partial^k v$ in this equation are given by

$$\tilde{T}^{(k-1)} \partial_t \partial^{k-1} v + \tilde{X}^{(k-1)} \cdot \partial_x \partial^{k-1} v = F^{(k)} \partial^k v.$$

Expand out $F^{(k)}\partial^k v = F^{(k-1,t)}\partial_t\partial^{k-1}v + F^{(k-1,x)} \cdot \partial_x \partial^{k-1} v$, and collect the terms $\partial_t \partial^{k-1} v$ and $\partial_x \partial^{k-1} v$ in the equation, giving

$$(\tilde{T}^{(k-1)} - F^{(k-1,t)})\partial_t\partial^{k-1}v + (\tilde{X}^{(k-1)} - F^{(k-1,x)}) \cdot \partial_x\partial^{k-1}v = 0.$$

The same analysis used in the proof of Lemma 4.3 then yields

$$\begin{aligned}(\tilde{T}^{(k-1)}\partial^{k-1}v, \tilde{X}^{(k-1)}\partial^{k-1}v) = {} & (F^{(k-1,t)}\partial^{k-1}v, F^{(k-1,x)}\partial^{k-1}v) \\ & + D \cdot \varPsi^{(k-2)}(v) + \text{ lower order terms}\end{aligned}$$

where

$$\varPsi^{(k-2)}(v) = \begin{pmatrix} 0 & \varTheta^{(k-2)}(v) \\ -\varTheta^{(k-2)}(v) & \varGamma^{(k-2)}(v) \end{pmatrix}$$

with $\varTheta^{(k-2)}(v) = \theta^{(k-2)}\partial^{k-2}v$ and $\varGamma^{(k-2)}(v) = \gamma^{(k-2)}\partial^{k-2}$ being given by some differential vector function $\theta^{(k-2)}$ and some differential antisymmetric tensor function $\gamma^{(k-2)}$. Hence the highest derivative terms in $\varUpsilon(v)$ involving v have the form

$$\varUpsilon(v) = (F^{(k-1,t)}\partial^{k-1}v, F^{(k-1,x)}\partial^{k-1}v) + D \cdot \varPsi^{(k-2)}(v) + \tilde{\varUpsilon}(v)$$

where $\tilde{\varUpsilon}(v)$ comprises all remaining terms, which contain derivatives of v up to order $\partial^{k-2}v$, and where $D \cdot \varPsi^{(k-2)}(v)$ is a total curl, which has a vanishing total divergence. Substitution of this expression $\varUpsilon(v)$ into the Fréchet derivative equation gives

$$\begin{aligned}(\tilde{T}^{(k-2)} + D \cdot F^{(k-1,t)})\partial_t\partial^{k-2}v + (\tilde{X}^{(k-2)} + D \cdot F^{(k-1,x)}) \cdot \partial_x\partial^{k-2}v \\ = F^{(k-1)}\partial^{k-1}v + \text{ lower order terms}\end{aligned}$$

where $\tilde{T}^{(k-2)}$ and $\tilde{X}^{(k-2)}$ are a set of differential scalar functions and a set of differential vector functions given by the coefficients of the terms $\partial^{k-2}v$ in $\tilde{\varUpsilon}(v)$. After $F^{(k-1)}\partial^{k-1}v = F^{(k-2,t)}\partial_t\partial^{k-2}v + F^{(k-2,x)} \cdot \partial_x\partial^{k-2}v$ is expanded out, the terms containing highest derivatives of v in this equation are given by

$$(\tilde{T}^{(k-2)} - F^{(k-2,t)} + D \cdot F^{(k-1,t)})\partial_t \partial^{k-2} v \\ + (\tilde{X}^{(k-2)} - F^{(k-2,x)} + D \cdot F^{(k-1,x)}) \cdot \partial_x \partial^{k-2} v = 0$$

which has the same form as the equation solved previously. This completes the first step in the descent argument.

Next, continuing to all lower orders, the descent argument yields

$$\Upsilon(v) = \sum_{l=1}^{k-1} D \cdot \Psi^{(l-1)}(v) + \sum_{l=0}^{k-1} \sum_{j=l}^{k-1} ((-D)^{j-l} \cdot F^{(j,t)}, (-D)^{j-l} \cdot F^{(j,x)})\partial^l v.$$

Note the terms in the first sum are a total curl, and the terms in the second sum vanish on $\mathscr{E}$ since $F(l)|_{\mathscr{E}} = 0$.

The final step is to apply the general line integral (67) to the Fréchet derivative $\delta_v \tilde{\Phi} = \Upsilon(v)$ evaluated on $\mathscr{E}$. Since $\Upsilon(v)|_{\mathscr{E}} = \sum_{l=1}^{k-1} D \cdot \Psi^{(l-1)}(v)|_{\mathscr{E}}$, this gives

$$\tilde{\Phi}|_{\mathscr{E}} - \tilde{\Phi}|_{u=0} = \int_0^1 \sum_{l=1}^{k-1} D \cdot \Psi^{(l-1)}(v)\Big|_{u=u_{(\lambda)}, v=\partial_\lambda u_{(\lambda)}} d\lambda$$

where $u_{(\lambda)}(t,x)$ is a homotopy curve in the solution space $\mathscr{E}$ of the regular PDE system, with $u_{(1)} = u(t,x)$ being an arbitrary solution and $u_{(0)} = u_0(t,x)$ being any particular solution. Thus $\tilde{\Phi}|_{\mathscr{E}} - \tilde{\Phi}_0 = D \cdot \Psi$ is a total curl, where

$$\Psi = \int_0^1 \sum_{l=1}^{k-1} \Psi^{(l-1)}(v)\Big|_{u=u_{(\lambda)}, v=\partial_\lambda u_{(\lambda)}} d\lambda$$

has the form (81). Now, substitution of $\tilde{\Phi}|_{\mathscr{E}} = \tilde{\Phi}|_{u=u_0} + D \cdot \Psi$ into $D \cdot \tilde{\Phi} = GQ$ yields $0 = (D \cdot \tilde{\Phi} - GQ)|_{\mathscr{E}} = D \cdot (\tilde{\Phi}|_{u=u_0})$. This immediately establishes that $\tilde{\Phi}|_{u=u_0} = D \cdot \Psi_0$ is an ordinary curl, by Poincare's lemma. Thus,

$$\tilde{\Phi}|_{\mathscr{E}} = D \cdot (\Psi + \Psi_0)$$

is a locally trivial conserved current, which completes the proof of Proposition 1.

□

The correspondence stated in Proposition 1 no longer holds when a PDE system possesses a differential identity (88). In particular, for a given differential identity, multiplication by an arbitrary differential function χ, followed by integration by parts, yields

$$0 = \chi\mathscr{D}(G) = G\mathscr{D}^*(\chi) + D \cdot \Phi(\chi, G) \tag{106}$$

where $\Phi(\chi, G)$ is a conserved current that vanishes on the solution space of the PDE system,

$$\Phi(\chi, G)|_{\mathscr{E}} = \Phi(\chi, 0) = 0. \tag{107}$$

Hence

$$Q = \mathscr{D}^*(\chi) \tag{108}$$

is a multiplier which determines a locally trivial conserved current. This derivation can be reversed, showing that the existence of a multiplier (108) is necessary and sufficient for a PDE system to possess a differential identity (88).

Multipliers of the form (108), given by a linear differential operator acting on an arbitrary differential function χ, will be called *gauge multipliers* [16], in analogy with gauge symmetries. Note that a gauge multiplier is non-vanishing on the solution space $\mathscr{E}$ of the PDE system whenever the differential identity is non-trivial, since $\mathscr{D}|_{\mathscr{E}} \neq 0$ implies $Q|_{\mathscr{E}} \neq 0$ for $\chi \neq 0$. Two multipliers that differ by a gauge multiplier will be called *gauge equivalent*.

Running Ex. (3) The Euler equations for constant density, inviscid fluids in two dimensions comprise an evolution equation for $\mathbf{u} = (u^1, u^2)$,

$$\mathbf{G} = \mathbf{u}_t + \mathbf{u} \cdot \nabla \mathbf{u} + (1/\rho)\nabla p = 0,$$

a spatial equation relating $\mathbf{u}$ to p,

$$G^p = (1/\rho)\Delta p + (\nabla \mathbf{u}) \cdot (\nabla \mathbf{u})^{\mathrm{t}} = 0,$$

and a spatial constraint equation on $\mathbf{u}$,

$$G^{\mathrm{div}} = \nabla \cdot \mathbf{u} = 0.$$

This PDE system obeys a differential identity

$$\mathrm{Div}\,\mathbf{G} - D_t G^{\mathrm{div}} - G^p = 0$$

which has the form (88) where $\mathscr{D} = \mathrm{diag}(\mathrm{Div}\,, -1, -D_t)$ and $G = (\mathbf{G}, G^p, G^{\mathrm{div}})$. The corresponding gauge multiplier is given by

$$Q = (\mathbf{Q}, Q^p, Q^{\mathrm{div}})^{\mathrm{t}}, \quad \mathbf{Q} = -\mathrm{Grad}\,\chi, \quad Q^p = -\chi, \quad Q^{\mathrm{div}} = D_t\chi$$

where χ is an arbitrary differential scalar function. The characteristic equation yields

$$GQ = -(\mathrm{Grad}\,\chi) \cdot \mathbf{G} - \chi G^p + (D_t\chi)G^{\mathrm{div}} = D_t(\chi G^{\mathrm{div}}) + D_x \cdot (-\chi \mathbf{G})$$

which is a locally trivial conservation law, where $T = \chi G^{\mathrm{div}}$ is the conserved density and $\mathbf{X} = -\chi \mathbf{G}$ is the spatial flux. If χ is chosen to be a constant, $\chi = 1$, then the conserved density becomes a total spatial divergence $T = D_x \cdot \mathbf{u}$ which produces a boundary conservation law

$$\frac{d}{dt}\int_{\Omega} TdV|_{\mathscr{E}} = \frac{d}{dt}\oint_{\partial\Omega} \mathbf{u}\cdot\boldsymbol{\nu}dA|_{\mathscr{E}} = 0$$

on any closed spatial domain $\Omega \in \mathbb{R}^2$, since the flux vanishes on the solution space of the system, $\mathbf{X}|_{\mathscr{E}} = 0$. This boundary conservation law represents conservation of streamlines in the fluid.

Running Ex. (4) The magnetohydrodynamics equations for a compressible, infinite conductivity plasma in three dimensions comprise evolution equations for $\rho, \mathbf{u} = (u^1, u^2, u^3)$ and $\mathbf{B} = (B^1, B^2, B^3)$,

$$G^{\rho} = \rho_t + \nabla\cdot(\rho\mathbf{u}) = 0,$$
$$\mathbf{G}^u = \mathbf{u}_t + \mathbf{u}\cdot\nabla\mathbf{u} + (1/\rho)(P'(\rho)\nabla\rho - \mathbf{J}\times\mathbf{B}) = 0, \quad 4\pi\mathbf{J} = \nabla\times\mathbf{B},$$
$$\mathbf{G}^B = \mathbf{B}_t - \nabla\times(\mathbf{u}\times\mathbf{B}) = 0,$$

and a spatial constraint equation on $\mathbf{B}$,

$$G^{\text{div}} = \nabla\cdot\mathbf{B} = 0.$$

This PDE system obeys a differential identity

$$\text{Div}\,(\mathbf{G}^B) - D_t G^{\text{div}} = 0$$

which has the form (88) where $\mathscr{D} = \text{diag}(0, 0, \text{Div}\,, -D_t)$ and $G = (G^{\rho}, \mathbf{G}^u, \mathbf{G}^B, G^{\text{div}})$. The corresponding gauge multiplier is given by

$$Q = (Q^{\rho}, \mathbf{Q}^u, \mathbf{Q}^B, Q^{\text{div}})^{\text{t}}, \quad Q^{\rho} = 0, \quad \mathbf{Q}^u = 0, \quad \mathbf{Q}^B = -\text{Grad}\,\chi, \quad Q^{\text{div}} = D_t\chi$$

where χ is an arbitrary differential scalar function. The characteristic equation yields

$$GQ = -(\text{Grad}\,\chi)\cdot\mathbf{G}^B + (D_t\chi)G^{\text{div}} = D_t(\chi G^{\text{div}}) + D_x\cdot(-\chi\mathbf{G}^B)$$

which is a locally trivial conservation law, where $T = \chi G^{\text{div}}$ is the conserved density and $\mathbf{X} = -\chi\mathbf{G}^B$ is the spatial flux. If χ is chosen to be a constant, $\chi = 1$, then the conserved density becomes a total spatial divergence $T = D_x \cdot \mathbf{B}$ which produces a boundary conservation law

$$\frac{d}{dt}\int_{\Omega} TdV|_{\mathscr{E}} = \frac{d}{dt}\oint_{\partial\Omega} \mathbf{B}\cdot\boldsymbol{\nu}dA|_{\mathscr{E}} = 0$$

on any closed spatial domain $\Omega \in \mathbb{R}^3$, since the flux vanishes on the solution space of the system, $\mathbf{X}|_{\mathscr{E}} = 0$. This boundary conservation law represents conservation of magnetic flux in the plasma.

The following natural generalization of Proposition 1 will now be established.

Proposition 2 *For any regular PDE system* (39) *that possesses a differential identity* (88)*, a conserved current is locally trivial* (44) *iff its corresponding multiplier* (96) *evaluated on the solution space of the system is equal to a gauge multiplier* (108) *for some differential function* χ.

The same steps used in the proof for Proposition 1 go through with only two changes. For the "if part", suppose a multiplier satisfies $Q|_{\mathscr{E}} = \mathscr{D}^*(\chi)$, which implies $Q = \hat{Q}(G) + \mathscr{D}^*(\chi)$ by Lemma 4.4, where $\hat{Q}$ is some linear differential operator whose coefficients are differential functions that are non-singular when evaluated on $\mathscr{E}$. Then the conservation law identity (106) combined with the characteristic equation (94) yields

$$G\hat{Q}(G) = G(Q - \mathscr{D}^*(\chi)) = D_t(\tilde{T} + \Phi^t(\chi, G)) + D_x \cdot (\tilde{X} + \Phi^x(\chi, G)).$$

This equation can be solved by the same steps used in proving the "if part" of Proposition 1, thus showing that $\tilde{\Phi} + \Phi(\chi, G)$ is a locally trivial current. Since $\Phi(\chi, G)$ itself is a locally trivial current, the conservation law given by $\tilde{\Phi}$ is therefore locally trivial (44). For the "only if" part, suppose a conserved current is locally trivial (44), so then, by Lemma 4.4, the conserved density and the spatial flux will have the respective forms $T = D_x \cdot \Theta + \hat{T}(G)$ and $X = -D_t\Theta + D_x \cdot \Gamma + \hat{X}(G)$ for some linear differential operators $\hat{T}$ and $\hat{X}$ whose coefficients are differential functions that are non-singular when evaluated on $\mathscr{E}$. As the PDE system is assumed to satisfy a differential identity (88), the divergence identity (93) will be unique only up to the addition of a multiple of this differential identity, $\chi\mathscr{D}(G) = 0$. This implies from the homotopy integral formula for the operator R_Φ that the characteristic equation (94)–(95) holds with $\tilde{T} = T - \hat{T}(G) - \Phi^t(\chi, G)$, $\tilde{X} = X - \hat{X}(G) - \Phi^x(\chi, G)$, and $GQ = G\mathscr{D}^*(\chi)$. The equation $G(Q - \mathscr{D}^*(\chi)) = 0$ can be solved by the same steps used in proving the "only if part" of Proposition 1, thereby showing $(Q - \mathscr{D}^*(\chi))|_{\mathscr{E}} = 0$, so $Q|_{\mathscr{E}}$ is equal to $\mathscr{D}^*(\chi)|_{\mathscr{E}}$. This completes the proof of Proposition 2. □

The characterization of locally trivial conservation laws in Proposition 1 and Proposition 2 establishes an important general correspondence result which underlies the usefulness of multipliers.

For a given regular PDE system, the set of multipliers forms a vector space on which the symmetries of the system have a natural action [27, 28]. A multiplier is called *trivial* if yields a locally trivial conservation law, and two multipliers are said to be *equivalent* if they differ by a trivial multiplier. When the PDE system has no differential identities, then a multiplier Q is trivial iff it vanishes on the solution space, $Q|_{\mathscr{E}} = 0$, whereas when the PDE system possesses a differential identity (88), a multiplier Q is trivial iff it equals a gauge multiplier (108) on the solution space, $Q|_{\mathscr{E}} = \mathscr{D}^*(\chi)$. A set of multipliers is linearly independent if no linear combination of the multipliers is trivial. Likewise, a set of conservation laws is linearly independent if no linear combination of the conserved currents is locally trivial.

Theorem 5.1 *(i) For any regular PDE system* (39)*, whether or not it possesses a differential identity, there is a one-to-one correspondence between its admitted equivalence classes of linearly-independent local conservation laws and its admitted equivalence classes of linearly-independent multipliers. (ii) An explicit formulation of this correspondence is given by the homotopy integral formula* (103)—(104) *for conserved currents in terms of multipliers.*

Infinitesimal symmetries have a well-known action on conserved currents [1, 3]. This action induces a corresponding action of infinitesimal symmetries on multipliers [27, 28], and there are several equivalent formulas [6, 13, 14, 28–32] for the conserved current obtained from the action of a given infinitesimal symmetry applied to a given multiplier. It is worth noting that this action does not preserve linear independence of equivalence classes. For example [28, 29], any non-trivial conserved current that does not explicitly contain at least one of the independent variables in a PDE system is mapped into a locally trivial current under any translation symmetry.

5.2 Low-Order Conservation Laws

For any given regular PDE system, the correspondence between local conservation laws and multipliers stated in Theorem 5.1 gives a straightforward way using the following three steps to find all of the non-trivial local conservation laws (up to equivalence) admitted by the PDE system. Step 1: solve the determining condition (101) to obtain all multipliers. Step 2: find all linearly independent equivalence classes of non-trivial multipliers. Step 3: apply the homotopy integral formula (105) to a representative multiplier in each equivalence class to obtain a corresponding conserved current.

In practice, for solving the determining condition (101), it is very useful to know at which differential orders the non-trivial multipliers will be found. As seen in the examples in Sect. 2, physically important conservation laws, such as energy and momentum, always have a low differential order for the conserved density T and the spatial flux X, whereas conservation laws having a high differential order are typically connected with integrability. A general pattern emerges from these conservation law examples when their multipliers are examined.

In Ex 1 and Ex 2, mass conservation for the transport equation (11) and net heat conservation for the diffusion/heat conduction equation (13) both have $Q = 1$ which does not involve u or its derivatives.

In Ex 3, energy conservation for the telegraph equation (15) has $Q = \exp(2\int a(t)dt)u_t$, while the leading derivative in this equation is u_{tt} or u_{xx}.

In Ex 4, for the nonlinear dispersive wave equation (17), mass conservation, L^2-norm conservation, and energy conservation respectively have $Q = 1$, $Q = 2u$, and $Q = g(u) + u_{xx}$. The leading derivative in this equation is u_t or u_{xxx}.

In Ex 5, for the viscous fluid equations (19), mass conservation has $Q^{\mathrm{t}} = (1, 0)$, momentum conservation has $Q^{\mathrm{t}} = (u, 1)$, and Galilean momentum conservation has $Q^{\mathrm{t}} = (tu, t)$, while $\{\rho_t, u_t\}$ is a set of leading derivatives in this system.

In Ex 6, energy conservation for the barotropic gas flow/compressible inviscid fluid equations (21), has $Q^{\mathrm{t}} = (\frac{1}{2}u^2, \rho u)$, and again $\{\rho_t, u_t\}$ is a set of leading derivatives in this system.

In Ex 7, momentum conservation, energy conservation, and energy-momentum conservation for the breaking wave equation (23) respectively have $Q = 1$, $Q = u$, $Q = -(u_{tx} - u(u_m + \frac{1}{2}u) + \frac{1}{2}u_x^2)$, while the Hamiltonian Casimir has $Q = \frac{1}{2}(u - u_{xx})^{1/2}$. The leading derivative in this equation is u_{txx} or u_{xxx}.

In Ex 8, mass conservation for the porous media equation (25) has $Q = \alpha(x)$.

In Ex 9, angular momentum conservation and boost momentum conservation for the non-dispersive wave equation (27) respectively have $Q = (\mathbf{a} \cdot \mathbf{x}) \cdot \nabla u$, $Q = \mathbf{b} \cdot x u_t + c^2 t \mathbf{b} \cdot \nabla u$, while the leading derivative in this equation is u_{tt} or Δu.

In all of these examples, each variable $\partial^k u$ that appears in the conservation law multiplier is related to some leading derivative of u in the PDE system by differentiation of this variable $\partial^k u$ with respect to t, x.

In contrast, the conservation laws for the higher-derivative quantities (22b) in Ex 6 and (24e) in Ex 7 have, respectively, $Q^{\mathrm{t}} = ((2u_x\rho_x/u_x^2 - p'\rho_x^2/(\rho^2)^2)\ (u_x^2/(u_x^2 - p'\rho x^2/\rho^2)^2) + p'/\rho^2(\rho_x^2/(u_x^2 - p'\rho x_2/\rho^2)^2)_x)$ and $Q = \frac{5}{2}m^{-7/2}m_x^2 - 2m^{-5/2}m_{xx} - 2m^{-3/2}$, which involve variables of higher differential order than the leading derivatives.

An exceptional case is the conservation laws for local helicity and local enstrophy in Ex 10. These conservation laws for the inviscid (compressible/incompressible) fluid equation (29) have, respectively, $\mathbf{Q} = 2\nabla \times \mathbf{u}$ which involves a variable with the same differential order as the leading derivative $\mathbf{u}_t$, and $\mathbf{Q} = f''((\operatorname{curl} \mathbf{u})/\rho)\nabla \cdot ((\nabla \wedge \mathbf{u})/\rho)$ which involves a higher-derivative variable. Note, however, if the fluid equation is expressed as a system for the velocity $\mathbf{u}$ and the vorticity vector $\boldsymbol{\omega} = \nabla \times \mathbf{u}$ in three dimensions or the vorticity scalar $\omega = \operatorname{curl} \mathbf{u}$ in two dimensions, then the multipliers for helicity and enstrophy conservation are given by, respectively, $Q^{\mathrm{t}} = (\boldsymbol{\omega}, \mathbf{u})$ and $Q^{\mathrm{t}} = (\mathbf{0}, f'(\omega/\rho))$ in which the variables are related to the leading derivatives $\mathbf{u}_t$ and $\boldsymbol{\omega}_t$ by differentiation with respect to t.

This pattern motivates introducing the following general class of multipliers. A multiplier Q for a regular PDE system (39) will be called *low-order* if each jet variable $\partial^k u^\alpha$ that appears in $Q|_{\mathscr{E}}$ is related to some leading derivative of u^α by differentiations with respect to t, x^i. (Note that, therefore, the differential order r of $Q|_{\mathscr{E}}$ must be strictly less than the differential order N of the PDE system.) Correspondingly, a conservation law is said to be of *low-order* if its multiplier is low-order when evaluated on the solution space of the PDE system.

For a given regular PDE system, the explicit form for low-order conservation laws can be determined from the form for low-order multipliers by inverting the relation (96) which defines a multiplier in terms of a conserved current.

Running Ex. (1) The gKdV equation (56) is a time evolution PDE whose leading derivative is u_t or u_{xxx}. Its low-order conservation laws $(D_tT + D_xX)|_{\mathscr{E}} = 0$ are given by multipliers that have the form

$$Q(t, x, u, u_x, u_{xx})$$

since, in the jet space $J = (t, x, u, u_t, u_x, u_{tt}, u_{tx}, u_{xx}, \ldots)$, the only variables that can be differentiated with respect to t or x to obtain a leading derivative are u, u_x, u_{xx}. To derive the corresponding form for low-order conserved currents $\varPhi = (T, X)$, the first step is to expand out D_tT and D_xX starting from general expressions for T and X in which a leading derivative u_t or u_{xxx} has been eliminated along with all of its differential consequences. If u_t is chosen, then the starting expressions will be $T(t, x, u, u_x, u_{xx}, \ldots)$ and $X(t, x, u, u_x, u_{xx}, \ldots)$, which gives

$$D_tT = T_t + u_tT_u + u_{tx}T_{u_x} + u_{txx}T_{u_{xx}} + \cdots ,$$
$$D_xX = X_x + u_xX_u + u_{xx}X_{u_x} + u_{xxx}X_{u_{xx}} + \cdots .$$

The second step is to obtain the operator $R_\varPhi$ from the terms in the divergence expression $D_tT + D_xX$ containing u_t (and its differential consequences). This yields

$$D_tT+D_xX = (T_u+T_{u_x}D_x+T_{u_{xx}}D_x^2+\cdots)u_t+T_t+X_x+u_xX_u+u_{xx}X_{u_x}+u_{xxx}X_{u_{xx}}+\cdots$$

and hence

$$R_\varPhi = T_u + T_{u_x}D_x + T_{u_{xx}}D_x^2 + \cdots$$

since $u_t = G - u^pu_x - u_{xxx}$ is the solved form for the PDE expression. Then the main steps are, first, to equate Q with the expression $E_G(R_\varPhi(G))$ and, next, to use the resulting equation together with the characteristic equation $D_t\tilde{T} + D_x\tilde{X} = GQ$ to determine the dependence of T and X on all jet variables that do not appear in Q. This gives, first,

$$R_\varPhi(G) = T_uG + T_{u_x}D_xG + T_{u_{xx}}D_x^2G + \cdots = \delta_GT = GE_u(T) + D_x\varUpsilon^x(G)$$

by using the Euler-Lagrange relation (70), which yields the equation

$$Q(t, x, u, u_x, u_{xx}) = E_G(R_\varPhi(G)) = E_G(\delta_GT) = E_u(T).$$

Comparison of the differential order of both sides of this equation directly determines

$$T = \tilde{T}(t,x,u,u_x) + D_x\Theta(t,x,u,u_x,\ldots).$$

This implies $\Upsilon^x(G) = \tilde{T}_{u_x}G$. Next, the characteristic equation then yields

$$GQ = D_tT + D_x(X - G\tilde{T}_{u_x}) = D_t\tilde{T} + D_x\tilde{X}$$

which gives

$$D_x(X + D_t\Theta) = D_x(\tilde{X} + G\tilde{T}_{u_x}) = -\tilde{T}_t + (u^p u_x + u_{xxx})\tilde{T}_u + (u^p u_x + u_{xxx})_x\tilde{T}_{u_x}.$$

Comparison of both sides of this equation now determines

$$X = \tilde{X}(t,x,u,u_t,u_x,u_{xx}) - D_t\Theta(t,x,u,u_x,\ldots) - G\tilde{T}_{u_x}(t,x,u,u_x).$$

The same result can be shown to hold if u_{xxx} is chosen as the leading derivative instead of u_t. Hence, all low-order conserved currents have the general form

$$\Phi|_{\mathscr{E}} = (\tilde{T}(t,x,u,u_x), \tilde{X}(t,x,u,u_t,u_x,u_{xx}))$$

modulo locally trivial conserved currents.

Running Ex. (2) The breaking wave equation (58) is a regular PDE whose leading derivative is u_{txx} or u_{xxx}. All low-order conservation laws of this PDE are given by multipliers that have the second-order form

$$Q(t,x,u,u_t,u_x,u_{tx},u_{xx}) \tag{109}$$

where u_{tt} is excluded because it cannot be differentiated to obtain a leading derivative u_{txx} or u_{xxx}. The corresponding form for low-order conserved currents $\Phi = (T,X)$ is derived by starting from general expressions for T and X in which a leading derivative u_{txx} or u_{xxx} has been eliminated along with all of its differential consequences. It is simplest to use the pure derivative u_{xxx}, which implies T and X are functions only of t, x, u, u_x, u_{xx}, and their t-derivatives. Then the terms in the divergence expression $D_tT + D_xX$ containing the leading derivative u_{xxx} (and its differential consequences) are given by

$$\begin{aligned}D_tT + D_xX = {} & (X_{u_{xx}} + X_{u_{txx}}D_t + X_{u_{ttxx}}D_t^2 + \cdots)u_{xxx}\\ & + T_t + X_x + u_tT_u + u_xX_u + u_{tt}T_{u_t} + u_{tx}(T_{u_x} + X_{u_t}) + u_{xx}X_{u_x}\\ & + u_{ttt}T_{u_{tt}} + u_{ttx}(T_{u_{tx}} + X_{u_{tt}}) + u_{txx}(T_{u_{xx}} + X_{u_{tx}}) + \cdots .\end{aligned}$$

This expression yields the operator

$$R_\Phi = (X_{u_{xx}} + X_{u_{txx}}D_t + X_{u_{ttxx}}D_t^2 + \cdots)u^{-1}$$

since $u_{xxx} = -u^{-1}(G + bu_x(u_{xx} - u) - u_{txx} + u_t) + u_x$ is the solved form for the PDE expression. Now, the main steps consist of first, equating Q with the expression $E_G(R_\Phi(G))$ and, next, using the characteristic equation $D_t\tilde{T} + D_x\tilde{X} = GQ$ to determine the dependence of T and X on all jet variables that do not appear in Q. The first step gives

$$R_\Phi(G) = (X_{u_{xx}} + X_{u_{txx}} D_t + X_{u_{ttxx}} {D_t}^2 + \cdots)(-u^{-1}G) = -u^{-1}GE_{u_{xx}}(X) - D_t\Upsilon^t(u^{-1}G)$$

after using the relation (70), which yields the equation

$$Q(t, x, u, u_t, u_x, u_{tx}, u_{xx}) = E_G(R_\Phi(G)) = E_G(-u^{-1}GE_{u_{xx}}(X)) = -u^{-1}E_{u_{xx}}(X).$$

Comparison of the differential order of both sides of this equation directly determines

$$X = \tilde{X}(t, x, u, u_t, u_x, u_{tx}, u_{xx}) - D_t\Theta(t, x, u, u_t, u_x, u_{tt}, u_{tx}, u_{xx}, \ldots)$$

which implies $\Upsilon^t(u^{-1}G) = 0$. Then, for the next step, the characteristic equation yields

$$GQ = D_tT + D_xX = D_t\tilde{T} + D_x\tilde{X}$$

giving

$$D_t(T - D_x\Theta) = D_t\tilde{T} = -\tilde{X}_x - u_x\tilde{X}_u - u_{tx}\tilde{X}_{u_t} - u_{xx}\tilde{X}_{u_x} - u_{txx}\tilde{X}_{u_{tx}}.$$

Comparison of both sides of this equation now determines

$$T = \tilde{T}(t, x, u, u_x, u_{xx}) + D_x\Theta(t, x, u, u_t, u_x, u_{tt}, u_{tx}, u_{xx}, \ldots).$$

Hence, all low-order conserved currents have the general form

$$\Phi|_{\mathscr{E}} = (\tilde{T}(t, x, u, u_x, u_{xx}), \tilde{X}(t, x, u, u_t, u_x, u_{tx}, u_{xx}))$$

modulo locally trivial conserved currents.

6 Variational Symmetries and Noether's Theorem in Modern Form

A PDE system (39) is *globally variational* if it is given by the critical points of a variational principle defined on some spatial domain $\Omega \subseteq \mathbb{R}^n$ and some time interval $[t_0, t_1] \subseteq \mathbb{R}$. In typical applications, this will involve specifying a function space for $u(t, x)$ with $x \in \Omega$ and also posing boundary conditions on $u(t, x)$ for

$x \in \partial\Omega$. Noether's theorem is usually formulated in this context, where it shows that every transformation group leaving invariant the variational principle yields a corresponding conserved integral (42) for solutions of the PDE system with $u(t,x)$ belonging to the specified function space.

However, for the purpose of obtaining local conservation laws (40), a global variational principle is not necessary, and a PDE system instead needs to have just a local variational principle.

A PDE system (39) is *locally variational* if it is given by the Euler-Lagrange equations

$$0 = G = E_u(L)^{\mathrm{t}} \tag{110}$$

for some differential function $L(t, x, u, \partial u, \ldots, \partial^k u)$, called a *Lagrangian*. Note that, as shown by Lemma 4.2, a Lagrangian is unique up to addition of an arbitrary total divergence. In particular, L and $\tilde{L} = L + D_t\Psi^t + D_x \cdot \Psi^x$ have the same Euler-Lagrange equations, for any differential scalar function Ψ^t and differential vector function Ψ^x.

There is a well-known condition for a given PDE system to be locally variational [1, 3].

Lemma 6.6 *$G = E_u(L)^{\mathrm{t}}$ holds for some Lagrangian $L(t, x, u, \partial u, \ldots, \partial^k u)$ iff*

$$\delta_v G^{\mathrm{t}} = \delta_v^* G^{\mathrm{t}} \tag{111}$$

holds for all differential functions $v(t, x)$.

The "only if" part of the proof has two steps. First, $\delta_v E_u(L) = E_u(\delta_v L)$ can be directly verified to hold, due to v having no dependence on u and derivatives of u. Next, the Euler-Lagrange relation (70) combined with Lemma 4.2 yields $E_u(\delta_v L) = E_u(vE_u(L)) = \delta_v^* E_u(L)$, after again using the fact that v has no dependence on u and derivatives of u. Hence, $\delta_v E_u(L) = \delta_v^* E_u(L)$, which completes this part of the proof.

The "if" part of the proof proceeds by first inverting the relation $G^{\mathrm{t}} = E_u(L)$ through applying Lemma 4.1 to $f = L$. This yields $L = \tilde{L} + D \cdot F$, with $\tilde{L} = \int_0^1 \partial_\lambda u_{(\lambda)} G^{\mathrm{t}}\big|_{u=u_{(\lambda)}} d\lambda$. Then the remaining steps consist of showing that $E_u(L) = E_u(\tilde{L}) = G^{\mathrm{t}}$ holds for this Lagrangian when $\delta_v G^{\mathrm{t}} = \delta_v^* G^{\mathrm{t}}$. First, the Fréchet derivative of $\tilde{L}$ gives $\delta_v \tilde{L} = \int_0^1 \big(\partial_\lambda v_{(\lambda)} G^{\mathrm{t}}\big|_{u=u_{(\lambda)}} + \partial_\lambda u_{(\lambda)} \delta_{v_{(\lambda)}} G^{\mathrm{t}}\big|_{u=u_{(\lambda)}}\big) d\lambda$ where $v_{(\lambda)} = \delta_v u_{(\lambda)}$. Next, substitute $\delta_{v_{(\lambda)}} G^{\mathrm{t}} = \delta^*_{v_{(\lambda)}} G^{\mathrm{t}}$ and use the Fréchet derivative relation (64), which yields

$$\begin{aligned}
\delta_v \tilde{L} &= \int_0^1 \big(\partial_\lambda v_{(\lambda)} G^{\mathrm{t}}\big|_{u=u_{(\lambda)}} + v_{(\lambda)} \partial_\lambda G^{\mathrm{t}}\big|_{u=u_{(\lambda)}} - D \cdot \Psi(\partial_\lambda u_{(\lambda)}, v_{(\lambda)}; G^{\mathrm{t}})\big|_{u=u_{(\lambda)}}\big) d\lambda \\
&= vG^{\mathrm{t}} - v_0 G^{\mathrm{t}}\big|_{u=u_0} - D \cdot \int_0^1 \Psi(\partial_\lambda u_{(\lambda)}, v_{(\lambda)}; G^{\mathrm{t}})\big|_{u=u_{(\lambda)}} \, d\lambda
\end{aligned}$$

where Ψ is given by expression (66). Finally, apply E_v to $\delta_v\tilde{L}$ to get $E_v(\delta_v\tilde{L}) = G^{\mathrm{t}}$, and use the identity $E_v(\delta_v\tilde{L}) = E_v(vE_u(\tilde{L})) = E_u(\tilde{L})$ which follows from the Euler-Lagrange relation (70). This yields $E_u(\tilde{L}) = G^{\mathrm{t}}$, which completes the proof. □

The condition (111) for a PDE system $G = 0$ to be locally variational states that the linearization of G^{t} must be self-adjoint. From the relations (63) and (78), or equivalently (65) and (77), this condition splits with respect to $v, \partial v, \ldots, \partial^k v$ into a linear overdetermined system of equations on G:

$$\frac{\partial G}{\partial(\partial^k u)} = (-1)^k \big(E_u^{(k)}(G)\big)^{\mathrm{t}}, \quad k = 0, 1, \ldots, N \tag{112}$$

where N is the differential order of the PDE system $G = 0$. These equations are called the *Helmholtz conditions*. Note the appearance of the transpose implies that the Helmholtz conditions cannot hold if u and G have a different number of components. Also, the expression (76) for the higher Euler operators $E_u^{(k)}$ shows that the Helmholtz condition for $k = N$ reduces to the equation

$$(1 - (-1)^N)\Big(\frac{\partial G}{\partial(\partial^N u)} + \Big(\frac{\partial G}{\partial(\partial^N u)}\Big)^{\mathrm{t}}\Big) = 0 \tag{113}$$

which cannot hold if N is odd. Consequently, a necessary condition for a PDE system to be locally variational is that its differential order N must be even and the number M of PDEs must be the same as the number m of dependent variables.

When a PDE system satisfies the Helmholtz conditions (112), a Lagrangian L for the system can be recovered from the expressions $G = (G^1, \ldots, G^M)$ by the general homotopy integral formula

$$L = \int_0^1 \partial_\lambda u_{(\lambda)} G^{\mathrm{t}}\big|_{u=u_{(\lambda)}} d\lambda \tag{114}$$

(as shown in the proof of Lemma 6.6). A total divergence can be added to this Lagrangian to obtain an equivalent Lagrangian that has the lowest possible differential order, which is $N/2$.

Running Ex. (1) The gKdV equation (56) is an odd-order PDE. Hence, it cannot be locally variational as it stands. To verify there is no local variational principle, note $G = G^{\mathrm{t}} = u_t + u^p u_x + u_{xxx}$ gives

$$\delta_v G^{\mathrm{t}} = v_t + u^p v_x + pu^{p-1}u_x v + v_{xxx}, \quad \delta_v^* G^{\mathrm{t}} = -v_t - u^p v_x - v_{xxx}$$

and hence $\delta_v G^{\mathrm{t}} - \delta_v^* G^{\mathrm{t}} = 2v_t + 2u^p v_x + pu^{p-1}u_x v + 2v_{xxx} \neq 0$ whereby G^{t} fails to have a self-adjoint linearization. Equivalently, the Helmholtz conditions are not satisfied:

$$
\begin{aligned}
&(k=0) \quad \frac{\partial G}{\partial u} = \rho u^{p-1} u_x \neq E_u(G) = 0,\\
&(k=1) \quad \frac{\partial G}{\partial u_t} = 1 \neq -E_u^{(t)}(G) = -1, \quad \frac{\partial G}{\partial u_x} = u^p \neq -E_u^{(x)}(G) = -u^p,\\
&(k=2) \quad \frac{\partial G}{\partial u_{xx}} = 0 = E_u^{(x,x)}(G) = -D_x(1),\\
&(k=3) \quad \frac{\partial G}{\partial u_{xxx}} = 1 \neq -E_u^{(x,x,x)}(G) = -1.
\end{aligned}
$$

However, if a potential variable w is introduced by putting $u = w_x$, then the PDE becomes $w_{tx} + w_x^p w_{xx} + w_{xxxx} = 0$ which has even order. Repetition of the previous steps, with $G = G^{\mathrm{t}} = w_{tx} + w_x^p w_{xx} + w_{xxxx}$, now gives

$$\delta_v G^{\mathrm{t}} = v_{tx} + w_x^p v_{xx} + p w_x^{p-1} w_{xx} v_x + v_{xxxx} = \delta_v^* G^{\mathrm{t}}$$

and

$$
\begin{aligned}
&(k=0) \quad \frac{\partial G}{\partial w} = 0 = E_w(G),\\
&(k=1) \quad \frac{\partial G}{\partial w_t} = 0 = -E_w^{(t)}(G) = -D_x(1),\\
&\qquad\qquad \frac{\partial G}{\partial w_x} = p w_x^{p-1} w_{xx} = -E_w^{(x)}(G)\\
&\qquad\qquad\qquad = -p w_x^{p-1} w_{xx} + 2D_x(w_x^p) + D_t(1) + D_x^3(1),\\
&(k=2) \quad \frac{\partial G}{\partial w_{tx}} = 1 = E_w^{(t,x)}(G),\\
&(k=3) \quad \frac{\partial G}{\partial w_{xxx}} = 0 = -E_w^{(x,x,,x)}(G) = D_x(1),
\end{aligned}
$$

$$(k=4) \quad \frac{\partial G}{\partial w_{xxxx}} = 1 = E_w^{(x,x,x,x)}(G).$$

Hence, the potential gKdV equation is locally variational. A Lagrangian is given by the homotopy integral

$$L = \int_0^1 w(\lambda w_{tx} + \lambda^{p+1} w_x^p w_{xx} + \lambda w_{xxxx})\, d\lambda = \tfrac{1}{2} w w_{tx} + \tfrac{1}{p+2} w w_x^p w_{xx} + \tfrac{1}{2} w w_{xxxx}$$

using $w_{(\lambda)} = \lambda w$. The addition of a total divergence $D_t \Psi^t + D_x \Psi^x$ given by

$$\Psi^t = -\tfrac{1}{2} w w_x, \quad \Psi^x = -\tfrac{1}{2}(w w_{xxx} - w_x w_{xx}) - \tfrac{1}{(p+1)(p+2)} w w_x^{p+1}$$

yields an equivalent Lagrangian that has minimal differential order,

$$\tilde{L} = -\tfrac{1}{2}w_x w_t - \tfrac{1}{(p+1)(p+2)}w_x^{p+2} + \tfrac{1}{2}w_{xx}^2.$$

For a locally variational PDE system, a global variational principle on a spatial domain Ω and a time interval $[t_0, t_1]$ can be defined in terms of a Lagrangian by

$$S[u] = \int_{t_0}^{t_1} \int_{\Omega} (L(t, x, u, \partial u, \ldots, \partial^k u) + D_x \cdot \Theta(t, x, u, \partial u, \ldots))\, dV\, dt \qquad (115)$$

where the spatial divergence term is chosen to let spatial boundary conditions be posed on $u(t, x)$ for $x \in \partial\Omega$. The critical points of the variational principle (115) are given by the vanishing of the variational derivative of $S[u]$,

$$\begin{aligned} 0 = S'[u] &= \frac{\partial}{\partial\epsilon} S[u + \epsilon v]\big|_{\epsilon=0} \\ &= \int_{t_0}^{t_1} \int_{\Omega} v E_u(L)\, dV\, dt + \int_{t_0}^{t_1} \oint_{\partial\Omega} (\delta_v \Theta + \Upsilon_L(v)) \cdot \nu\, dA \end{aligned} \qquad (116)$$

where $v(t, x)$ is an arbitrary differential function that satisfies the same spatial boundary conditions as $u(t, x)$. Here ν denotes the outward unit normal vector on $\partial\Omega$, and Υ_L is given by the Euler-Lagrange relation (70). Provided Θ is chosen so that the boundary integral vanishes, then $S'[u] = 0$ yields the PDE system $G = E_u(L)^{\mathrm{t}} = 0$ on the spatial domain Ω.

6.1 Variational Symmetries

A *variational symmetry* [1, 2] of a given variational principle (115) is a generator (47) whose prolongation leaves invariant the variational principle. This invariance condition has both a global aspect, which involves the spatial domain and the spatial boundary conditions, and a local aspect, which involves only the Lagrangian.

For a local variational principle (110), a *variational (divergence) symmetry* [1, 2] is a generator (47) whose prolongation satisfies the invariance condition

$$\mathrm{pr}\mathbf{X}(L) = \tau D_t L + \xi \cdot D_x L + D_t \Psi^t + D_x \cdot \Psi^x \qquad (117)$$

for some for differential scalar function Ψ^t and differential vector function Ψ^x. This condition can be expressed alternatively as

$$\mathrm{pr}\mathbf{X}(L) = D_t \tilde{\Psi}^t + D_x \cdot \tilde{\Psi}^x - (D_t \tau + D_x \cdot \xi)L \qquad (118)$$

with $\tilde{\Psi}^t = \Psi^t + L\tau$ and $\tilde{\Psi}^x = \Psi^x + L\xi$, where $D_t\tau + D_x \cdot \xi$ represents the infinitesimal conformal change in the space-time volume element $dVdt$ under the symmetry generator $\mathbf{X}$.

A simpler formulation of a variational symmetry is given by using the characteristic form (89) for the symmetry generator. Then an infinitesimal symmetry (89) is a variational symmetry iff its prolongation leaves invariant the Lagrangian modulo total a divergence,

$$\mathrm{pr}\hat{\mathbf{X}}(L) = D_t\Psi^t_P + D_x \cdot \Psi^x_P \tag{119}$$

for some for differential scalar function Ψ^t_P and differential vector function Ψ^x_P depending on the characteristic function P of the symmetry. Note that, since any total divergence is annihilated by the Euler operator E_u, a variational symmetry preserves the critical points of the Lagrangian L. As a consequence, every variational symmetry is an infinitesimal symmetry of the PDE system $G = E_u(L) = 0$. The converse is not true in general, since (for example) scaling symmetries of Euler-Lagrange equations need not always preserve the Lagrangian.

There is an equivalent, modern formulation of the variational symmetry condition (119) which uses only the Euler-Lagrange equations and not the Lagrangian itself.

Proposition 3 *For any locally variational PDE system* (110)*, an infinitesimal symmetry in characteristic form* $\hat{\mathbf{X}} = P(t, x, u, \partial u, \ldots, \partial^r u)\partial_u$ *is a variational symmetry iff*

$$\delta_P G^{\mathrm{t}} = -\delta^*_G P^{\mathrm{t}} \tag{120}$$

holds identically.

To prove this result, first note that $E_u(\mathrm{pr}\hat{\mathbf{X}}(L))$ vanishes identically iff $\mathrm{pr}\hat{\mathbf{X}}(L)$ is a total divergence, by Lemma 4.2. Next, $E_u(\mathrm{pr}\hat{\mathbf{X}}(L)) = E_u(\delta_P L) = E_u(PE_u(L)) = \delta^*_P G^{\mathrm{t}} + \delta^*_{G^{\mathrm{t}}} P$ directly follows from the Euler-Lagrange relation (70) combined with the product rule shown in Lemma 4.2 for the Euler operator. Finally, $\delta^*_P G^{\mathrm{t}} = \delta_P G^{\mathrm{t}}$ holds by Lemma 6.6, and $\delta^*_{G^{\mathrm{t}}} P = \delta^*_G P^{\mathrm{t}}$ holds as an identity. Hence $E_u(\mathrm{pr}\hat{\mathbf{X}}(L)) = \delta_P G^{\mathrm{t}} + \delta^*_G P^{\mathrm{t}}$ is an identity. This completes the proof. □

An importance consequence of equation (120) is that it provides a determining condition to find *all* variational symmetries for a given locally variational PDE system, without the explicit use of a Lagrangian. In particular, this formulation avoids the need to consider the "gauge terms" $D_t\Psi^t + D_x \cdot \Psi^x$ which arise in the Lagrangian formulation (119).

Running Ex. (1) The Lie symmetries of the gKdV equation (56) consist of a time translation $\hat{\mathbf{X}} = -u_t\partial_u$, a space translation $\hat{\mathbf{X}} = -u_x\partial_u$, a scaling $\hat{\mathbf{X}} = -(\frac{2}{p}u + 3tu_t + xu_x)\partial_u$, and a Galilean boost $\hat{\mathbf{X}} = (1 - tu_x)\partial_u$ if $p \neq 1$. These symmetries project to corresponding Lie symmetries of the potential gKdV equation $w_{tx} + w^p_x w_{xx} + w_{xxxx} = 0$ through the relation $u = w_x$. This yields the generator $\hat{\mathbf{X}} = P\partial_w$ with

$$P = P^{\mathrm{t}} = (1 - \tfrac{2}{p})c_3 w + c_4 - (c_1 + 3c_3 t)w_t - (c_2 + c_3 x + c_4 t)w_x. \tag{121}$$

The variational Lie symmetries can be easily found by checking the condition (120). Using $G = G^{\mathrm{t}} = w_{tx} + w_x^p w_{xx} + w_{xxxx}$, a simple computation yields

$$\begin{aligned}\delta_P G^{\mathrm{t}} &= D_t D_x P + p w_x^{p-1} w_{xx} D_x P + w_x^p D_x^2 P + D_x^4 P\\ &= -(c_1 + 3c_3 t)D_t G - (c_2 + c_3 x + c_4 t)D_x G - (3 + \tfrac{2}{p})c_3 G\end{aligned} \tag{122}$$

and also

$$\begin{aligned}\delta_G^* P^{\mathrm{t}} &= G\frac{\partial P}{\partial w} - D_t\Big(G\frac{\partial P}{\partial w_t}\Big) - D_x\Big(G\frac{\partial P}{\partial w_x}\Big)\\ &= (5 - \tfrac{2}{p})c_3 G + (c_1 + 3c_3 t)D_t G + (c_2 + c_3 x + c_4 t)D_x G.\end{aligned} \tag{123}$$

Hence, $0 = \delta_P G^{\mathrm{t}} + \delta_G^* P^{\mathrm{t}} = (2 - \frac{4}{p})c_3 G$ determines $(p-2)c_3 = 0$. This shows that all of the Lie symmetries except the scaling symmetry are variational symmetries for an arbitrary nonlinearity power $p \neq 0$, and that the scaling symmetry is a variational symmetry only for the special power $p = 2$.

6.2 Noether's Theorem in Modern Form

Variational symmetries have a direct relationship to local conservation laws through the variational identity

$$\begin{aligned}\mathrm{pr}\hat{\mathbf{X}}(L) &= D_t \Psi_P^t + D_x \cdot \Psi_P^x\\ &= \delta_P L = P E_u(L) + D \cdot \Upsilon_L(P)\end{aligned} \tag{124}$$

holding due to the Euler-Lagrange relation (70). The identity (124) yields

$$P E_u(L) = D \cdot \Phi, \quad \Phi = (\Psi_P^t - \Upsilon_L^t(P), \Psi_P^x - \Upsilon_L^x(P)) \tag{125}$$

which is a conservation law in characteristic form for the PDE system given by $E_u(L) = 0$. When combined with the formula (105) for conserved currents, this provides a modern, local form of Noether's theorem, which does not explicitly use the Lagrangian.

Theorem 6.2 *For any locally variational PDE system* $G = E_u(L)^{\mathrm{t}} = 0$, *variational symmetries* $\hat{\mathbf{X}} = P\partial_u$ *and local conservation laws in characteristic form* $D_t\tilde{T} + D_x \cdot \tilde{X} = GQ$ *have a one-to-one correspondence given by the relation*

$$P = Q^{\mathrm{t}}. \tag{126}$$

Equivalently, this correspondence is given by the homotopy integral

$$\tilde{\Phi} = (\tilde{T}, \tilde{X}) = \int_0^1 \sum_{j=1}^{k} \left(\partial_\lambda \partial^{j-1} u_{(\lambda)} \left(\sum_{l=j}^{k} (-D)^{l-j} \cdot \left(\frac{\partial (PG^{\mathrm{t}})}{\partial(\partial^l u)} \right) \Big|_{u=u_{(\lambda)}} \right) \right) d\lambda \tag{127}$$

modulo a total curl, along a homotopy curve $u_{(\lambda)}(t,x)$, with $u_{(1)} = u$ and $u_{(0)} = u_0$ such that $(GQ)|_{u=u_0}$ is non-singular. Here $k = \max(r,N)$.

The Noether correspondence stated in Theorem 6.2 has a sharper formulation using the additional correspondence between multipliers and local conservation laws provided by Theorem 5.1. This formulation depends on whether a given variational PDE system possesses differential identities or not.

In particular, when a PDE system satisfies a differential identity (88), there will exist *gauge symmetries*

$$\hat{\mathbf{X}} = (\mathscr{D}^*(\chi))^{\mathrm{t}} \partial_u \tag{128}$$

corresponding to gauge multipliers (108), where $\mathscr{D}$ is the linear differential operator defining the given differential identity (88), and χ is an arbitrary differential function. Two symmetries that differ by a gauge symmetry will be called *gauge equivalent*.

Recall, for any regular PDE system, a symmetry is trivial iff its characteristic function vanishes on the solution space of the PDE system, and two symmetries are equivalent iff they differ by a trivial symmetry.

Corollary 1 *(i) If a locally variational, regular PDE system* (110) *has no differential identities, then there is a one-to-one correspondence between its admitted equivalence classes of linearly-independent local conservation laws and its admitted equivalence classes of linearly-independent variational symmetries. (ii) If a locally variational, regular PDE system* (110) *satisfies a differential identity, then its admitted equivalence classes of linearly-independent local conservation laws are in one-to-one correspondence with its admitted equivalence classes of linearly-independent variational symmetries modulo gauge symmetries.*

6.3 Computation of Variational Symmetries and Noether Conservation Laws

Whenever a locally variational PDE system (110) is regular, the determining condition (120) for finding variational symmetries $\hat{\mathbf{X}} = P(t,x,u,\partial u,\ldots,\partial^r u)\partial_u$ can be converted into a linear system of equations for $P(t,x,u,\partial u,\ldots,\partial^r u)$ by the following steps.

On the solution space $\mathscr{E}$ of the PDE system, the Fréchet derivative adjoint operator $\delta^*_G|_{\mathscr{E}}$ vanishes. Thus, the determining condition (120) implies $(\delta_P G^{\mathrm{t}})|_{\mathscr{E}} = 0$ which coincides with the determining equation (91) for an infinitesimal symmetry of the PDE system. This shows that P is the characteristic function of an infinitesimal symmetry. From Lemma 4.4, it then follows that P satisfies the relation

$$\delta_P G^{\mathrm{t}} = R_P(G^{\mathrm{t}}) \tag{129}$$

for some linear differential operator

$$R_P = R_P^{(0)} + R_P^{(1)} \cdot D + R_P^{(2)} \cdot D^2 + \cdots + R_P^{(r)} \cdot D^r \tag{130}$$

whose coefficients are non-singular on $\mathscr{E}$, as the PDE system is assumed to be regular, where r is the differential order of P. Note that if the PDE system satisfies a differential identity (88) then R_P is determined by P only up to $\chi\mathscr{D}^{\mathrm{t}}$ where χ is an arbitrary differential function and $\mathscr{D}$ is the linear differential operator defining the identity. Substitution of the relation (129) into the determining condition (120) yields

$$0 = R_P(G^t) + \delta^*_G P^{\mathrm{t}}. \tag{131}$$

Note that $\delta^*_G P^{\mathrm{t}}$ can be expressed in an operator form

$$\delta^*_G P^{\mathrm{t}} = E_u(P)G^{\mathrm{t}} - E_u^{(1)}(P) \cdot (DG^{\mathrm{t}}) + \cdots + E_u^{(r)}(P) \cdot (-D)^r G^{\mathrm{t}} \tag{132}$$

using the relation (78). Consequently, when the PDEs $G = (G^1, \ldots, G^M)$ are expressed in a solved form (54)–(55) for a set of leading derivatives, equation (131) can be split with respect to these leading derivatives and their differential consequences. This yields a linear system of equations

$$0 = R_P^{(k)} + (-1)^k E_u^{(k)}(P), \quad k = 0, 1, \ldots, r. \tag{133}$$

Note that these equations are similar in structure to the Helmholtz conditions (112).

Hence, the following result has been established.

Theorem 6.3 *The determining equation (120) for variational symmetries $\hat{\mathbf{X}} = P(t, x, u, \partial u, \ldots, \partial^r u)\partial_u$ of any locally variational, regular PDE system (110) is equivalent to a linear system of equations consisting of the determining condition (91) for $\hat{\mathbf{X}}$ to be an infinitesimal symmetry of the PDE system, and Helmholtz-type conditions (133) for $\hat{\mathbf{X}}$ to leave any Lagrangian of the PDE system invariant modulo a total divergence. This linear determining system (91), (133) is formulated entirely in terms of the symmetry characteristic function P and the PDE expressions $G = (G^1, \ldots, G^M)$, without explicit use of a Lagrangian.*

It is important to emphasize that the determining system (91), (133) can be solved computationally by the same standard procedure [1–3] that is used to solve the standard determining equation (90) for symmetries.

7 Main Results

For any regular PDE system (39), whether or not it has a variational principle, all local conservation laws have a characteristic form given by multipliers, as shown by the general correspondence stated in Theorem 5.1. In the case of regular PDE systems that are locally variational, the modern form of Noether's theorem given by Theorem 6.2 shows that multipliers for local conservation laws are the same as characteristic functions for variational symmetries. These symmetries satisfy a determining equation (120) which can be split into an equivalent determining system for the symmetry characteristic functions, without explicit use of a Lagrangian, as shown in Theorem 6.3. A similar determining system can be derived for multipliers, by splitting the multiplier determining equation (101) in the same way.

On the solution space $\mathscr{E}$ of a given regular PDE system (39), the Fréchet derivative adjoint operator $\delta^*_G|_{\mathscr{E}}$ vanishes. Thus, the multiplier determining equation (101) implies

$$(\delta^*_Q G)|_{\mathscr{E}} = 0 \tag{134}$$

which is the adjoint of the symmetry determining equation (91), and its solutions $Q(t, x, u, \partial u, \dots, \partial^r u)$ are called *adjoint-symmetries* [6–8] (or sometimes *cosymmetries*). Then Q satisfies the identity

$$\delta^*_Q G = \delta^*_{Q^{\mathrm{t}}} G^{\mathrm{t}} = R_{Q^{\mathrm{t}}}(G^{\mathrm{t}}) \tag{135}$$

from Lemma 4.4, where

$$R_{Q^{\mathrm{t}}} = R^{(0)}_{Q^{\mathrm{t}}} + R^{(1)}_{Q^{\mathrm{t}}} \cdot D + R^{(2)}_{Q^{\mathrm{t}}} \cdot D^2 + \cdots + R^{(r)}_{Q^{\mathrm{t}}} \cdot D^r \tag{136}$$

is some linear differential operator whose coefficients are non-singular on $\mathscr{E}$, and r is the differential order of Q. Note that if the PDE system satisfies a differential identity (88) then $R_{Q^{\mathrm{t}}}$ is determined by Q only up to $\chi\mathscr{D}^{\mathrm{t}}$ where χ is an arbitrary differential function and $\mathscr{D}$ is the linear differential operator defining the identity. The determining equation (101) now becomes

$$0 = R_{Q^{\mathrm{t}}}(G^{\mathrm{t}}) + \delta^*_G Q. \tag{137}$$

From the relation (78), note that $\delta^*_G Q$ can be expressed in an operator form

$$\delta_G^* Q = E_u(Q^{\mathrm{t}})G^{\mathrm{t}} - E_u^{(1)}(Q^{\mathrm{t}}) \cdot (DG^{\mathrm{t}}) + \cdots + E_u^{(r)}(Q^{\mathrm{t}}) \cdot (-D)^r G^{\mathrm{t}}. \tag{138}$$

Consequently, when the PDEs $G = (G^1, \ldots, G^M)$ are expressed in a solved form (54)–(55) in terms of a set of leading derivatives, equation (137) can be split with respect to these leading derivatives and their differential consequences. This yields a linear system of equations

$$0 = R_{Q^{\mathrm{t}}}^{(k)} + (-1)^k E_u^{(k)}(Q^{\mathrm{t}}), \quad k = 0, 1, \ldots, r \tag{139}$$

which is similar in form to the Helmholtz conditions (112).

Thus, the following result has been established.

Theorem 7.4 *The determining equation* (101) *for conservation law multipliers of any regular PDE system* (39) *is equivalent to the linear system of equations* (134)*,* (139)*. In particular, multipliers are adjoint-symmetries* (134) *satisfying Helmholtz-type conditions* (139)*, where these conditions are necessary and sufficient for an adjoint-symmetry* $Q(t, x, u, \partial u, \ldots, \partial^r u)$ *to have the variational form* (98) *derived from a conserved current* $\Phi = (T, X^i)$*.*

A comparison of the determining systems formulated in Theorem 7.4 and Theorem 6.3 shows how the correspondence between the local conservation laws and the multipliers for regular PDE systems is related to the Noether correspondence between the local conservation laws and the variational symmetries for locally variational, regular PDE systems.

Corollary 2 *When a regular PDE system is locally variational* (110)*, the adjoint-symmetry determining equation* (134) *is the same as the symmetry determining equation* (91)*, and the Helmholtz-type conditions* (139) *under which an adjoint-symmetry is a multiplier are equivalent to the variational conditions* (133) *under which a symmetry is a variational symmetry.*

Thus, Theorems 5.1 and 7.4 provide a direct generalization of the modern form of Noether's theorem given by Theorems 6.2 and 6.3, in which the role of symmetries in the derivation of local conservation laws for variational PDE systems is replaced by adjoint-symmetries in the derivation of local conservation laws for non-variational PDE systems.

7.1 Computation of Multipliers and Conserved Currents

For any given regular PDE system, all of its non-trivial local conservation laws (up to equivalence) can be obtained by the following three steps.

Step 1: Solve the determining system (134), (139) to obtain all multipliers.
Step 2: Find all linearly independent equivalence classes of non-trivial multipliers.
Step 3: Construct the conserved current determined by a representative multiplier in each equivalence class.

The multiplier determining system (134), (139) can be solved computationally by the same standard procedure [1–3] that is used to solve the determining equation (91) for symmetries. Moreover, for multipliers of a given differential order r, the multiplier determining system is, in general, more overdetermined than is the symmetry determining equation for infinitesimal symmetries of the same differential order r. Consequently, the computation of multipliers is typically easier than the computation of symmetries.

As an alternative to solving the whole multiplier determining system together, only the adjoint-symmetry determining equation can be solved first, and the Helmholtz-type conditions (139) then can be checked for each adjoint-symmetry to obtain all multipliers.

In practice, it can be computationally hard to obtain the complete solution to the multiplier determining system (or the adjoint-symmetry determining equation) because this will involve going to an arbitrarily high differential order for the dependence of the multiplier (or the adjoint-symmetry) on the derivatives of the dependent variables in the PDE system. Moreover, for computations using computer algebra, this differential order must be specified in advance. The same issue arises when symmetries are being sought, but often these obstacles are set aside by looking for just Lie symmetries, or higher symmetries of a special form.

A similar approach can be used for multipliers, by looking just for all low-order conservation laws or by looking just for higher order conservation laws with a special form or with a particular differential order. In physical applications, there is often a specific class of conserved densities that is of interest. The form for multipliers corresponding to a given class of conserved densities can be derived directly by balancing derivatives on both sides of the characteristic equation, as shown in the running examples in Sect. 5.2.

For each non-trivial multiplier, the construction of a corresponding non-trivial conserved current can be carried out by several different methods.

First, the homotopy integral formula (103)–(104) can be applied. An advantage of this formula compared to the standard linear-homotopy formula in the literature [1, 7, 8] is that the homotopy curve can be adapted to the structure of the expressions for the multiplier Q and the PDE system G, which allows avoiding integration singularities.

Second, the characteristic equation (94) can be converted into a linear system of determining equations for the conserved density $\tilde{T}$ and the flux $\tilde{X}$. The determining equations are derived in a straightforward way starting from the expression for the multiplier Q, similarly to the derivation of the form for low-order conservation laws explained in Sect. 5.2. This method is computationally advantageous as it can be implemented in the same way as setting up and solving the determining system for multipliers [3, 10].

Third, if a given PDE system possesses a scaling symmetry then an algebraic formula that yields a scaling multiple of the conserved current $\Phi = \tilde{\Phi}|_{\mathcal{E}} = (\tilde{T}, \tilde{X})|_{\mathcal{E}}$ is available [9], where the scaling multiple is simply the scaling weight of the corresponding conserved integral. The formula can be derived by applying the scaling relation (84)–(85) directly to the function $f = GQ$. This gives

$$T = \omega\tilde{T}|_{\mathscr{E}} = \Big(P\sum_{l=1}^{k}(-D)^{l-1}\cdot\Big(\frac{\partial G}{\partial(\partial^{l-1}\partial_t u)}Q\Big) + (DP)\cdot\Big(\sum_{l=2}^{k}(-D)^{l-2}\cdot\Big(\frac{\partial G}{\partial(\partial^{l-1}\partial_t u)}Q\Big)\Big) + \cdots + (D^{k-1}P)\cdot\Big(\frac{\partial G}{\partial(\partial^{k-1}\partial_t u)}Q\Big)\Big)\Big|_{\mathscr{E}}, \tag{140}$$

$$X = \omega\tilde{X}|_{\mathscr{E}} = \Big(P\sum_{l=1}^{k}(-D)^{l-1}\cdot\Big(\frac{\partial G}{\partial(\partial^{l-1}\partial_x u)}Q\Big) + (DP)\cdot\Big(\sum_{l=2}^{k}(-D)^{l-2}\cdot\Big(\frac{\partial G}{\partial(\partial^{l-1}\partial_x u)}Q\Big)\Big) + \cdots + (D^{k-1}P)\cdot\Big(\frac{\partial G}{\partial(\partial^{k-1}\partial_x u)}Q\Big)\Big)\Big|_{\mathscr{E}}, \tag{141}$$

modulo a locally trivial current $\Phi_{\text{triv}} = (D_x\Theta, -D_t\Theta + D_x\cdot\Theta)$, where

$$P = \eta - u_t\tau - u_x\cdot\xi, \quad \tau = at, \quad \xi = (b_{(1)}x^1, \ldots, b_{(n)}x^n), \quad \eta = (c_{(1)}u^1, \ldots, c_{(m)}u^m) \tag{142}$$

are the characteristic functions in the generator of the scaling symmetry (82). Here

$$\omega = s + D_t\tau + D_x\cdot\xi = s + a + \sum_{i=1}^{n} b_{(i)} \tag{143}$$

is a scaling factor, with s being the scaling weight of the function GQ. Note, as seen from the characteristic equation (94), ω is equal to the scaling weight of the conserved integral $\int_{\Omega}\tilde{T}|_{\mathscr{E}}\,dV$, as defined on any given spatial domain $\Omega \subseteq \mathbb{R}^n$.

This algebraic formula (140)–(142) has the advantage that it does not require any integrations. However, it assumes that the scaling multiple ω is non-zero, which means that it can be used only for constructing conserved currents whose corresponding conserved integral has a non-zero scaling weight, $\omega \neq 0$.

A more general algebraic construction formula can be derived by utilizing dimensional analysis, which is applicable to PDE systems without a scaling symmetry. Any given PDE system arising in physical applications will be scaling homogeneous under dimensional scaling transformations that act by rescaling the fundamental physical units of all variables and all parametric constants [1, 2] (whether or not the PDE system admits a scaling symmetry). In particular, these dimensional scaling transformations will comprise independent rescalings of length, time, mass, charge, and so on. For each dimensional scaling transformation, a scaling formula will arise for T and X, generalizing the algebraic formula (140)–(142) in a way that involves the dependence of Q and G on all of the dimensionful parametric constants appearing in their expressions. If a conserved integral represents a dimensionful physical quantity, then the scaling multiple in the resulting formula will be non-zero.

A derivation of this general construction formula will be given elsewhere [33]. Here, it will be illustrated in a running example.

Running Ex. (1) All low-order conservation laws will now be derived for the gKdV equation (56). As shown previously, low-order conserved currents correspond to low-order multipliers, which have the general form $Q(t, x, u, u_x, u_{xx})$. Multipliers are adjoint-symmetries that satisfy Helmholtz-type conditions. To set up the determining system for multipliers, first note $\delta^*_Q G = -(D_t Q + D_x^3 Q + u^p D_x Q)$, where $G = u_t + u^p u_x + u_{xxx}$. Hence the adjoint-symmetry determining equation for Q is given by

$$(D_t Q + D_x^3 Q + u^p D_x Q)|_{\mathscr{E}} = 0.$$

Next look at the terms that contain the leading derivative u_t and its x-derivatives in this equation. This yields

$$-D_t Q + D_x^3 Q + u^p D_x Q = -\frac{\partial Q}{\partial u} - \frac{\partial Q}{\partial u_x} D_x G - \frac{\partial Q}{\partial u_{xx}} D_x^2 G = R_Q(G)$$

holding off of the gKdV solution space, where the components of the operator R_Q are given by

$$R_Q^{(0)} = -\frac{\partial Q}{\partial u}, \quad R_Q^{(x)} = -\frac{\partial Q}{\partial u_x}, \quad R_Q^{(x,x)} = -\frac{\partial Q}{\partial u_{xx}}.$$

Then the Helmholtz-type equations on Q consist of

$$\begin{aligned} 0 =& R_Q^{(0)} + E_u(Q) = -D_x \frac{\partial Q}{\partial u_x} + D_x^2 \frac{\partial Q}{\partial u_{xx}}, \\ 0 =& R_Q^{(x)} - E_u^{(x)}(Q) = -2\frac{\partial Q}{\partial u_x} + 2D_x \frac{\partial Q}{\partial u_{xx}}, \\ 0 =& R_Q^{(x,x)} + E_u^{(x,x)}(Q^{\mathrm{t}}) = 0, \end{aligned}$$

which reduce to a single equation

$$D_x \frac{\partial Q}{\partial u_{xx}} - \frac{\partial Q}{\partial u_x} = 0.$$

This Helmholtz-type equation and the adjoint-symmetry equation can be split with respect to all derivatives of u which do not appear in Q, with u_t eliminated through the gKdV equation. This gives, after some simplifications, a linear overdetermined system of 8 equations:

$$\frac{\partial Q}{\partial u_x} = 0, \quad \frac{\partial^2 Q}{\partial u_{xx}^2} = 0, \quad \frac{\partial^2 Q}{\partial x \partial u_{xx}} = 0, \quad \frac{\partial^2 Q}{\partial u \partial u_{xx}} = 0, \quad \frac{\partial^3 Q}{\partial x \partial u^2} = 0,$$

$$\frac{\partial^3 Q}{\partial u^3} - p(p-1)u^{p-2}\frac{\partial Q}{\partial u_{xx}} = 0, \quad \frac{\partial^3 Q}{\partial x^2 \partial u} + u_{xx}\frac{\partial^2 Q}{\partial u^2} - pu^{p-1}u_{xx}\frac{\partial Q}{\partial u_{xx}} = 0,$$

$$\frac{\partial Q}{\partial t} + u^p\frac{\partial Q}{\partial x} + \frac{\partial^3 Q}{\partial x^3} + 3u_{xx}\frac{\partial^2 Q}{\partial x \partial u} = 0.$$

These equations can be solved for Q, with p treated as an unknown, to get

$$Q = c_1 + c_2 u + c_3(u_{xx} + \tfrac{1}{p+1}u^{p+1}) + c_4(x - tu) + c_5(t(3u_{xx} + u^3) - xu)$$

with $c_4 = 0$ if $p \neq 1$, and $c_5 = 0$ if $p \neq 2$. Hence, 5 low-order multipliers are obtained,

$$Q_1 = 1, \quad Q_2 = u, \quad Q_3 = u_{xx} + \tfrac{1}{p+1}u^{p+1}, \quad p > 0,$$
$$Q_4 = x - tu, \quad p = 1,$$
$$Q_5 = t(3u_{xx} + u^3) - xu, \quad p = 2.$$

The corresponding low-order conserved currents will now be derived using the three different construction methods. First is the homotopy integral method. The simplest choice for the homotopy is $u_{(\lambda)} = \lambda u$ since the gKdV equation is a homogeneous PDE, $G|_{u=0} = 0$. Hence the homotopy integral is simply given by

$$\begin{aligned}\tilde{T} &= \int_0^1 u\frac{\partial(GQ)}{\partial u_t}\Big|_{u=u_{(\lambda)}} d\lambda \\ &= \int_0^1 u\big(c_1 + c_4 x + (c_2 - c_4 t - c_5 x)u\lambda + (c_3 + c_5 3t)u_{xx}\lambda \\ &\qquad + c_5 t u^3\lambda^3 + c_3\tfrac{1}{p+1}u^{p+1}\lambda^{p+1}\big)\, d\lambda \\ &= (c_1 + c_4 x)u + \tfrac{1}{2}(c_2 - c_4 t - c_5 x)u^2 + \tfrac{1}{2}(c_3 + c_5 3t)uu_{xx} \\ &\qquad + c_5\tfrac{1}{4}tu^4 + c_3\tfrac{1}{(p+1)(p+2)}u^{p+2}\end{aligned}$$

and

$$\begin{aligned}\tilde{X} = \int_0^1 \Big(& u\Big(\frac{\partial(GQ)}{\partial u_x}\Big|_{u=u_{(\lambda)}} - D_x\frac{\partial(GQ)}{\partial u_{xx}}\Big|_{u=u_{(\lambda)}} + D_x^2\frac{\partial(GQ)}{\partial u_{xxx}}\Big|_{u=u_{(\lambda)}}\Big) \\ &+ u_x\Big(\frac{\partial(GQ)}{\partial u_{xx}}\Big|_{u=u_{(\lambda)}} - D_x\frac{\partial(GQ)}{\partial u_{xxx}}\Big|_{u=u_{(\lambda)}}\Big) + u_{xx}\frac{\partial(GQ)}{\partial u_{xxx}}\Big|_{u=u_{(\lambda)}}\Big) d\lambda\end{aligned}$$

$$\begin{aligned}
&= \int_0^1 \Big(u\big((c_4xu - 2c_5u_x - (c_3+c_5t)u_{tx} - (c_4t+c_5x - c_2)u_{xx})\lambda - c_4tu^2\lambda^2 \\
&\quad + c_5(3u^2u_{xx} - xu^3)\lambda^3 + c_5tu^5\lambda^5 + c_1 \\
&\quad + u^p\lambda^p - (c_5 3ptu^{p-1}u_x^2 + (c_2u^{p+1} + c_3u^pu_{xx})\lambda^{p+1} \\
&\quad + c_3\tfrac{1}{p+1}u^{2p+1}\lambda^{2p+1}\big) + u_x\big(- c_4 + (c_5u + (c_3 + c_5 3t)u_t \\
&\quad + (c_4t + c_5x - c_2)u_x)\lambda\big) + u_{xx}\big(c_1 + c_4x + (c_2 - c_4t - c_5x)u \\
&\quad + (c_3 + c_5 3t)u_{xx})\lambda + c_5tu^3\lambda^3 + c_3\tfrac{1}{p+1}u^{p+1}\lambda^{p+1}\big)\Big)d\lambda
\end{aligned}$$

which is easiest to evaluate when separated into the non-overlapping cases $p = 1$ with $c_5 = c_1 = c_2 = c_3 = 0$, $p = 2$ with $c_4 = c_1 = c_2 = c_3 = 0$, and $p > 0$ with $c_4 = c_5 = 0$. This yields the 5 low-order conserved currents

$$\begin{aligned}
&\tilde{T}_1 = u, \quad \tilde{X}_1 = \tfrac{1}{p+1}u^{p+1} + u_{xx} \\
&\tilde{T}_2 = \tfrac{1}{2}u^2, \quad \tilde{X}_2 = \tfrac{1}{p+2}u^{p+2} + uu_{xx} - \tfrac{1}{2}u_x^2 \\
&\tilde{T}_3 = \tfrac{1}{2}uu_{xx} + \tfrac{1}{(p+1)(p+2)}u^{p+2}, \quad \tilde{X}_3 = \tfrac{1}{2(p+1)^2}u^{2p+2} + \tfrac{1}{p+1}u^{p+1}u_{xx} \\
&\qquad\qquad\qquad\qquad\qquad\qquad + \tfrac{1}{2}(u_{xx}^2 + u_tu_x) - uu_{tx} \\
&\tilde{T}_4 = xu - \tfrac{1}{2}tu^2, \quad \tilde{X}_4 = t(\tfrac{1}{2}u_x^2 - uu_{xx} - \tfrac{1}{3}u^3) + x(u_{xx} + \tfrac{1}{2}u^2) - u_x, \quad p = 1 \\
&\tilde{T}_5 = \tfrac{1}{2}(3tuu_{xx} - xu^2) + \tfrac{1}{4}tu^4, \quad \tilde{X}_5 = t(\tfrac{3}{2}(u_{xx}^2 + u_tu_x) + u^3u_{xx} - \tfrac{3}{2}uu_{tx} + \tfrac{1}{6}u^6) \\
&\qquad\qquad\qquad\qquad\qquad\qquad + x(\tfrac{1}{2}u_x^2 - uu_{xx} - \tfrac{1}{4}u^4) - \tfrac{1}{2}uu_x, \quad p = 2
\end{aligned}$$

whose respective multipliers are $Q_1, \ldots, Q_5$. Each of these conserved currents is in characteristic form, namely $D_t\tilde{T}_i + D_x\tilde{X}_i = Q_iG$.

Second is the integration method using the characteristic equation $D_t\tilde{T} + D_x\tilde{X} = GQ$, where

$$\begin{aligned}
GQ = &\big(c_1 + c_2u + c_3(u_{xx} + \tfrac{1}{p+1}u^{p+1}) + c_4(x - tu) \\
&+ c_5(t(3u_{xx} + u^3) - xu)\big)(u_t + u^pu_x + u_{xxx})
\end{aligned}$$

with $c_4 = 0$ if $p \neq 1$, and $c_5 = 0$ if $p \neq 2$. There are three steps in this method. First, as shown previously from balancing derivatives on both sides of the characteristic equation, the general form for all low-order conserved currents $\tilde{\Phi} = (\tilde{T}, \tilde{X})$ is found to be given by

$$\tilde{\Phi}|_{\mathscr{E}} = (\tilde{T}(t, x, u, u_x), \tilde{X}(t, x, u, u_t, u_x, u_{xx})).$$

Second, the characteristic equation can then be split with respect to u_{tx} and u_{xxx}, which yields (after simplifications) a linear overdetermined system of three equations:

$$\frac{\partial \tilde{T}}{\partial u_x} + \frac{\partial \tilde{X}}{\partial u_t} = 0,$$

$$\frac{\partial \tilde{X}}{\partial u_{xx}} = c_3 \tfrac{1}{p+1} u^{p+1} + c_1 + c_4 x + (c_2 - c_4 t - c_5 x)u + (c_3 + c_5 3t)u_{xx} + c_5 t u^3,$$

$$\begin{aligned} \frac{\partial \tilde{T}}{\partial t} + \frac{\partial \tilde{X}}{\partial x} + u_t \frac{\partial \tilde{T}}{du} + u_x \frac{\partial \tilde{X}}{\partial u} + u_{xx} \frac{\partial \tilde{X}}{\partial u_x} = (u_t +_u P_{u_x}) \\ (c_3 \tfrac{1}{p+1} u^{p+1} + c_1 + c_4 x + (c_2 - c_4 t - c_5 x)u \\ + (c_3 + c_5 3t)u_{xx} + c_5 t u^3). \end{aligned}$$

These equations can be integrated directly. It is simplest to consider separately the non-overlapping cases $p = 1$ with $c_5 = c_1 = c_2 = c_3 = 0$, $p = 2$ with $c_4 = c_1 = c_2 = c_3 = 0$, and $p > 0$ with $c_4 = c_5 = 0$. The first case is found to reproduce $\tilde{T}_4$ and $\tilde{X}_4$; the second case yields $\tilde{T}_5 - D_x \tilde{\Theta}_5$ and $\tilde{X}_5 + D_t \tilde{\Theta}_5$ where $\tilde{\Theta}_5 = \frac{3}{2} tuu_x$. Similarly, the third case with $c_3 = 0$ is found to reproduce $\tilde{T}_1$, $\tilde{T}_2$, $\tilde{X}_1$, $\tilde{X}_2$, and with $c_3 \neq 0$ it yields $\tilde{T}_3 - D_x \tilde{\Theta}_3$ and $\tilde{X}_3 + D_t \tilde{\Theta}_3$ where $\tilde{\Theta}_3 = \frac{1}{2} uu_x$. Thus, the resulting conserved currents agree with those obtained from the homotopy integral, up to locally trivial currents. In particular, path of these currents is in characteristic form.

Third is the scaling symmetry method. The gKdV equation possesses a scaling symmetry

$$t \to \lambda^3 t, \quad x \to \lambda x, \quad u \to \lambda^{-2/p} u, \quad \lambda \neq 0$$

with the characteristic function $P = -(2/p)u - 3tu_t - xu_x$. Note the multipliers $Q_1, \ldots, Q_5$ are each homogeneous under the scaling symmetry, with respective scaling weights $q_1 = 0$, $q_2 = -2/p$, $q_3 = -2 - 2/p$, $q_4 = 1$, $q_5 = 0$. Hence the corresponding scaling factors (143) are given by $\omega_1 = 1 - 2/p$, $\omega_2 = 1 - 4/p$, $\omega_3 = -1 - 4/p$, $\omega_4 = 0$, $\omega_5 = 0$, where $s_i = q_i + c - a, a = 3, b = 1, c = -2/p$. Then the scaling symmetry formula is given by

$$T_i = \omega_i \tilde{T}_i|_{\mathscr{E}} = \Big(P \frac{\partial G}{\partial u_t} Q_i\Big)\Big|_{\mathscr{E}},$$

$$\begin{aligned} X_i = \omega_i \tilde{X}_i|_{\mathscr{E}} = \Big(P\Big(\frac{\partial G}{\partial u_x} Q_i - D_x\Big(\frac{\partial G}{\partial u_{xx}} Q_i\Big) + D_x^2\Big(\frac{\partial G}{\partial u_{xxx}} Q_i\Big)\Big) \\ + D_x P\Big(\frac{\partial G}{\partial u_{xx}} Q_i - D_x\Big(\frac{\partial G}{\partial u_{xxx}} Q_i\Big)\Big) + D_x^2 P\Big(\frac{\partial G}{\partial u_{xxx}} Q_i\Big)\Big)\Big|_{\mathscr{E}}, \end{aligned}$$

modulo a locally trivial current. For $i = 1, 2, 3$, this yields the conserved density expressions

$$
\begin{gathered}
T_1 = -(\tfrac{2}{p}u + 3tu_t + xu_x)|_{\mathscr{E}} = (1-2/p)\tilde{T}_1|_{\mathscr{E}} + D_x\Theta_1, \quad \Theta_1 = 3t\tilde{X}_1 - x\tilde{T}_1, \\
T_2 = -((\tfrac{2}{p}u + 3tu_t + xu_x)u)|_{\mathscr{E}} = (1-4/p)T_2|_{\mathscr{E}} + D_x\Theta_2, \quad \Theta_2 = 3t\tilde{X}_2 - x\tilde{T}_2, \\
T_3 = -(\tfrac{2}{p}u + 3tu_t + xu_x)(u_{xx} + \tfrac{1}{p+1}u^{p+1})|_{\mathscr{E}} = (-1-4/p)T_3|_{\mathscr{E}} + D_x\Theta_3, \\
\Theta_3 = \tfrac{1}{2}(1+4/p)uu_x + 3t(\tilde{X}_3 + D_t\tilde{\Theta}_3) - \tilde{X}(\tilde{T}_3 - D_x\tilde{\Theta}_3).
\end{gathered}
$$

Note their scaling factors are non-zero when $p \neq 2$, $p \neq 4$, and $p \neq -4$, respectively. When $p = 2$, T_1 reduces to a locally trivial conserved density $D_x\Theta_1$ and when $p = 4$, T_2 reduces to a locally trivial conserved density $D_x\Theta_2$. Likewise, when $p = -4$, T_3 reduces to a locally trivial conserved density $D_x\Theta_3$.

The expressions given by the scaling symmetry formula for $i = 4, 5$ yield

$$
\begin{aligned}
T_4 &= -(2u + 3tu_t + xu_x)(x - tu)|_{\mathscr{E}} = D_x\Theta_4, \\
\Theta_4 &= (x - tu)(t(3u_{xx} + u^2) - xu) + \tfrac{3}{2}(tu_x - 1)^2
\end{aligned}
$$

and

$$
\begin{aligned}
T_5 &= -(u + 3tu_t + xu_x)(3t(u_{xx} + u^3) - xu)|_{\mathscr{E}} = D_x\Theta_5, \\
\Theta_5 &= \tfrac{1}{2}(t(3u_{xx} + 4^3) - xu)^2.
\end{aligned}
$$

These cases for $p > 0$ in which the scaling symmetry formula yields locally trivial currents are called the critical powers for the corresponding conserved currents. To obtain the conserved currents for a critical power, it is necessary to use the more general dimensional scaling formula.

Several steps are needed to set up the dimensional scaling formula.

The first step is to introduce dimensionful constants into the gKdV equation so that it is homogeneous under separate dimensional scalings of t [time], x [length], and u [mass]. Thus, let

$$
\tilde{G} = u_t + \mu u^p u_x + \nu u_{xxx}, \quad \mu, \nu = \text{const.}
$$

where μ has dimensions of $[\text{time}]^{-1}[\text{length}][\text{mass}]^{-p}$, and ν has dimensions of $[\text{time}]^{-1}[\text{length}]^3$. Note $\tilde{G} = G$ will be the gKdV equation when these constants have the numerical values $\mu = 1$ and $\nu = 1$.

The next step is to insert factors of μ and ν into the expressions for the low-order multipliers so that $Q_1, \ldots, Q_5$ are each dimensionally homogeneous:

$$
\begin{gathered}
Q_1 = 1, \quad Q_2 = u, \quad Q_3 = \nu u_{xx} + \tfrac{1}{p+1}\mu u^{p+1}, \quad p > 0, \\
Q_4 = x - \mu tu, \quad p = 1, \\
Q_5 = t(3\nu u_{xx} + \mu u^3) - xu, \quad p = 2.
\end{gathered}
$$

The main step consists of generalizing the scaling relation (84)–(85) so that it applies to dimensional scaling transformations. These transformations are given by

$$t \to \lambda t, \quad \mu \to \lambda^{-1}\mu, \quad \nu \to \lambda^{-1}\nu;$$
$$x \to \lambda x, \quad \mu \to \lambda\mu, \quad \nu \to \lambda^{3}\nu;$$
$$u \to \lambda u, \quad \mu \to \lambda^{-p}\mu, \quad \nu \to \nu;$$

as determined by the dimensions of μ and ν. Since the scaling relation (84)–(85) only holds for variables in jet space, the constants μ and ν now must be treated as variables by introducing the equations

$$\tilde{G}^{(\mu)} = (\mu_t, \mu_x) = 0, \quad \tilde{G}^{(\nu)} = (\nu_t, \nu_x) = 0.$$

Then the augmented PDE system

$$\tilde{G} = 0, \quad \tilde{G}^{(\mu)} = 0, \quad \tilde{G}^{(\nu)} = 0$$

will admit each of the three scaling transformations as symmetries formulated in the augmented jet space $\tilde{J} = (t, x, u, \mu, \nu, u_t, u_x, \mu_t, \mu_x, \nu_t, \nu_x, \ldots)$. Note that the characteristic equation for conserved currents will have additional multiplier terms

$$D_t\tilde{T} + D_x\tilde{X} = \tilde{G}Q + \tilde{G}^{(\mu)}\tilde{Q}_{(\mu)} + \tilde{G}^{(\nu)}\tilde{Q}_{(\nu)}$$

for some expressions $\tilde{Q}_{(\mu)} = (\tilde{Q}^t_{(\mu)}, \tilde{Q}^x_{(\mu)})^{\mathrm{t}}$ and $\tilde{Q}_{(\nu)} = (\tilde{Q}^t_{(\nu)}, \tilde{Q}^x_{(\nu)})^{\mathrm{t}}$, where Q is unchanged. These expressions can be found in a straightforward way by setting up and solving the multiplier determining system, with $Q = Q_i$ being the previously derived low-order multipliers for the gKdV equation. Since μ and ν appear linearly in each Q_i as well as in the PDE expression $\tilde{G}$, note $\tilde{Q}_{(\mu)}$ and $\tilde{Q}_{(\nu)}$ can have at most linear dependence on these variables and cannot contain any derivatives of these variables. Also, since Q_i depends on u, u_x, u_{xx}, u_{xxx}, and $\tilde{G}$ depends on u, u_t, u_x, u_{xxx}, note $\tilde{Q}_{(\mu)}$ and $\tilde{Q}_{(\nu)}$ can depend on only $u, u_t, u_x, u_{tx}, u_{xx}$ in addition to t, x and μ, ν:

$$\tilde{Q}_{(\mu)}(t, x, u, \mu, \nu, u_t, u_x, u_{tx}, u_{xx}), \quad \tilde{Q}_{(\nu)}(t, x, u, \mu, \nu, u_t, u_x, u_{tx}, u_{xx}).$$

The multiplier determining system is then given by

$$E_u(\tilde{G}Q_i + \tilde{G}^{(\mu)}\tilde{Q}_{(\mu)} + \tilde{G}^{(\nu)}\tilde{Q}_{(\nu)}) = 0,$$
$$E_{(\mu)}(\tilde{G}Q_i + \tilde{G}^{(\mu)}\tilde{Q}_{(\mu)} + \tilde{G}^{(\nu)}\tilde{Q}_{(\nu)}) = 0, \quad E_{(\nu)}(\tilde{G}Q_i + \tilde{G}^{(\mu)}\tilde{Q}_{(\mu)} + \tilde{G}^{(\nu)}\tilde{Q}_{(\nu)}) = 0$$

for $i = 1, \ldots, 5$. This system splits with respect to all derivatives of u, μ, ν which do not appear in $\tilde{Q}_{(\mu)}$ and $\tilde{Q}_{(\nu)}$. Integration of the resulting equations yields

$$
\begin{gathered}
\tilde{Q}^t_{1(\mu)} = 0, \quad \tilde{Q}^x_{1(\mu)} = \tfrac{1}{p+1}u^{p+1}, \quad \tilde{Q}^t_{1(\nu)} = 0, \quad \tilde{Q}^x_{1(\nu)} = u_{xx},\\
\tilde{Q}^t_{2(\mu)} = 0, \quad \tilde{Q}^x_{2(\mu)} = \tfrac{1}{p+2}u^{p+2}, \quad \tilde{Q}^t_{2(\nu)} = 0, \quad \tilde{Q}^x_{2(\nu)} = uu_{xx} - \tfrac{1}{2}u_x^2,\\
\tilde{Q}^t_{3(\mu)} = \tfrac{1}{(p+1)(p+2)}u^{p+2}, \quad \tilde{Q}^x_{3(\mu)} = \tfrac{1}{p+1}\nu u^{p+1} + \tfrac{1}{(p+1)(p+2)}\mu u^{2p+2},\\
\tilde{Q}^t_{3(\nu)} = -\tfrac{1}{2}u_x^2, \quad \tilde{Q}^x_{3(\nu)} = \nu u_{xx}^2 + u_t u_x + \mu u^{p+1}u_{xx},\\
\tilde{Q}^t_{4(\mu)} = -\tfrac{1}{2}tu^2, \quad \tilde{Q}^x_{4(\mu)} = t\nu(\tfrac{1}{2}u_x^2 - uu_{xx}) + \tfrac{1}{2}xu^2 - \tfrac{2}{3}\mu u^3,\\
\tilde{Q}^t_{4(\nu)} = 0, \quad \tilde{Q}^x_{4(\nu)} = t\mu(\tfrac{1}{2}u_x^2 - uu_{xx}) + xu_{xx} - u_x,\\
\tilde{Q}^t_{5(\mu)} = \tfrac{1}{4}tu^2, \quad \tilde{Q}^x_{5(\mu)} = t(\nu u^3 u_{xx} + \tfrac{1}{3}\mu u^6) - \tfrac{1}{4}xu^4,\\
\tilde{Q}^t_{5(\nu)} = -\tfrac{3}{2}tu_x^2, \quad \tilde{Q}^x_{5(\nu)} = t(\mu u^3 u_{xx} + 3\nu u_{xx}^2 + 3u_t u_x) + x(\tfrac{1}{2}u_x^2 - uu_{xx}) + uu_x.
\end{gathered}
$$

The scaling relation (84)–(85) can now be applied to the function $f_i = \tilde{G}Q_i + \tilde{G}^{(\mu)}\tilde{Q}_{i(\mu)} + \tilde{G}^{(\nu)}\tilde{Q}_{i(\nu)}$ in the augmented jet space $\tilde{J} = (t, x, u, \mu, \nu, u_t, u_x, \mu_t, \mu_x, \nu_t, \nu_x, \ldots)$ by using an infinitesimal scaling symmetry given by one of the scaling transformation generators

$$
\begin{aligned}
\hat{\mathbf{X}}_{\text{time}} &= -(\mu + t\mu_t)\partial_\mu - (\nu + t\nu_t)\partial_\nu - tu_t\partial_u,\\
\hat{\mathbf{X}}_{\text{length}} &= (\mu - x\mu_x)\partial_\mu + (3\nu - x\nu_x)\partial_\nu - xu_x\partial_u,\\
\hat{\mathbf{X}}_{\text{mass}} &= -p\mu\partial_\mu + u\partial_u.
\end{aligned}
$$

Let P, $P^{(\mu)}$, $P^{(\nu)}$ denote the characteristic functions in the selected scaling transformation generator $\hat{\mathbf{X}}$. Then this yields the dimensional scaling formula

$$
\begin{aligned}
T_i = \omega_i \tilde{T}_i|_{\mathscr{E}} &= \Big(P\frac{\partial \tilde{G}}{\partial u_t}Q_i + P^{(\mu)}\tilde{Q}^t_{i(\mu)} + P^{(\nu)}\tilde{Q}^t_{i(\nu)}\Big),\\
X_i = \omega_i \tilde{X}_i|_{\mathscr{E}} &= \Big(P\Big(\frac{\partial \tilde{G}}{\partial u_x}Q_i - D_x\Big(\frac{\partial \tilde{G}}{\partial u_{xx}}Q_i\Big) + D_x^2\Big(\frac{\partial \tilde{G}}{\partial u_{xxx}}Q_i\Big)\Big)\\
&\qquad + D_xP\Big(\frac{\partial \tilde{G}}{\partial u_{xx}}Q_i - D_x\Big(\frac{\partial \tilde{G}}{\partial u_{xxx}}Q_i\Big)\Big) + D_x^2P\Big(\frac{\partial \tilde{G}}{\partial u_{xxx}}Q_i\Big)\\
&\qquad + P^{(\mu)}\tilde{Q}^x_{i(\mu)} + P^{(\nu)}\tilde{Q}^x_{i(\nu)}\Big),
\end{aligned}
$$

modulo a locally trivial current, where

$$
\omega_i = q_i + s + D_t\tau + D_x\xi \tag{144}
$$

is a scaling factor defined in terms of the scaling weights q_i, s of $Q_i, \tilde{G}$ and the divergence factor $D_t\tau + D_x\xi$ arising from the selected dimensional scaling

Table 1 Properties of dimensional scaling transformations for the gKdV equation and its low-order multipliers

	P	$P^{(\mu)}$	$P^{(\nu)}$	$D_t\tau + D_x\xi$	s	q_1	q_2	q_3	q_4	q_5
Time	$-tu_t$	$-(\mu + t\mu_t)$	$-(\nu + t\nu_t)$	1	-1	0	0	-1	0	0
Length	$-xu_x$	$\mu - x\mu_x$	$3\nu - x\nu_x$	1	0	0	0	1	1	1
Mass	u	$-p\mu$	0	0	1	0	1	1	0	1

Table 2 Dimensional scaling weights for low-order conserved currents of the gKdV equation

	ω_1	ω_2	ω_3	ω_4	ω_5
Time	0	0	-1	0	0
Length	1	1	2	2	2
Mass	1	2	2	1	2

transformation. In particular, for each $i = 1, \ldots, 5$, there will be some (possibly combined) transformation such that the scaling factor w_i is non-zero, as seen from Tables 1 and 2.

The dimensional scaling formula will now be used to obtain the conserved currents $\Phi_i = (T_i, X_i)|_{\mathscr{E}}$ that were missed previously by the scaling symmetry formula. These cases are: $i = 4, 5$; and, $i = 1, 2$ when p is a critical power. From the form of the dimensional scaling generators, the mass scaling transformation is simplest choice to use. Then the formula becomes

$$T_i = \omega_i T_i = (uQ_i - p\tilde{Q}^t_{i(\mu)})|_{\mu=\nu=1},$$

$$X_i = \omega_i X_i = (u^{p+1}Q_i + uD_x^2 Q_i - u_x D_x Q_i + u_{xx} Q_i - p\tilde{Q}^x_{i(\mu)})|_{\mu=\nu=1},$$

modulo a locally trivial current. This mass scaling formula yields the conserved density and flux expressions

$$T_1 = \tilde{T}_1, \quad X_1 = \tilde{X}_1,$$

$$T_2 = 2\tilde{T}_2, \quad X_2 = \tilde{X}_2,$$

which hold for all powers $p > 0$ (including the critical powers $p = 2$ and $p = 4$, respectively), and also

$$T_4 = \tilde{T}_4 \qquad X_4 = \tilde{X}_4,$$

$$-u_x - D_t\Theta_4$$

$$T_5 = 2\tilde{T}_5, \quad X_5 = 2\tilde{X}_5,$$

8 Concluding Remarks

The main results presented in Sect. 7 provide a broad generalization of Noether's theorem in modern form using multipliers, yielding a general method which is applicable to all typical PDE systems arising in physical applications. In this generalization, the problem of finding all conservation laws for a given PDE system becomes an adjoint version of the problem of finding all infinitesimal symmetries of the PDE system.

For any given variational PDE system, conservation laws arise from variational symmetries, which are infinitesimal symmetries that satisfy variational conditions corresponding to invariance of any variational principle for the PDE system. Noether's theorem shows that the characteristic functions in a variational symmetry are precisely the component functions in a multiplier. For any given non-variational PDE system, the role of symmetries in the derivation of conservation laws is replaced by adjoint-symmetries, and the variational conditions under which an infinitesimal symmetry is a variational symmetry are replaced by Helmholtz-type conditions under which an adjoint-symmetry is a multiplier. Also, the role of a Lagrangian in constructing a conserved integral from a variational symmetry is replaced by several different constructions: an explicit integral formula, an explicit algebraic scaling formula, and a system of determining equations, all of which use only a multiplier and the given PDE system itself.

Most importantly, the completeness of this general method in finding all conservation laws for a given PDE system is established by working with the system expressed in a solved-form for a set of leading derivatives without restricting it to have a generalized Cauchy-Kovalevskaya form. This means that the method applies equally well to PDE systems that possess differential identities.

As a consequence, there is no need to use special methods or ansatzes for determining the conservation laws of any given PDE system, just as there is no necessity to use special methods or ansatzes for finding its symmetries.

The formulation of the general method as a generalization of Noether's theorem rests on the adjoint relationship between variational symmetries and multipliers, which originates from the algebraic relationship between symmetries and adjoint-symmetries. An interesting question is whether this algebraic relationship has a geometrical interpretation.

As will be shown in more detail elsewhere [34], adjoint-symmetries indeed can be given a simple geometrical meaning. In the case of PDE systems comprised of dynamical evolution equations, $G = \partial_t u - g(t, x, u, \partial_x u, \partial_x^2 u, \ldots, \partial_x^N u) = 0$, an adjoint-symmetry defines a 1-form (or covector field) $\boldsymbol{Q} \boldsymbol{d} \boldsymbol{u}$ that is invariant under the dynamical flow on $u(t,x)$, similarly to how a symmetry $P\partial_u$ defines an invariant vector field. This geometrical statement essentially relies on the number of dependent variables being the same as the number of equations in the PDE system. For general PDE systems $G = 0$, it seems necessary to use the well-known procedure [1] of embedding the PDE system into a larger, variational system defined by a Lagrangian $L = Gv^{\mathrm{t}}$ where v denotes additional dependent variables which are paired with the equations $G = 0$ in the given PDE system. In this setting, an adjoint-symmetry defines a symmetry vector field $Q\partial_v$ of the enlarged system, $G = 0$ and $G'^*(v) = 0$, where G' is the Fréchet derivative of G, and $G'*$ is its adjoint. Then, it is straightforward to show that an adjoint-symmetry is a multiplier precisely when $Q\partial_v$ is a variational symmetry.

Acknowledgements S.C. Anco is supported by an NSERC research grant. The referees are thanked for valuable comments which have improved this work.

References

1. P. Olver, *Applications of Lie Groups to Differential Equations* (Springer-Verlag, New York, 1986).
2. G. Bluman and S.C. Anco, *Symmetry and Integration Methods for Differential Equations*, Springer Applied Mathematics Series 154 (Springer-Verlag, New York, 2002).
3. G. Bluman, A. Cheviakov, S.C. Anco, *Applications of Symmetry Methods to Partial Differential Equations*, Springer Applied Mathematics Series 168, (Springer, New York, 2010).
4. L. Martinez Alonso, *Lett. Math. Phys.* 3, 419–424 (1979).
5. Y. Kosmann-Schwarzbach, *The Noether theorems. Invariance and conservation laws in the twentieth century*, Sources and Studies in the History of Mathematics and Physical Sciences (translated, revised, and augmented by B.E. Schwarzbach), (Springer, New York, 2011).
6. S.C. Anco and G. Bluman, *Phys. Rev. Lett.* 78, 2869–2873 (1997).
7. S.C. Anco and G. Bluman, *Euro. J. Appl. Math.* 13, 545–566 (2002).
8. S.C. Anco and G. Bluman, *Euro. J. Appl. Math.* 13, 567–585 (2002).
9. S.C. Anco, *J. Phys. A: Math. and Gen.* 36, 8623–8638 (2003).
10. T. Wolf, *Euro. J. Appl. Math.* 13, 129–152 (2002).
11. C. Morawetz, *Bulletin Amer. Math. Soc.* 37(2), 141–154 (2000).
12. A.H. Kara and F.M. Mahomed, *Nonlin. Dyn.* 45, 367–383 (2006).
13. N.H. Ibragimov, *J. Math. Anal. Appl.* 333, 311–328 (2007).
14. N.H. Ibragimov, *J. Phys. A: Math. and Theor.* 44, 432002 (2011).
15. D. Poole and W. Hereman, *J. Symbolic Computation* 46, 1355–1377 (2011).
16. S.C. Anco and J. Pohjanpelto, Classification of local conservation laws of Maxwell's equations, *Acta. Appl. Math.* 69, 285–327 (2001).
17. I.S. Krasil'shchik and A.M. Vinogradov (eds.), *Symmetries and Conseervation Laws for Differential Equations of Mathematical Physics*, Translations of Mathematical Monographs 182, Amer. Math. Soc.:Providence, 1999.
18. A. Verbotevsky, in: *Secondary Calculus and Cohomological Physics*s, 211–232, Contemporary Mathematics 219, Amer. Math. Soc.:Providence, 1997.
19. G. Barnich, F. Brandt, and M. Henneaux, *Commun. Math. Phys.* 174 (1994), 57–91.
20. L.C. Evans, *Partial Differential Equations*, Graduate Studies in Mathematics 19, (Amer. Math. Soc., Providence, 1998).
21. A. Degaspersis, A.N.W. Hone, D.D. Holm, in: *Nonlinear Physics: Theory and Experiment II*, 37–43, (eds. M.J. Ablowitz, M. Boiti, F. Pempinelli, and B. Prinari) World Scientific (2003).
22. L.V. Ovsiannikov, *Group Analysis of Differential Equations* (Academic Press, New York, 1982).
23. R.L. Anderson and N.H. Ibragimov, *Lie-Bäcklund Transformations in Applications* (SIAM Studies in Applied Mathematics, 1979).
24. N.H. Ibragimov, *Transformation groups applied to mathematical physics* (Reidel, Dordrecht, 1985).

25. N.H. Ibragimov (ed.), *CRC Handbook of Lie Group Analysis of Differential Equations* Volumes I,II,III (CRC Press, 1994–1996).
26. J. Nestruev, *Smooth manifolds and observables*, (Springer, New York, 2002).
27. S.C. Anco and A. Kara, Euro. J. Appl. Math. (in press) (2017).
28. S.C. Anco, Int. J. Mod. Phys. B 30, 164004 (2016).
29. S.C. Anco, Symmetry 9(3), 33 (2017).
30. R.S. Khamitara, Teoret, Mat. Fiz. 52(2), 244–251 (1982); English translations Theoret. and Math. Phys. 52(2), 777–781, (1982).
31. N.H. Ibragimov, A.H. Kara, F.M. Mahomed, Non. Dyn. 15(2), 115–136 (1998).
32. G. Bluman, Tempurchaalu, S.C. Anco, J. Math. Anal. Appl. 322, 233–250 (2006).
33. S.C. Anco and W. Hereman, in preparation.
34. S.C. Anco, in preparation.

Part V
Materials Science, Engineering, and New Technologies

Adaptive Simulation Selection for the Discovery of the Ground State Line of Binary Alloys with a Limited Computational Budget

Jesper Kristensen, Ilias Bilionis, and Nicholas Zabaras

Abstract First principles calculations are computationally expensive. This information acquisition cost, combined with an exponentially high number of possible material configurations, constitutes an important roadblock towards the ultimate goal of materials by design. To overcome this barrier, one must devise schemes for the automatic and maximally informative selection of simulations. Such information acquisition decisions are task-dependent, in the sense that an optimal information acquisition policy for learning about a specific material property will not necessarily be optimal for learning about another. In this work, we develop an information acquisition policy for learning the ground state line (GSL) of binary alloys. Our approach is based on a Bayesian interpretation of the cluster expanded energy. This probabilistic surrogate of the energy enables us to quantify the epistemic uncertainty induced by the limited number of simulations which, in turn, is the key to defining a function of the configuration space that quantifies the expected improvement to the GSL resulting from a hypothetical simulation. We show that optimal information acquisition policies should balance the maximization of the expected improvement of the GSL and the minimization of the size of the simulated structure. We validate our approach by learning the GSLs of NiAl and TiAl binary alloys, where to establish the ground truth GSL we use the embedded-atom method (EAM) for the calculation of the energy of a given alloy configuration. Note that the proposed policies are directly applicable to the discovery of generic phase diagrams, if one can construct a probabilistic surrogate of the relevant thermodynamic potential.

J. Kristensen (✉)
School of Applied and Engineering Physics, Cornell University, 271 Clark Hall, Ithaca, NY 14853-3501, USA

I. Bilionis
School of Mechanical Engineering, Purdue University, 585 Purdue Mall, West Lafayette, IN 47907-2088, USA
e-mail: ibilion@purdue.edu

N. Zabaras
Department of Aerospace and Mechanical Engineering, University of Notre Dame, 365 Fitzpatrick Hall, Notre Dame, IN 46556, USA
e-mail: nzabaras@gmail.com

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_6

1 Introduction

Technological innovations require high-accuracy analysis of existing materials as well as the discovery of novel ones with extreme/desired properties. Material innovation, however, relies on expensive, time consuming, and risky experiments: (1) There is a significant monetary cost associated with material experiments: hiring and training staff, purchasing supplies, buying and maintaining the instruments, implementing and enforcing safety conditions, etc.; (2) Experiments may take a large amount of time which can be measured in days, weeks, or even months; (3) The discovery process is, in general, intuition-driven and resembles a combinatorial trial-and-error search with no guarantees of success. Note that the -necessarily- finite budget allocated to experimentation amplifies the difficulty of both the analysis and the discovery processes.

The cost associated with physical experiments led to the development of computer codes that could presumably replace the physical experiments at significantly reduced monetary and temporal costs [12]. The added benefits of computer codes include parallelization, thereby accelerating the discovery process, and that they can analyze arbitrary configurations including configurations which are, as of now, impossible to construct experimentally. However, increasing the accuracy of a computer code, e.g., by implementing more of the physics associated with any given application, results in an increase in the computational cost associated with running it. The cost of performing a single simulation may become so high that we are able to afford only few. This is indeed the norm when the simulator is based on *ab initio* calculations.

In binary alloys, the most accurate energy calculation is achieved via the *ab initio* method of *density functional theory* (DFT) [26, 35] for which, e.g., the *Vienna ab initio simulation package* (VASP) [36–39] is popular, and has been available for many years. Other methods can be used as well, such as the *embedded-atom method* (EAM) [14, 15], which approximates the DFT energy landscape. VASP takes as input a binary alloy with atomic positions, and associated identities/atomic types, defined via a unit cell and a set of basis atoms. It then relaxes the ionic and electronic *degrees of freedom* (DOFs) and outputs the quantum mechanical energy of the relaxed structure. Optimization and thermodynamic characterization of binary alloys demand billions of VASP simulations to explore the configuration space near-exhaustively. Since a single VASP simulation needs hours to run, even on a modern supercomputer, directly using the full-fledged quantum mechanical model in such calculations is a futile task.

To circumvent the high cost of the accurate simulator, the common approach is to replace it with an inexpensive surrogate surface based on a finite number of expensive simulations. For alloys on fixed lattices, a popular surrogate is the cluster expansion [66, 67]. The cluster expansion expresses a configuration-dependent property as a linear combination of so-called correlation functions that account for interactions between clusters of atoms. The unknown coefficients of this basis expansion, the *effective cluster interactions* (ECI), are fitted to few expensive observations of the property. The current standard in the field is to fit the ECI using least squares [34, 76] (potentially coupled with genetic algorithms [72]). Other techniques, include linear programming methods [23], compressive sensing [52] and its Bayesian version [53], relative entropy [41], Bayesian linear regression with Gaussian prior on the ECI [51], Bayesian linear regression with Laplace prior on the ECI and Poisson prior on the number of clusters [40], and others [17, 18].

Compared to the many different techniques that have been proposed to fit the ECI, little attention has been paid to the design of algorithms *specifically* for the problem of selecting the design of the computational experiments, i.e., to the data acquisition problem. This is, in the cluster expansion community, known as the *structure selection problem* [76]. In the statistics literature [13, 64], the same problem is known as *design of (computer) experiments* the machine learning community calls it *active learning* [68], while in operations research the term commonly used is *optimal learning* [58]. Intuitively, the data acquisition problem attempts to answer the following question: *How should we design our computational experiments so that we get the maximum amount of information out of them while staying within a finite computational budget?* The answer to this question depends strongly on what we want to learn, i.e., it is task-specific, as well as on the type of budget constraints that are imposed on us. For example, we might be interested in a design which improves the overall predictive capabilities of the surrogate model/surface, study the sensitivity of the response surface with respect to perturbations of the input, or locate extreme properties. All these objectives should, in principle, be addressed with different data acquisition policies. Budget constraints may restrict the total number of simulations that can be performed or the total amount of computational time, which, in turn, may be spent sequentially or in parallel, etc.

A *data acquisition policy* is a decision rule that helps us select the next simulation(s) we should run given our design so far. Note that the stopping rule, i.e., whether it is valuable to continue gathering data or not, is also part of a data acquisition policy. The most popular data acquisition policy is the *uncertainty sampling* policy. Uncertainty sampling simply selects the structure about which the current surrogate is maximally uncertain, i.e., the simulation that exhibits the largest predictive error bar. Intuitively, the objective met by this policy is the overall improvement of the surrogate that represents the underlying response surface. Indeed, it can be shown that, under special assumptions, this policy can be derived from the maximization of the expected information gain about the parameters of the model [45].

In binary alloys modeling, the Alloy Theoretic Automated Toolkit (ATAT) uses a variation of uncertainty sampling that attempts to focus on the discovery of the ground states at each concentration, i.e., it attempts to discover the ground state line (GSL). Specifically, the ATAT policy starts from some initial design defined by the user. At each step, the MAPS algorithm chooses the structure that maximizes a combination of the *least squares variance* and the expected cost of a structure that the authors in Ref. [76] call the *gain*. There is an additional layer of complexity which attempts to address the exploration/exploitation trade-off. At each step in the algorithm, a set of structures are generated that are not part of the training data for the CE. If a subset of these structures are predicted to lower the currently known GSL (we also say that these structures breach the GSL), then the next structure is chosen only among this set. Thus, the algorithm goes from a cost-effective exploration to a cost-effective exploitation mode in the case that we have breaching structures. While we would have preferred to compare our developments with MAPS, the surrogate used in MAPS is not compatible with the Bayesian arguments set forth in this paper. A fair comparison can only be accomplished through extensive modifications to either MAPS or our approach, which would simultaneously defeat the point of the comparison. That is not to say that ideas from MAPS cannot benefit our approach or vice versa, but developing a way to compare the two on equal footing is outside the scope of this work.

The balance between exploration and exploitation is a key concept in the field of global optimization of expensive objective functions. In this field, the *Bayesian global optimization* (BGO) approach has been successful in providing a solution to the data acquisition problem [30, 44, 49, 50, 73]. In BGO, the objective function is replaced by a surrogate based on *Gaussian process regression* (GPR) [62]. One of the marking differences between GPR and classical regression techniques is its Bayesian nature which allows the quantification of the predictive uncertainty of the surrogate [4–6]. It is exactly this *epistemic* uncertainty that can be exploited in various ways to propose adaptive data acquisition policies. These policies are, typically, *myopic*, i.e., they make a decision by considering the result of only one hypothetical future simulation (one-step-look-ahead strategies), and they rely on the maximization of an acquisition function that depends on the surrogate of the objective function. Some popular acquisition functions are the *expected improvement* (EI) [31], the *probability of improvement* [43], and the *knowledge gradient* [21]. From a mathematical perspective, it is straightforward to extend these acquisition functions to construct non-myopic multi-step-look-ahead strategies, albeit this approach requires the solution of a hard dynamic programming problem [3].

In this work, we focus on data acquisition policies for materials discovery and design. In particular, we develop an extension of the EI suitable for gathering information about any *quantity of interest* (QoI) which is defined through the minimization of a functional of the *ab initio* system's thermodynamic potential with respect to a set of parameters [42] specifying the state of the system such as concentration, temperature, pressure, etc. That is, the core idea of our methodology can be used to construct data acquisition policies suitable for the discovery of phase diagrams. Despite the generality of our proposal, we focus on a simple representative example: the discovery of the GSL of a binary alloy. Knowing the GSL with high accuracy forms an essential starting point in constructing the entire phase diagram, e.g., by using thermodynamic integration [22]. We purposefully chose this application because it allows us to validate our results using a highly accurate ground truth, and to systematically compare to traditional, albeit *ad hoc*, data acquisition policies. We demonstrate that the proposed policy can lead to computational savings with simultaneously increased accuracy.

This chapter is organized as follows. We start by developing the theoretical framework, presenting the energy computation scheme used in this work and introducing the surrogate modeling technique. Then, we present the theory underlying our proposed method of selecting structures and summarize the structure acquisition algorithm. The framework is then extended by considering the effect of alloy structure costs. Other structure acquisition algorithms are also considered in order to compare our proposed framework. We present next the results by first describing our validation setup followed by a comparison of our framework with other acquisition strategies. We finally provide some brief conclusions.

2 Methodology

We will study alloy compounds and consider data acquisition strategies which maximize our knowledge about QoIs of such systems. Among possible QoIs are the ground state line, phase diagrams, particular phase transition temperatures, largest-band-gap structures, etc.

The binary NiAl and TiAl alloy compounds will be specifically considered. NiAl was chosen because it plays a central technological role in, e.g., aircraft and rocket engines, power generation turbines, nuclear-power generation, due to its high-temperature strength, toughness, and degradation resistance in oxidizing environments among other useful properties [57]. While other elements can be added to NiAl such as Ti [60], Fe [61, 65], Cr [71], Ta, and Nb [19, 70] to engineer specific properties, we restrict our attention to pure NiAl, since it forms one of the most important binary bases for superalloys [27]. The elements Ni and Al crystallize in fcc lattices. NiAl has been observed to form in both fcc and bcc lattices. Secondly, TiAl was chosen because of its application in gas turbines due to its high strength-to-weight ratio and excellent corrosion resistance. It is used in aerospace applications, specifically in landing gear beams on the Boeing 747 and 757 replacing steel which has too high a density [8]. We consider TiAl on an fcc lattice, but note that it can crystallize in hcp and bcc lattices as well.

Let the atomic identities occupying the lattice of the A_xB_{1-x} alloy with N_s sites be summarized in an N_s-length boolean vector called a *configuration* and denote it by $\boldsymbol{\sigma}$. Accordingly, call the set of all possible configurations the *configuration space*. Now denote a set of thermodynamic parameters which specify the state of a system, e.g., temperature, pressure, concentration, etc., by ω. We are interested in the characterization of QoIs of the form:

$$\boldsymbol{\sigma}^*(\omega) = \arg\min_{\boldsymbol{\sigma}} G(\boldsymbol{\sigma}, \omega), \tag{1}$$

where the minimization takes place over the $\boldsymbol{\sigma}$'s that are compatible with ω. The function $G(\boldsymbol{\sigma}, \omega)$ is just the natural thermodynamic potential of the system whose minimization gives us the thermodynamically stable structure of the system at ω [42].

As two examples, consider first a system at zero temperature and with the concentration as the only thermodynamic parameter, i.e., $\omega = \{x\}$. For a binary A_xB_{1-x} alloy with N_s lattice sites, we map A to -1 and B to $+1$. Thus, $2x = 1 - \sum_{n=1}^{N_s} \sigma_n / N_s$, where σ_n is $-1(+1)$ if A(B) occupies site n. The thermodynamic potential forming the energy-composition ground state line is $\Delta E_{\mathrm{form}}(\boldsymbol{\sigma}) \equiv E(\boldsymbol{\sigma}) - (xE_{\mathrm{A}} + (1-x)E_{\mathrm{B}})$, where E_{C} is the internal energy of the structure containing only C-atoms. In other words, $G(\boldsymbol{\sigma}, x) = \Delta E_{\mathrm{form}}(\boldsymbol{\sigma})$ ($\boldsymbol{\sigma}$ implies x but we leave the notation general) and Eq. (1) forms the binary alloy GSL. As a second example, assume the thermodynamic parameters are the pressure P and the temperature T, i.e., $\omega = \{P, T\}$. Then, the natural thermodynamic potential $G(\boldsymbol{\sigma}, P, T)$ is the Gibbs free energy of a closed system and Eq. (1) provides the stable structures versus ω, i.e., the temperature-pressure phase diagram. More generally, Eq. (1) constructs all the possible phase diagrams from various subsets of ω.

In this work, we restrict the DOFs of the alloys by considering structures having fixed lattices, but our methodology extends to multi-lattice settings. Furthermore, only alloy DOFs which characterize the atomic type/identity on each lattice site are considered. We will be interested in QoIs at zero temperature the implication being that $\omega = \{x\}$ is the lone thermodynamic parameter.

2.1 Computing Alloy Energies with an EAM Relaxation Scheme

In order to determine the energy landscape and the GSL of the system, we need to calculate the energy of an arbitrary (Ni,Ti)Al configuration. There are various ways of approximating this energy. Here, we choose not to use a high-accuracy *ab initio* approach, since this would limit our ability to validate our predictions. For example, establishing the ground truth GSL by evaluating the ground state energy using density functional theory (DFT) would require tremendous computational resources. Therefore, we use the embedded atom method (EAM) which provides an energy landscape approximating that of DFT, but which is navigable with significantly reduced computational resources. We use the parameters found in Ref. [59]. These parameters are highly transferable in the NiAl system and they reproduce fairly well the GSL of the system as opposed to reproducing just single isolated phases [48]. As an example of the increased computational benefits in using EAM over, say, DFT, the temporal cost of a single high-accuracy structure with DFT is approximately, on our machines, equivalent to computing $5,000$ structures with EAM, even after accounting for the extra efforts from using the EAM relaxation scheme to be discussed next.

Each configuration was relaxed under the EAM potential using the following iterative procedure [11]. The fcc unit cell size was initially guessed to have lattice constant 4.01 Å for both NiAl and TiAl. Since the preferred size of the unit cell and the interatomic distances vary depending on the atomic environment, we implemented a 2-step loop to find the energetically most favorable atomic arrangement. In step 1 (unit cell relaxation), the unit cell was isotropically scaled until reaching equilibrium in the EAM energy landscape using a bounded (with bounds 50 % to 200 % of the lattice constant) Brent–Dekker method [9, 16]. In step 2 (atomic positions relaxation), a Broyden–Fletcher–Goldfarb–Shanno quasi-Newton algorithm [10, 20, 24, 69] relaxed the interatomic force vectors to zero. These two steps were repeated until the energy and the magnitude of the largest force vector did not change in two consecutive steps by more than 10^{-8}. Less than eight iterations of the loop were typically enough to converge. To verify the above implementation, we applied it to fcc pure Al(Ni) and obtained −3.36 eV/atom(−4.45 eV/atom) with relaxed-structure-lattice-constants of 4.0500 Å(3.5199 Å) all in excellent agreement with Ref. [59]. This framework was readily implemented relying on the ASE [1] PYTHON [47, 54] package developed at the Technical University of Denmark.

2.2 *Cluster Expansion Surrogate Model*

In this section, we briefly discuss the details of the cluster expansion (CE). We refer to Refs. [63, 66], and [67] for an introduction to this topic. The CE expands the configuration-dependent (Ni,Ti)Al energy $E(\cdot)$ as:

$$\begin{aligned} E(\boldsymbol{\sigma}) \approx E(\boldsymbol{\sigma};\boldsymbol{\gamma}) &= \textstyle\sum_{i=1}^{M} \gamma_i \phi_i(\boldsymbol{\sigma}) \\ &= \boldsymbol{\gamma}^T \boldsymbol{\phi}(\boldsymbol{\sigma}), \end{aligned} \tag{2}$$

where $\boldsymbol{\gamma} = \{\gamma_i\}$ are the unknown expansion coefficients called the effective cluster interactions (ECI), $\boldsymbol{\phi}(\boldsymbol{\sigma}) = \{\phi_i(\boldsymbol{\sigma})\}$, and the ith basis function is given by:

$$\phi_i(\boldsymbol{\sigma}) = \langle \Gamma_{\boldsymbol{\alpha}'}(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\alpha}' \sim \boldsymbol{\alpha}_i}, \tag{3}$$

where $\boldsymbol{\alpha}_i$ is a vector, the ith subset among all subsets of lattice sites we have chosen to consider in the sum (e.g., $\boldsymbol{\alpha}_2$ could be the first two lattice sites; under some arbitrary numbering of the sites), also called a *cluster* and $\langle\cdot\rangle_{\boldsymbol{y}\sim\boldsymbol{x}}$ is taken to mean an average over all clusters $\boldsymbol{y}$ that are symmetrically equivalent to $\boldsymbol{x}$ under a space group operation of the empty lattice (lattice points without atomic identities). Thus, the sum in Eq. (2) is over all symmetrically inequivalent clusters. The $\Gamma_{\cdot}(\cdot)$'s are known as *correlation functions*. They are defined to be monomials of the spin variables

$$\Gamma_{\boldsymbol{\alpha}}(\boldsymbol{\sigma}) = \prod_{i\in\boldsymbol{\alpha}} \sigma_i, \tag{4}$$

where the product is over all sites in cluster $\boldsymbol{\alpha}$ and σ_i is the atomic identity on site i.

Notationally, a single alloy configuration $\boldsymbol{\sigma}$ is associated with a set of M basis functions collected in the vector $\boldsymbol{\phi}(\boldsymbol{\sigma})$. When we consider, say, N configurations, $\{\boldsymbol{\sigma}^{(j)}\}_{j=1}^{N}$, we collect the basis vectors associated with each configuration in an $N \times M$ design matrix $\boldsymbol{\Phi}$ where the jth row is $\boldsymbol{\phi}(\boldsymbol{\sigma}^{(j)})$. The kth column in the jth row is $\phi_k(\boldsymbol{\sigma}^{(j)})$ and given by Eq. (3).

If a cluster contains n sites it is called an n-point cluster. Our CE of fcc (Ni,Ti)Al included $M = 49$ symmetrically inequivalent clusters with maximum spatial extents of 10, 7, and 4 Å for the 2, 3, and 4-point clusters, respectively. Interestingly, if one lets $M \to \infty$ the CE becomes *exact*, however, a truncation of the clusters to sum over is needed and carried out by fixing the maximum number of points present in any cluster as well as its maximum spatial extent (measured, in this work, as the largest distance between any two lattice sites in the cluster) [76].

At this point, note that $G(\boldsymbol{\sigma}, \omega)$, Eq. (1) depends on the configurational energy surface $E(\cdot; \boldsymbol{\gamma})$, and thus, as a consequence, on the ECI, $\boldsymbol{\gamma}$. That is, we can write the following:

$$G(\boldsymbol{\sigma}, \omega) \equiv G(\boldsymbol{\sigma}, \omega, E(\cdot; \boldsymbol{\gamma})) \equiv G(\boldsymbol{\sigma}, \omega, \boldsymbol{\gamma}). \tag{5}$$

Similarly, the stable structure of Eq. (1) depends on $\boldsymbol{\gamma}$: $\boldsymbol{\sigma}^*(\omega) \equiv \boldsymbol{\sigma}^*(\omega, E(\cdot; \boldsymbol{\gamma})) \equiv \boldsymbol{\sigma}^*(\omega, \boldsymbol{\gamma}) = \operatorname{argmin}_{\sigma} G(\boldsymbol{\sigma}, \omega, \boldsymbol{\gamma})$. If the thermodynamic potential $G(\boldsymbol{\sigma}, \omega)$ is expensive to evaluate, e.g., if it requires thermodynamic integration, a solution is to cluster expand it but with ω-dependent ECI as discussed in Ref. [75].

2.3 Learning the Effective Cluster Interactions Using Bayesian Linear Regression

To learn the ECI in Eq. (2) and to enable a quantification of the uncertainty in energies computed from these parameters, we adopt a Bayesian approach [29]. We start with our prior belief about the ECI, represented here as a continuous probability distribution [7]. We believe more in smaller valued ECI [52]. In other words, we favor smoother energy surfaces, so the distribution should put more of its mass closer to zero. A distribution satisfying this, and which simultaneously simplifies the mathematics ahead, is a zero-mean isotropic Gaussian:

$$\begin{aligned} p(\boldsymbol{\gamma}|\alpha) &= \mathcal{N}(\boldsymbol{\gamma}|\mathbf{0}, \alpha^{-1}\mathbf{1}) \\ &= \left(\tfrac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\tfrac{\alpha}{2}\boldsymbol{\gamma}^T\boldsymbol{\gamma}\right), \end{aligned} \tag{6}$$

where M is the total number of ECI, and α is known as the *precision* hyper-parameter. The precision hyper-parameter, is the inverse variance associated with the prior probability we assign to the ECI. That is, the greater the precision hyper-parameter, the more certain we are *a priori* that the ECI are closer to zero.

The next ingredient that we need to specify is the likelihood of the data. The likelihood of a configuration-energy couple, denoted $(\boldsymbol{\sigma}, E)$, is defined conditional on the ECI, $\boldsymbol{\gamma}$. In common data analysis, the likelihood models the measurement process. Here, since our measurement process is essentially deterministic, it quantifies the model discrepancy. That is, it quantifies the discrepancy between the cluster expansion and the actual energy. In lack of a better alternative, we assume that this discrepancy is distributed normally with a noise precision β. Mathematically:

$$p(E|\boldsymbol{\sigma},\boldsymbol{\gamma},\beta) = \mathcal{N}\left(E|\boldsymbol{\gamma}^T\boldsymbol{\phi}(\boldsymbol{\sigma}),\beta^{-1}\right).$$

Assuming independence of each observation, the likelihood of a set of N observed configuration-energy couples,

$$\mathcal{D}_N = \left\{\left(\boldsymbol{\sigma}^{(i)},E^{(i)}\right)\right\}_{i=1}^{N}, \tag{7}$$

is given by

$$p(\mathcal{D}_N|\boldsymbol{\gamma},\beta) = \prod_{i=1}^{N} p\left(E^{(i)}|\boldsymbol{\sigma}^{(i)},\boldsymbol{\gamma},\beta\right). \tag{8}$$

Combining our prior belief, Eq. (6), with our observations, Eq. (8), using Bayes' rule [2] results in the *posterior* probability density:

$$p(\boldsymbol{\gamma}|\mathcal{D}_N,\alpha,\beta) \propto p(\mathcal{D}_N|\boldsymbol{\gamma},\beta)p(\boldsymbol{\gamma}|\alpha), \tag{9}$$

which corresponds to our updated beliefs about the ECI. For these specific prior and likelihood choices, it can be shown that the posterior is Gaussian,

$$p(\boldsymbol{\gamma}|\mathcal{D}_N,\alpha,\beta) = \mathcal{N}\left(\boldsymbol{\gamma}|\mathbf{m}_N,\mathbf{S}_N\right), \tag{10}$$

where the mean vector and covariance matrix are given by

$$\boldsymbol{m}_N = \beta\mathbf{S}_N\boldsymbol{\Phi}^T\mathbf{E}, \tag{11}$$

and

$$\mathbf{S}_N = \left(\alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}, \tag{12}$$

respectively [7], with $\mathbf{I}$ being the unit matrix, and $\mathbf{E} = \left(E^{(1)},\dots,E^{(N)}\right)$.

So far, we have tacitly assumed that the *hyper-parameters* α and β are given. In general, however, they are unknown *a priori*, and we should have assigned a prior, $p(\alpha,\beta)$, to them. Having done that, we would have had to characterize the joint posterior probability density:

$$p(\alpha,\beta,\boldsymbol{\gamma}|\mathcal{D}_N) \propto p(\mathcal{D}_N|\boldsymbol{\gamma},\beta)p(\boldsymbol{\gamma}|\alpha)p(\alpha,\beta),$$

but unfortunately, the resulting posterior would not be analytically available. Therefore, we resort to the *evidence approximation* [7]. To motivate this approximation, notice that, by repeated applications of the Bayes' rule, we may write:

$$p(\alpha, \beta, \boldsymbol{\gamma}|\mathcal{D}_N) = p(\boldsymbol{\gamma}|\mathcal{D}_N, \alpha, \beta)p(\alpha, \beta|\mathcal{D}_N), \tag{13}$$

with

$$p(\alpha, \beta|\mathcal{D}_N) \propto p(\mathcal{D}_N|\alpha, \beta)p(\alpha, \beta). \tag{14}$$

Here $p(\mathcal{D}_N|\alpha, \beta)$ is known as the *marginal likelihood* and, using the sum rule of probability,

$$p(\mathcal{D}_N|\alpha, \beta) = \int p(\mathcal{D}_N|\boldsymbol{\gamma}, \beta)p(\boldsymbol{\gamma}|\alpha)\mathrm{d}\boldsymbol{\gamma}. \tag{15}$$

Intuitively, the evidence approximation assumes that the prior $p(\alpha, \beta)$ is relatively flat, and that the marginal likelihood $p(\mathcal{D}_N|\alpha, \beta)$ has a well separated global maximum. This justifies an approximation of $p(\alpha, \beta|\mathcal{D}_N)$ of the form

$$p(\alpha, \beta|\mathcal{D}_N) \approx \delta(\alpha - \hat{\alpha})\delta(\beta - \hat{\beta}), \tag{16}$$

where $\delta(\cdot)$ is the Dirac δ-function, and the $\hat{\alpha}$ and $\hat{\beta}$ are set by maximizing the marginal likelihood:

$$(\hat{\alpha}, \hat{\beta}) = \arg\max_{(\alpha,\beta)} p(\mathcal{D}_N|\alpha, \beta). \tag{17}$$

See Ref. [7] for an expectation-maximization algorithm that converges to a local maximum of the marginal likelihood. Repeated restarts of this algorithm, provide a good approximation to the solution of Eq. (17).

Having characterized the posterior, Eq. (13), via the evidence approximation, we can now make predictions about the energy, $\tilde{E}$, we may observe at an arbitrary configuration $\tilde{\boldsymbol{\sigma}}$. The *predictive* probability density is:

$$\begin{aligned} &p(\tilde{E}|\tilde{\boldsymbol{\sigma}}, \mathcal{D}_N) \\ &\quad = \int p(\tilde{E}|\tilde{\boldsymbol{\sigma}}, \boldsymbol{\gamma}, \beta)p(\alpha, \beta, \boldsymbol{\gamma}|\mathcal{D}_N)\mathrm{d}\alpha\mathrm{d}\beta\mathrm{d}\boldsymbol{\gamma} \\ &\quad \approx \int p(\tilde{E}|\tilde{\boldsymbol{\sigma}}, \boldsymbol{\gamma}, \hat{\beta})p(\boldsymbol{\gamma}|\mathcal{D}_N, \hat{\alpha}, \hat{\beta})\mathrm{d}\boldsymbol{\gamma}, \end{aligned}$$

where to derive the last equation we used Eq. (13) and Eq. (16). Since the two probability densities inside the last integral are Gaussian, see Eqs. (8) and (10), it is possible to evaluate it analytically. The result is:

$$p(\tilde{E}|\tilde{\boldsymbol{\sigma}}, \mathcal{D}_N) \approx \mathcal{N}\left(\tilde{E}|\mu_{E,N}(\tilde{\boldsymbol{\sigma}}), \nu^2_{E,N}(\tilde{\boldsymbol{\sigma}})\right), \tag{18}$$

where the *predictive mean* is

$$\mu_{E,N}(\tilde{\boldsymbol{\sigma}}) = \mathbf{m}_N^T \boldsymbol{\phi}(\tilde{\sigma}), \tag{19}$$

and the *predictive variance* is

$$\nu_{E,N}^2(\tilde{\boldsymbol{\sigma}}) = \frac{1}{\hat{\beta}} + \boldsymbol{\phi}(\tilde{\boldsymbol{\sigma}})^T \mathbf{S}_N \boldsymbol{\phi}(\tilde{\boldsymbol{\sigma}}). \tag{20}$$

The predictive mean can be thought of as a mean surrogate energy surface. The predictive variance quantifies our uncertainty about the predictions of this surrogate for any given alloy structure.

2.4 Optimal Selection of Input Structures

2.4.1 The Expected Improvement Policy

In this subsection, we develop an informed data acquisition policy that enables us to select simulations that are maximally informative about the thermodynamically stable structures versus ω, specified in Eq. (1). We assume that we have made a total of N observations, $\mathcal{D}_N$, as in Eq. (7), and that we have at hand a Gaussian approximation to the predictive distribution of the thermodynamic potential $G(\boldsymbol{\sigma}, \omega)$:

$$p(\tilde{G}|\tilde{\boldsymbol{\sigma}}, \tilde{\omega}, \mathcal{D}_N) \approx \mathcal{N}\left(\tilde{G}|\mu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}), \nu_{G,N}^2(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})\right), \tag{21}$$

where $\mu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})$ and $\nu_{G,N}^2(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})$ are the predictive mean and variance, respectively. What follows is independent of the way this predictive distribution was obtained. Remember that for the GSL, $\omega = x$ and $G(\boldsymbol{\sigma}, x)$ is just the formation energy. Thus, in this case, Eq. (21) can be obtained trivially from Eq. (18). For a general thermodynamic potential Eq. (21) has to be obtained by directly cluster expanding $G(\boldsymbol{\sigma}, \omega)$ with ω-dependent ECI [75].

The current observed minimum thermodynamic potential at a given ω, $G_N(\omega)$, is

$$G_N(\omega) = \min_{1 \leq n \leq N} G(\boldsymbol{\sigma}^{(n)}, \omega), \tag{22}$$

where the minimum is taken only over ω-compatible $\boldsymbol{\sigma}^{(n)}$'s. An informative simulation at ω would, ideally, yield a lower thermodynamic potential than the currently observed minimum. To formalize this intuition, let us fix the thermodynamic parameters to $\tilde{\omega}$ and assume that we make a hypothetical simulation at an $\tilde{\omega}$-compatible structure $\tilde{\boldsymbol{\sigma}}$. If this simulation resulted in a measured thermodynamic potential equal to $\tilde{G}$, then this would yield an *improvement* of $\mathrm{I}(\cdot)$ equal to

$$\mathrm{I}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}, \tilde{G}) = \max\left\{0, G_N(\tilde{\omega}) - \tilde{G}\right\}. \tag{23}$$

Notice here, that we consider the new simulation an "improvement" only if it reduces the currently observed thermodynamic potential $G_N(\tilde{\omega})$, that is, if it finds structures that are thermodynamically more stable.

$\tilde{G}$ is a hypothetical measurement. Therefore, in order to eliminate the dependence of the improvement on $\tilde{G}$, we take its expectation over the predictive distribution of $\tilde{G}$ conditional on $\tilde{\sigma}$, $p(\tilde{G}|\tilde{\boldsymbol{\sigma}}, \tilde{\omega}, \mathcal{D}_N)$. In this way, we define the *expected improvement* (EI):

$$\mathrm{EI}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}) = \mathbb{E}\left[\mathrm{I}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}, \tilde{G})|\tilde{\boldsymbol{\sigma}}, \tilde{\omega}, \mathcal{D}_N\right]. \tag{24}$$

The EI measures the expected change in the minimum observed thermodynamic potential value at $\tilde{\omega}$ after making an $\tilde{\omega}$-compatible simulation at $\tilde{\boldsymbol{\sigma}}$. Using Eq. (21), we have:

$$\begin{aligned}\mathrm{EI}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}) &= \mathbb{E}\left[\mathrm{I}\left(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}, \tilde{G}\right)|\tilde{\boldsymbol{\sigma}}, \tilde{\omega}, \mathcal{D}_N\right] \\ &= \int \max\left\{0, G_N(\tilde{\omega}) - \tilde{G}\right\} p(\tilde{G}|\tilde{\boldsymbol{\sigma}}, \tilde{\omega}, \mathcal{D}_N)\mathrm{d}\tilde{G} \\ &\approx \int \max\left\{0, G_N(\tilde{\omega}) - \tilde{G}\right\} \times \mathcal{N}\left(\tilde{G}|\mu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}), \nu^2_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})\right) \mathrm{d}\tilde{G} \\ &= \int_{-\infty}^{G_N(\tilde{\omega})} \left(G_N(\tilde{\omega}) - \tilde{G}\right) \times \mathcal{N}\left(\tilde{G}|\mu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}), \nu^2_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})\right) \mathrm{d}\tilde{G}.\end{aligned}$$

Employing standard normal integral identities, we obtain the following

$$\begin{aligned}\mathrm{EI}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}) = [G_N(\tilde{\omega}) - \mu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})]\, \Psi\left(\frac{G_N(\tilde{\omega}) - \mu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})}{\nu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})}\right) \\ +\nu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})\psi\left(\frac{G_N(\tilde{\omega}) - \mu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})}{\nu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})}\right),\end{aligned} \tag{25}$$

and $\mathrm{EI}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}) = 0$ if $\nu_{G,N}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}) = 0$, where $\Psi(\cdot)$ and $\psi(\cdot)$ are the cumulative distribution function and the probability density function of a standard normal random variable, respectively, and we note that the EI has the same units as the thermodynamic potential.

The EI acquisition policy adds to a current data pool of N structures the configuration that yields the maximum overall EI, i.e., the maximum EI over both the input space *and* the thermodynamic variables:

$$\left(\boldsymbol{\sigma}^{(N+1)}, \omega^{(N+1)}\right) = \underset{\tilde{\boldsymbol{\sigma}}, \tilde{\omega}}{\operatorname{argmax}}\, \mathrm{EI}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}). \tag{26}$$

Intuitively, this strategy chooses the simulation that yields the maximum change in our state of knowledge about the thermodynamically stable structures across all values of ω.

Solving the global maximization problem of Eq. (26) exactly is not trivial. Since we have an analytic approximation to the EI, Eq. (25), it is feasible to obtain approximate solutions to Eq. (26) through a random sampling strategy. Specifically, we consider a large pool of candidate simulations

Algorithm 1: EI structure acquisition strategy for learning the thermodynamic potential

```
Require: $\mathcal{D}_{N_0}$ (an initial pool of $N_0$ observed $\boldsymbol{\sigma}$-$\omega$-$G$ triples), $N_{\max}$ (maximum number of
         observations that can be afforded), $\epsilon$ (EI tolerance), $\mathcal{S}_{N_{\text{pool}}}$ (large pool of $\boldsymbol{\sigma}$-$\omega$ pairs
         to select simulations from).
1  $N \leftarrow N_0$
2  $\mathcal{D}_N \leftarrow \mathcal{D}_{N_0}$
3  repeat
4      Find: $\left(\boldsymbol{\sigma}^{(N+1)}, \omega^{(N+1)}\right) = \underset{(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}) \in \mathcal{S}_{N_{\text{pool}}}}{\operatorname{argmax}} \operatorname{EI}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})$.
5      if $\operatorname{EI}\left(\boldsymbol{\sigma}^{(N+1)}, \omega^{(N+1)}\right) < \epsilon$ then
6          Break loop
7      end
8      $G^{(N+1)} \leftarrow G\left(\boldsymbol{\sigma}^{(N+1)}, \omega^{(N+1)}\right)$
9      $\mathcal{D}_{N+1} \leftarrow \mathcal{D}_N \cup \left\{\left(\boldsymbol{\sigma}^{(N+1)}, \omega^{(N+1)}, G^{(N+1)}\right)\right\}$
10     $N \leftarrow N + 1$
11 until $N >= N_{max}$;
```

$$\mathcal{S}_{N_{\text{pool}}} = \left\{\left(\tilde{\boldsymbol{\sigma}}^{(n)}, \tilde{\omega}^{(n)}\right)\right\}_{n=1}^{N_{\text{pool}}},$$

and approximate Eq. (26) by:

$$\left(\boldsymbol{\sigma}^{(N+1)}, \omega^{(N+1)}\right) = \underset{(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}) \in \mathcal{S}_{N_{\text{pool}}}}{\operatorname{argmax}} \operatorname{EI}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}). \tag{27}$$

Importantly, assume that the candidate pool is generally attainable with minimal computational efforts, which is most typically the case.

Algorithm 1 outlines the EI sequential data acquisition policy for the thermodynamic potential. The policy sequentially selects the simulations that maximize the EI. The iterations stop when either the maximum EI falls below a specific threshold $\epsilon > 0$ or the simulation budget has been exhausted. At any given iteration of the algorithm, the best estimate of the thermodynamic potential, at some ω, is given by the current observed minimum-thermodynamic-potential structure at ω. In step 8 of Algorithm 1, the expensive computer code is run on the newly selected configuration $\boldsymbol{\sigma}^{(N+1)}$.

2.4.2 Dealing with Structures of Varying Cost

Let the cost of evaluating $G(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})$ be $C(\tilde{\boldsymbol{\sigma}}, \tilde{\omega})$, here $C(\tilde{\boldsymbol{\sigma}}, \omega)$ is the number of atoms in the configurational unit cell cubed. Obviously, if we had to choose among two structures with the same cost, we would pick the one with the maximum EI. Similarly, if we had to choose among two structures with the same EI, we would pick the one with the minimum cost. Therefore, the information acquisition problem must balance between the two, potentially, competing objectives of maximizing EI and minimizing cost. This multi-objective decision problem induces a Pareto front of optimal choices. An optimal information acquisition policy should only select for simulation optimal structures. To this end, we introduce a modified EI policy, selecting the structure that maximizes:

$$\mathrm{EI}_{\lambda}(\tilde{\boldsymbol{\sigma}}, \omega) \equiv \lambda \mathrm{EI}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}) - (1-\lambda)\mathrm{C}(\tilde{\boldsymbol{\sigma}}, \tilde{\omega}), \tag{28}$$

for some $\lambda \in [0, 1]$. Let this strategy be denoted $\mathcal{A}_{\mathrm{EI}_{\lambda}}$. When $\lambda = 0$ the cost is minimized and we always choose among the least expensive structures. When $\lambda = 1$, we follow the EI acquisition strategy with no regards to the cost of the structures. For other values of λ, we still attempt to choose structures with large values of EI, but at the same time, we are trying to minimize the cost. Each value of λ corresponds to a Pareto-frontier point. As a side note, numerically, it is important to compute Eq. (28) with scaled versions of the EI and the cost to make them comparable in size.

In the numerical examples, we show how $\mathcal{A}_{\mathrm{EI}_{\lambda}}$ and the Pareto frontier behave for the various acquisition policies introduced in the following section.

2.5 *Other Input Structure Acquisition Strategies*

We compare $\mathcal{A}_{\mathrm{EI}_{\lambda}}$ against three other methods to see how well it fares. We briefly discuss these trivial policies here. First, consider the acquisition strategy that randomly selects the next structure, denote this strategy $\mathcal{A}_{\mathrm{rnd}}$. All structures in the pool are picked with equal probability. Second, introduce a strategy which always chooses the smallest structure next. We order structure sizes by the number of atoms per unit cell first and then by the unit cell volume. If both these quantities tie, a random choice is made. Denote this strategy $\mathcal{A}_{\mathrm{sml}}$. Consider now a strategy which selects the next structure that has the largest predictive variance Eq. (20). We refer to this strategy as *uncertainty sampling* and denote it $\mathcal{A}_{\mathrm{us}}$.

In terms of the discussion in the previous subsection, these policies will not, generally speaking, lie on the Pareto frontier. We show this next.

3 Results

We developed a software package for performing the ground state search of binary alloys. The software is written in the PYTHON programming language using an amalgamation of PYMATGEN [55], ATAT [77], ENUMLIB [25], ASE [1], NUMPY [74], SCIPY [32], MATPLOTLIB [28], SCIKIT-LEARN [56], and PANDAS to handle the data in a concise database format [46].

3.1 Validation

To validate the proposed framework, a scenario is created in which ground truth is known. This is done by first computing the EAM energies of the first $34,368$ (Ni,Ti)Al symmetrically inequivalent configurations from smaller to larger multiples of the basic fcc unit cell, and then using them to approximate the GSL. Since $34,368$ configurations is a lot more than the sizes typically considered for such a task, we assume, for all intends and purposes, that $G_{34368}(x) \approx G_\infty(x)$, and to be concise we will be referring to it as the "true GSL" rather than as the "true GSL of the first $34,368$ symmetrically inequivalent configurations". Explicitly knowing the true GSL allows us to quantify the error of any data acquisition policy and is thus essential from a comparison perspective. Specifically, assume that we have made N observations, $\mathcal{D}_N$. We define the relative GSL error (GSLE) between the true GSL, $G_\infty(x)$, and a GSL formed from $\mathcal{D}_N$, $G_N(x)$, by:

$$\text{GSLE}(\mathcal{D}_N) = \frac{||G_\infty - G_N||_2}{||G_\infty||_2}, \tag{29}$$

where $||f||_2$ is the $\mathcal{L}_2$ norm of the function $f(x)$, i.e.,

$$||f||_2 = \sqrt{\int_{x=0}^{1} f^2(x)\mathrm{d}x}.$$

Along with Eq. (29), as a performance metric, we also keep track of *which* true ground states have been correctly predicted. Other performance metrics can be considered as well, see, e.g., Ref. [33] and the references therein.

3.2 Learning the Ground State Line Using the Expected Improvement Data Acquisition Policy

We now present the global EI acquisition process, Algorithm 1, when applied to the task of learning the TiAl fcc GSL. Figure 1 shows the evolution of the TiAl GSL (red dashed line) as observations are added. In the top right of each subplot a number identifies the iteration in our global EI algorithm. "iteration 0" in the top left subplot is the GSL of the initial data pool. In each subplot we find two plots, the top plot shows the true GSL (blue full line) with associated true ground states (blue upside-down triangles). The same plot, but for positive ordinate values, also shows as a black dashed line quantifying the error between the GSL of the initial data pool and the true GSL. This error is *not* Eq. (29), but simply the vertical distance between the GSLs versus concentration. For convenience, the GSLE, Eq. (29), is reported under the dashed line.

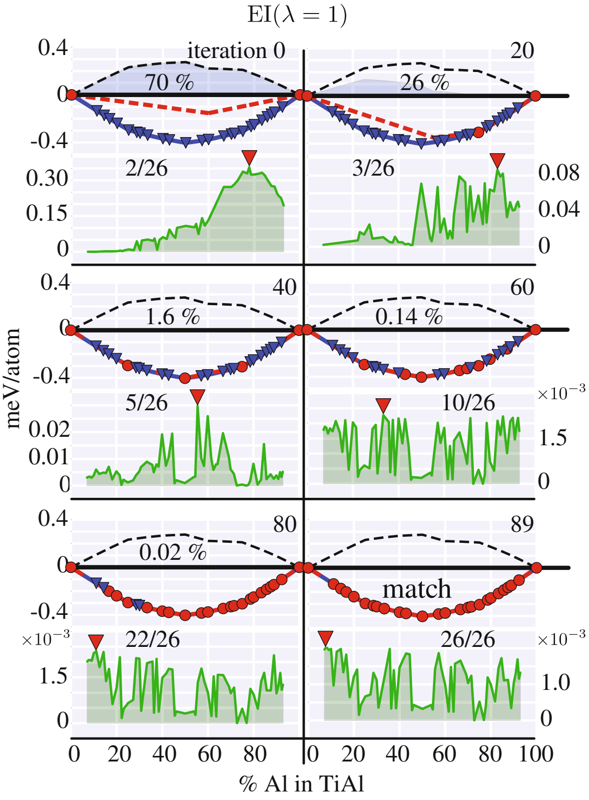


Fig. 1 (Color) For TiAl. Six different stages, shown as six separate subplots arranged in two columns and three rows, during the Bayesian global optimization algorithm for learning the true ground state line (GSL) with an initial data pool of six structures. The iteration number of the algorithm is shown in the top right of each subplot. E.g., the top left subplot shows the algorithm's behavior on the initial data set ("iteration 0"). Each subplot has two parts. The upper part shows, via a shaded blue area (for positive ordinate values), the error measured as the vertical distance between the GSL of the current seen structures (red dashed line at negative ordinate values) and the true GSL (blue full line at negative ordinate values) versus Al concentration. The dashed black line in all subplots is this error between the initial GSL (iteration 0) and the true GSL. The GSLE (Eq. (29)) is given as the text under this line. The true ground states are shown as blue upside-down triangles and the ground states which are correctly predicted by the global EI algorithm are shown as red circles. The number of correctly predicted ground states (out of the total possible of 26) is given in text under the true GSL. The lower plot in each subplot quantifies the EI versus Al concentration as a dark green full line. It is the largest point of this line, marked with an upside down red triangle, which is the global max EI, and hence where the next structure is selected for addition to the design. Note that the lower part of each subplot in the right (left) column is measured on the right (left) side of the figure. All ordinate values are in meV's/atom

The dashed line is present in all subsequent subplots for quick comparison of the current error (blue shaded area) to the initial error at iteration 0. In the final subplot (iteration 89), a "match" represents, instead of the GSLE, that all true ground states have been successfully learned. The red circles in Fig. 1 on top of the true GSL show ground states found by the thermodynamic EI algorithm which are also true ground states. The number of ground states found by the EI out of the total number of true ground states, the latter which in our experiment is 26, out of the $34,368$, is reported under the true GSL in each subplot. Remarkably, the thermodynamic EI finds all true ground states among the $34,368$ structures with just 89 structures. In each subplot, the second, lower, plot shows the EI versus Al concentration. A red upside-down triangle marks the global maximum of the EI, and hence which concentration the chosen structure has. Notice how, initially, the error is reduced for large Al concentrations. Once that part of the GSL has been learned, the algorithm automatically shifts its attention to lower Al concentrations and finally, the EI is comparable for all concentrations. Furthermore, we see that the global EI decreases in magnitude versus iteration. This is expected since we should expect a smaller and smaller difference to be made to the ground state line as we get closer to ground truth. We would like to emphasize the extremely small starting data pool of six structures using a simple Bayesian linear regression to capture the relaxed EAM energies. Furthermore, we do not use any basis-optimization, such as using cross-validation to select the best set of basis functions to use. In fact, we are using a fixed set of 39 clusters (13 2-point, 23 3-point, and one 4-point including the empty-point and 1-point clusters). We expect these results to be even better if coupling thermodynamic EI with a more advanced surrogate model such as GPR. To quantify the observed decrease in EI, we turn to Fig. 2 which shows the evolution of the thermodynamic EI, normalized to its initial value, along with a hypothetical 1 % threshold dashed line, which could act as a stopping rule in some cases.

3.3 Comparing Data Acquisition Policies for Learning the Ground State Line

The following results are all based on the Bayesian linear regression on the CE as a surrogate and can change if using different surrogates. In Fig. 3 we compare the GSLE in Eq. (29) of different structure acquisition strategies, including Eq. (28) for various values of λ, with the objective of learning the fcc NiAl and TiAl GSLs when starting from a (small) initial data pool of six structures spread across the concentration range as evenly as possible.

To better represent the total temporal cost of the various methods we do not plot against the total number of structures added to the design, but rather, we plot against the total expected temporal cost, which, for DFT, is proportional to the number of atoms in the configurational unit cell cubed. Although we are not using DFT, but

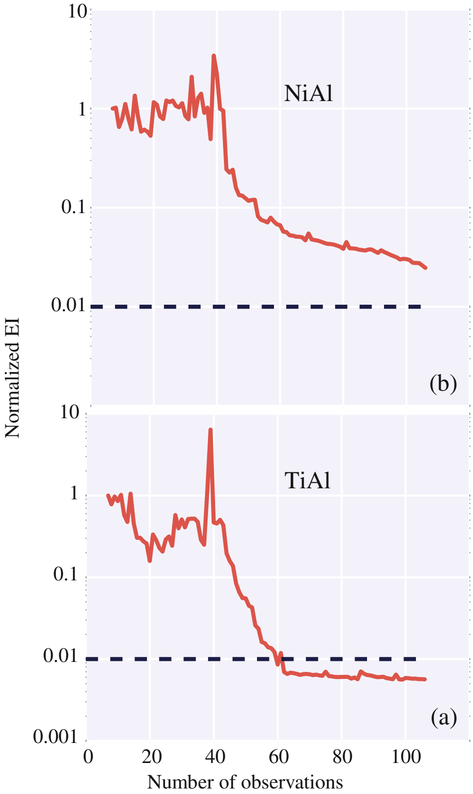


Fig. 2 (Color) The expected improvement (EI) in Eq. (26) for (**a**) NiAl and (**b**) TiAl, normalized to its value in the first iteration when starting from the initial data pool of six structures, plotted on a log scale. The abscissa shows the total number of observations in the design of the experiment and the black dashed line reports where the thermodynamic EI has dropped to 1 % of its initial value

rather an EAM relaxation scheme, we assume a DFT cost of the structures. So, e.g., if the first structure added has four atoms per configurational unit cell, the structure adds 64 to the current position on the abscissa.

Analyzing the graphs, we first note that $\mathcal{A}_{\mathrm{rnd}}$ is not a good strategy. We can understand this by consulting Fig. 4 where, in Fig. 4a, the logarithm of the total number of symmetrically inequivalent structures are plotted against the configurational unit cell size (measured as the number of atoms). In Fig. 4b we plot the number of ground states versus the same unit cell sizes as in Fig. 4a; this plot is highly dependent on what structure pool we have available. In our work, we have the first $34,368$ structures, but plot Fig. 4b could change if changing this pool. Finally, Fig. 4c shows the fraction in per cents of ground states at each configurational unit cell volume. We can now understand why $\mathcal{A}_{\mathrm{rnd}}$ fares poorly. There is exponentially more structures of larger unit cell sizes so the chances of choosing a large structure is much larger compared to choosing a smaller one. At the same time, the fraction of ground states decreases by orders of magnitude when going from smaller to larger structures. Therefore, most of the time, we do not choose a ground state when we pick a structure at random.

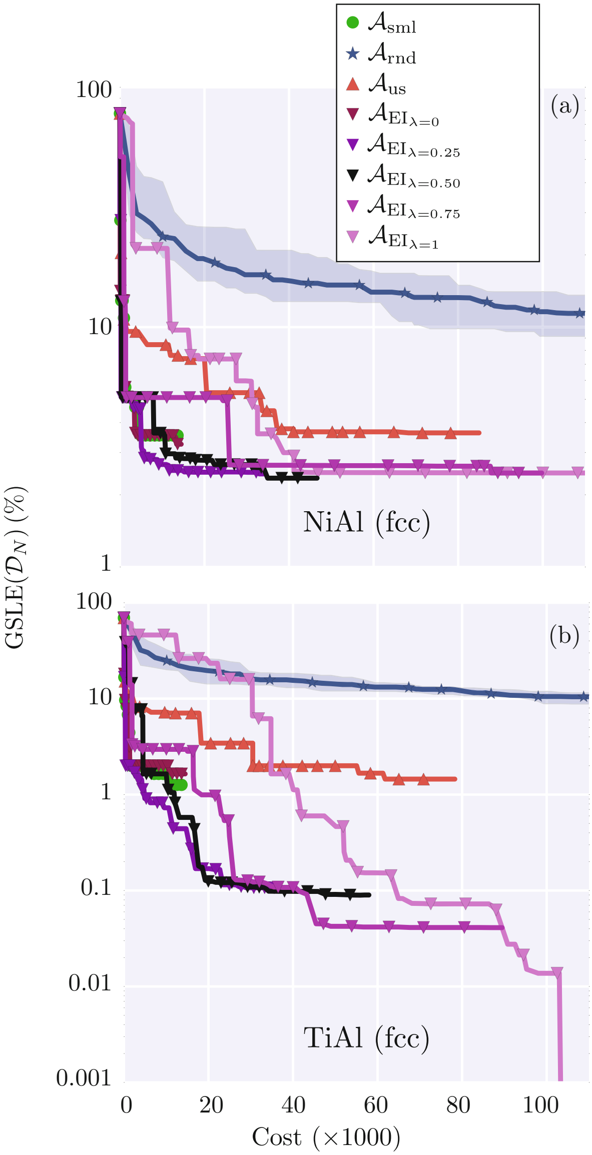


Fig. 3 (Color) Comparison of four different structure acquisition strategies when learning the true ground state line (GSL) for both the (**a**) NiAl and the (**b**) TiAl system. The abscissa shows the total number of observations in the design of the computer experiment and the ordinate reports the normed difference in percents between the GSL constructed from the observations on the abscissa and the true GSL, called GSLE (defined in Eq. (29)). The "random" graph (blue stars) has error bars from 10 different seeds of the random number generator for picking structures at random

Based on these arguments we then expect $\mathcal{A}_{\text{sml}}$ to perform relatively well since it always chooses the smallest structures of which the ratio of ground states is higher. In Fig. 3, we see that this is also largely the case. Notice also that, since it selects the smallest structures first, it never makes it far on the abscissa because the cost is kept at its lowest. The reason why a zero error is not achieved with this strategy is because the GSL is not only made up of smaller structures as is evident from Fig. 4b. Therefore, $\mathcal{A}_{\text{sml}}$ would only find all the true ground state structures by going

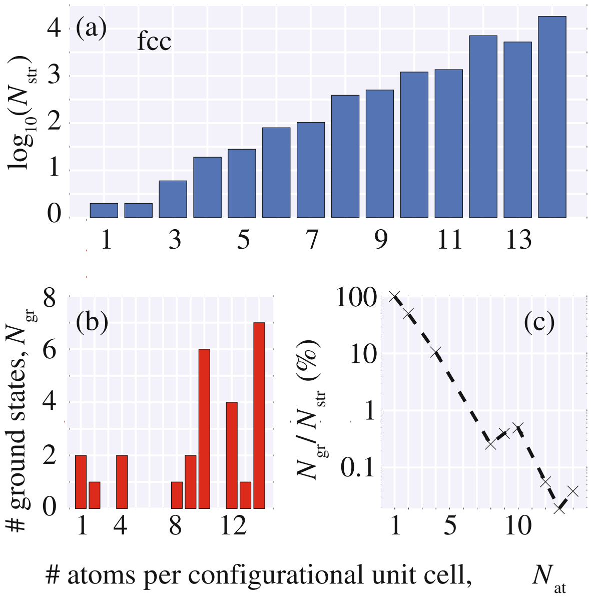


Fig. 4 (Color) (**a**) the logarithm of the number of symmetrically inequivalent structures N_{str} versus the number of atoms in their configurational unit cells N_{at}, for the fcc binary alloy lattices. (**b**) the number of ground states N_{gr} versus N_{at} for fcc TiAl. (**c**) the fraction $N_{\text{gr}}/N_{\text{str}}$, in per cents, of ground state structures, out of the total number of structures in (a) plotted on a logarithmic scale, versus N_{at} (black dashed line is a guide to the eye). All plots share the same abscissa label

through the entire pool of more than $34,000$ structures. Consider next $\mathcal{A}_{\text{us}}$. This method aims for global accuracy of the surrogate to fit the true EAM energy surface which, in light of a data budget, can and often will be a different objective than learning the ground state line. Achieving global accuracy means that we are also demanding good emulating capabilities of large energy structures. However, in order to determine the ground state line, we need not focus on high-accuracy predictions of high-energy structures. We see that the method does eventually achieve a low error.

Finally in this comparison, consider the thermodynamic EI cost-efficient method $\mathcal{A}_{\text{EI}_\lambda}$. Excitingly, this method achieves the lowest overall error of all methods simultaneously at a relatively low temporal cost. Moreover, in the case of TiAl, the thermodynamic EI learns the true ground state line for $\lambda = 1$. The reason is the mix between exploration (choosing large-predictive-variance structures) and exploitation (choosing lowest-predictive-mean structures).

We now address the cost-accuracy trade-off by considering the performance of the $\mathcal{A}_{\text{EI}_\lambda}$ strategy defined in Eq. (28) for various values of λ. First, consider $\lambda = 0$ which should yield a result very similar to the $\mathcal{A}_{\text{sml}}$. We see that this is indeed the case. Small discrepancies are due to the way ties of structure sizes are dealt with. Next, interestingly, we find that by changing λ, the rates in error reductions change dramatically. Between NiAl and TiAl, it is not the same value of λ which achieves the lowest overall error. Notice that, by mixing thermodynamic EI with the cost perspective we obtain lower error than $\mathcal{A}_{\text{sml}}$. We find that, with our surrogate model, structure pools, and alloy materials, a value of λ somewhere around, or less than, 0.5 seems to globally balance well the cost-accuracy trade-off.

3.3.1 Multi-Objective EI-Cost Trade-Off and the Associated Pareto Front

As has been previously mentioned, the structure acquisition strategy $\mathcal{A}_{\mathrm{EI}_\lambda}$, Eq. (28), is a multi-objective optimization task with an associated Pareto frontier. Each point on this frontier represents a distinct optimal structure selection strategy. Which point to choose is subjective and depends, loosely speaking, on how much cost matters compared to GSL accuracy. We now look further at this frontier and ask where the various acquisition strategies discussed earlier are relative to the frontier. We find the answers in Fig. 5 where the structures selected by the various acquisition strategies are marked in an EI-cost plot showing all structures (more than $34,000$) as gray crosses together with the Pareto frontier, shown as a black dashed line, all for the zeroth iteration, i.e., the situation where the first structure is added to the initial pool. The frontier is taking this shape because we wish to maximize the EI while minimizing the structure cost. Different colored diamonds show the particular structure chosen by each strategy. We expect to find $\mathcal{A}_{\mathrm{EI}}$ and $\mathcal{A}_{\mathrm{sml}}$ at the edge of the Pareto frontier since these methods correspond to extreme values of λ. We see that $\mathcal{A}_{\mathrm{rnd}}$ is far from Pareto optimal, but that $\mathcal{A}_{\mathrm{us}}$, at least in the first iteration, lies close to somewhere in between the extreme λ values. We note that $\mathcal{A}_{\mathrm{us}}$ is not on the Pareto frontier itself. At any given iteration, nothing prevents the random or the uncertainty

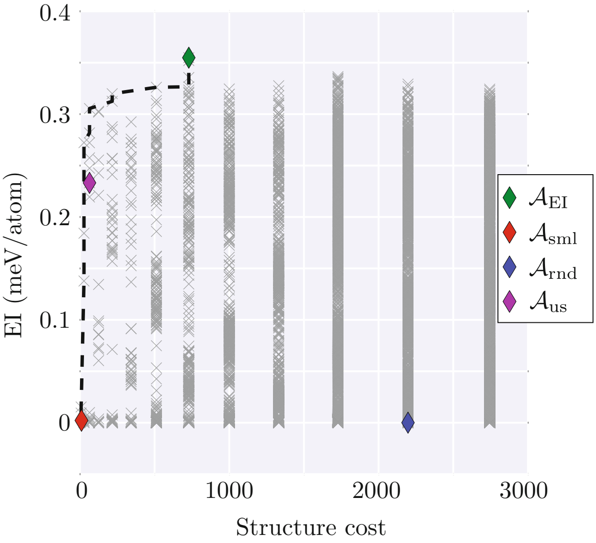


Fig. 5 (Color) The EI (as given by Eq. (25))-cost pareto frontier, built from the surrogate model fitted to the initial structure pool of six structures. Each gray cross is a structure from the unobserved large pool of more than $34,000$ structures. The black dashed line connects structures on the pareto frontier. Diamonds in various colors show where the ith structure acquisition strategy $\mathcal{A}_i$ chooses the first structure to be added to the data pool. Two points are on the frontier itself: $\mathcal{A}_{\mathrm{EI}}$ and $\mathcal{A}_{\mathrm{sml}}$

sampling strategy to be Pareto optimal, but it is unlikely that this will happen for all iterations. By choosing a value for λ in Eq. (28), we can ensure that Pareto optimal structures are always selected.

3.4 Assessing the Effect of the Initial Data Pool $\mathcal{D}_{N_0}$

At this point, we study the effect of the initial NiAl data pool $\mathcal{D}_{N_0}$ on the learning rate of the thermodynamic EI policy (all with $\lambda = 1$). These results are shown in Fig. 6. To make a relative comparison of how fast the lowest overall GSL error is obtained, the abscissa shows the number of observations *added* to the initial data pool, and thus not the total number of structures in the pool. We find that, for NiAl shown in Fig. 6a, starting with six structures the true GSL is learned in 89 iterations. Larger pools all achieve a similar error to each other but do not, in 100 iterations, learn the true GSL. Thus, the rate of learning the true GSL is dependent on the initial pool. This is expected because different initial pools will have different initial surrogate fits which, in turn, will direct the search for new structures into different regions of input space.

Considering then TiAl in Fig. 6b, all starting pools achieve similar error within a couple of percentage points after adding 30 structures, but are differing a lot for additions less than 30 structures. The explanation for this is similar to that given for NiAl.

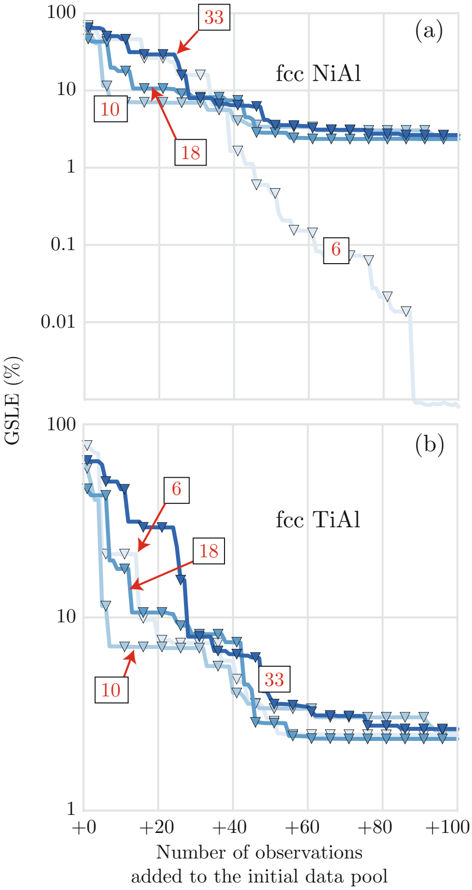
It interesting that a span of almost 30 observations in the initial pool does not appreciably change the lowest overall GSL error achieved within the range of 100 added structures in the case of TiAl. This is promising because we hope the EI method allows us to comfortably start with small datasets.

4 Conclusions

In this work, we introduced optimal information acquisition policies for the discovery of phase diagrams of binary alloys. We proposed policies that balance the maximization of the expected improvement in the thermodynamic potential and the minimization of the cost of simulations. We validated our methodology by learning the GSL of NiAl and TiAl binary alloys and comparing it to the ground truth. We found that the suggested policies outperform naively selected policies in every respect.

The strength of our approach lies on sound theoretical foundations and on its generality. From an application perspective, we plan to use it to actively select informative simulations for the discovery of phase diagrams, band gaps, and transition temperatures in binary alloys and beyond. From a theoretical perspective, we would like to (1) Diminish the reliance of the approach on a pool of structures by directly computing the Pareto front, e.g., via genetic algorithms; (2) Extend the

Fig. 6 (Color) The effect on the learning rate of the thermodynamic EI method (with $\lambda = 1$ in Eq. (28) when changing the initial data pool is investigated. The abscissa is the number of structures *added*. The boxes with red text report the size of the initial data pool for each graph. The ordinate is the per cent GSLE, defined in Eq. (29). Consider the graphs labeled "6" and "10": at the point "+40" on the abscissa, these designs contain a total of 46 and 50 structures, respectively. Results are shown for (**a**) NiAl and (**b**) TiAl



EI so that it can cope with noisy estimates of the thermodynamic potential, e.g., by prematurely stopping the thermodynamic integration required for its evaluation; (3) Allow for the ability to select between models of varying fidelity, e.g., select whether to compute energies with empirical potential or with density functional theory; (4) Design policies that are simultaneously optimal for learning different quantities; (5) Parallelizing information acquisition policies; and many more.

Acknowledgements The work of N.Z. was supported by EPSRC (Grant No. EP/L027682/1). N.Z. thanks the Royal Society for support through a Wolfson Research Merit Award and the Technische Universität München, Institute for Advanced Study for support through a Hans Fisher Senior Fellowship, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under Grant Agreement No. 291763. The work of J.K. at Cornell was supported by the US Department of Energy, Office of Science, Advanced Scientific Computing Research and the Computational Mathematics program of the National Science Foundation (NSF) (award DMS-1214282). I.B. acknowledges the startup support provided by the School of Mechanical Engineering at Purdue University.

References

1. Bahn SR, Jacobsen KW (2002) An object-oriented scripting interface to a legacy electronic structure code. Computing in Science and Engineering 4(3):56–66
2. Bayes M, Price M (1763) An essay towards solving a problem in the doctrine of chances. by the late rev. Mr. Bayes, frs communicated by Mr. Price, in a letter to John Canton, amfrs. Philosophical Transactions (1683–1775) pp 370–418
3. Bertsekas D (2007) Dynamic Programming and Optimal Control, 4th edn. Athena Scientific
4. Bilionis I, Zabaras N (2012) Multi-output local Gaussian process regression: Applications to uncertainty quantification. Journal of Computational Physics 231(17):5718–5746, DOI Doi10.1016/J.Jcp.2012.04.047, URL <GotoISI>://WOS:000305915400009http://ac.els-cdn.com/S0021999112002513/1-s2.0-S0021999112002513-main.pdf?_tid=38a01da2-53ab-11e4-b833-00000aab0f6c&acdnat=1413295738_53a4d40cd278f24f49ec27babcdcd03c
5. Bilionis I, Zabaras N (2012) Multidimensional adaptive relevance vector machines for uncertainty quantification. SIAM Journal on Scientific Computing 34(6):B881–B908, DOI Doi 10.1137/120861345, URL <GotoISI>://WOS:000312737900020http://epubs.siam.org/doi/pdf/10.1137/120861345
6. Bilionis I, Zabaras N, Konomi BA, Lin G (2013) Multi-output separable Gaussian process: Towards an efficient, fully Bayesian paradigm for uncertainty quantification. Journal of Computational Physics 241:212–239, DOI Doi 10.1016/J.Jcp.2013.01.011, URL <GotoISI>://WOS:000317186100012http://ac.els-cdn.com/S0021999113000417/1-s2.0-S0021999113000417-main.pdf?_tid=838a777c-53ab-11e4-950f-00000aab0f01&acdnat=1413295864_b643a385de014cfbd52c16ba5833af3b
7. Bishop CM (2006) Pattern Recognition and Machine Learning, vol 4. Springer New York
8. Boyer R (1996) An overview on the use of titanium in the aerospace industry. Materials Science and Engineering: A 213(1):103–114
9. Brent RP (1971) An algorithm with guaranteed convergence for finding a zero of a function. The Computer Journal 14(4):422–425
10. Broyden C (1969) A new double-rank minimization algorithm. Notices Amer Math Soc p 670
11. Ceder G (1993) A derivation of the ising model for the computation of phase diagrams. Computational Materials Science 1(2):144–150, DOI 10.1016/0927-0256(93)90005-8, URL http://www.sciencedirect.com/science/article/pii/0927025693900058
12. Christen JA, Sansó B (2011) Advances in the sequential design of computer experiments based on active learning. Communications in Statistics-Theory and Methods 40(24):4467–4483
13. Currin C, Mitchell T, Morris M, Ylvisaker D (1988) A Bayesian approach to the design and analysis of computer experiments. Report, Oak Ridge Laboratory
14. Daw MS, Baskes MI (1984) Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. Physical Review B 29(12):6443
15. Daw MS, Foiles SM, Baskes MI (1993) The embedded-atom method: a review of theory and applications. Materials Science Reports 9(7):251–310
16. Dekker T (1969) Finding a zero by means of successive linear interpolation. Constructive aspects of the fundamental theorem of algebra pp 37–51
17. Dreyssé H, Berera A, Wille L, De Fontaine D (1989) Determination of effective-pair interactions in random alloys by configurational averaging. Physical Review B 39(4):2442
18. Ducastelle F, Ducastelle F (1991) Order and phase stability in alloys. North-Holland Amsterdam
19. Durand-Charre M (1997) The microstructure of superalloys. Gordon and Breach Science Publishers, Amsterdam, The Netherlands
20. Fletcher R (1970) A new approach to variable metric algorithms. The computer journal 13(3):317–322

21. Frazier PI, Powell WB, Dayanik S (2008) A Knowledge-Gradient Policy for Sequential Information Collection. SIAM Journal on Control and Optimization 47(5):2410–2439, DOI Doi 10.1137/070693424, URL <GotoISI>://WOS:000260848200008http://epubs.siam.org/doi/pdf/10.1137/070693424
22. Frenkel D, Smit B (2001) Understanding molecular simulation: from algorithms to applications, vol 1. Academic press
23. Garbulsky G, Ceder G (1995) Linear-programming method for obtaining effective cluster interactions in alloys from total-energy calculations: Application to the fcc Pd-V system. Physical Review B 51(1):67
24. Goldfarb D (1970) A family of variable-metric methods derived by variational means. Mathematics of computation 24(109):23–26
25. Hart GL, Forcade RW (2008) Algorithm for generating derivative structures. Physical Review B 77(22):224,115
26. Hohenberg P, Kohn W (1964) Inhomogeneous electron gas. Physical review 136(3B):B864
27. Huang W, Chang Y (1998) A thermodynamic analysis of the Ni-Al system. Intermetallics 6(6):487–498
28. Hunter JD (2007) Matplotlib: A 2D graphics environment. Computing in science and engineering 9(3):90–95
29. Jaynes ET (2003) Probability Theory: The Logic of Science. Cambridge university press
30. Jones D (2001) A Taxonomy of Global Optimization Methods Based on Response Surfaces. Journal of Global Optimization 21(4):345–383, DOI 10.1023/A:1012771025575, URL http://dx.doi.org/10.1023/A%3A1012771025575
31. Jones D, Schonlau M, Welch W (1998) Efficient global optimization of expensive black-box functions. Journal of Global Optimization 13(4):455–492, DOI 10.1023/A:1008306431147, URL http://dx.doi.org/10.1023/A%3A1008306431147
32. Jones E, Oliphant T, Peterson P (2014) SciPy: Open source scientific tools for Python
33. Knowles J (2006) ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. Evolutionary Computation, IEEE Transactions on 10(1):50–66
34. Kohan A, Tepesch P, Ceder G, Wolverton C (1998) Computation of alloy phase diagrams at low temperatures. Computational Materials Science 9(3–4):389–396, DOI 10.1016/S0927-0256(97)00168-7, URL http://www.sciencedirect.com/science/article/pii/S0927025697001687
35. Kohn W, Sham LJ (1965) Self-consistent equations including exchange and correlation effects. Physical Review 140(4A):A1133
36. Kresse G, Furthmüller J (1996) Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. Computational Materials Science 6(1):15–50, URL http://www.sciencedirect.com/science/article/pii/0927025696000080
37. Kresse G, Furthmüller J (1996) Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. Phys Rev B 54:11,169–11,186, DOI 10.1103/PhysRevB.54.11169, URL http://link.aps.org/doi/10.1103/PhysRevB.54.11169
38. Kresse G, Hafner J (1993) Ab initio molecular dynamics for liquid metals. Physical Review B 47(1):558
39. Kresse G, Hafner J (1994) Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. Physical Review B 49(20):14,251
40. Kristensen J, Zabaras NJ (2014) Bayesian uncertainty quantification in the evaluation of alloy properties with the cluster expansion method. Computer Physics Communications 185(11):2885–2892
41. Kristensen J, Bilionis I, Zabaras N (2013) Relative entropy as model selection tool in cluster expansions. Physical Review B 87(17):174,112
42. Landau LD, Lifshitz E (1980) Statistical Physics. Part 1: Course of Theoretical Physics
43. Lizotte D (2008) Practical Bayesian Optimization. Thesis
44. Locatelli M (1997) Bayesian algorithms for one-dimensional global optimization. Journal of Global Optimization 10(1):57–76, URL <GotoISI>://WOS:A1997WJ71300004

45. MacKay DJC (1992) Information-based objective functions for active data selection. Neural Computation 4(4):590–604, URL <GotoISI>://WOS:A1992JF87200009
46. McKinney W (2010) Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, pp 51–56
47. Millman KJ, Aivazis M (2011) Python for scientists and engineers. Computing in Science and Engineering 13(2):9–12
48. Mishin Y (2004) Atomistic modeling of the γ and γ?-phases of the Ni–Al system. Acta Materialia 52(6):1451–1467
49. Mockus J (1972) On bayesian methods for seeking the extremum. Automatics and Computers (Avtomatika i Vychislitelnayya Tekchnika) 4(1):53–52
50. Mockus J (1994) Application of Bayesian approach to numerical methods of global and stochastic optimization. Journal of Global Optimization 4(4):347–365, DOI 10.1007/bf01099263, URL <GotoISI>://WOS:A1994NM81800001
51. Mueller T, Ceder G (2009) Bayesian approach to cluster expansions. Physical Review B 80(2):024,103
52. Nelson LJ, Hart GL, Zhou F, Ozoliņš V (2013) Compressive sensing as a paradigm for building physics models. Physical Review B 87(3):035,125
53. Nelson LJ, Ozoliņš V, Reese CS, Zhou F, Hart GL (2013) Cluster expansion made easy with Bayesian compressive sensing. Physical Review B 88(15):155,105
54. Oliphant TE (2007) Python for scientific computing. Computing in Science and Engineering 9(3):10–20
55. Ong SP, Richards WD, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier VL, Persson KA, Ceder G (2013) Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. Computational Materials Science 68:314–319
56. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al (2011) Scikit-learn: Machine learning in python. The Journal of Machine Learning Research 12:2825–2830
57. Pollock TM, Tin S (2006) Nickel-based superalloys for advanced turbine engines: chemistry, microstructure and properties. Journal of propulsion and power 22(2):361–374
58. Powell WB, Ryzhov IO (2012) Optimal Learning. Wiley Series in Probability and Statistics, Wiley
59. Purja Pun G, Mishin Y (2009) Development of an interatomic potential for the Ni-Al system. Philosophical Magazine 89(34–36):3245–3267, URL NISTInteratomicPotentialsRepository: http://www.ctcms.nist.gov/potentials
60. Raghavan V (2009) Al-Ni-Ti (Aluminum-Nickel-Titanium). Journal of Phase Equilibria and Diffusion 30(1):77–78
61. Raghavan V (2010) Al-Fe-Ni (Aluminum-Iron-Nickel). Journal of Phase Equilibria and Diffusion 31:455–458, DOI 10.1007/s11669-010-9745-1
62. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. Adaptive computation and machine learning, MIT Press, Cambridge, MA, URL Tableofcontentsonlyhttp://www.loc.gov/catdir/toc/fy0614/2005053433.html
63. Rosenbrock CW, Bieniek B, Blum V (2014) Hands-On Tutorial on Cluster Expansion. IPAM Los Angeles, California
64. Sacks J, Welch WJ, Mitchell T, Wynn HP (1989) Design and analysis of computer experiments. Statistical Science 4(4):409–423, URL http://www.jstor.org/stable/2245858
65. Sakiyama M, Tomaszewicz P, Wallwork G (1979) Oxidation of iron-nickel aluminum alloys in oxygen at 600–800 ° C. Oxidation of Metals 13(4):311–330
66. Sanchez J, Ducastelle F, Gratias D (1984) Generalized cluster description of multicomponent systems. Physica A: Statistical Mechanics and its Applications 128(1–2):334–350, DOI 10.1016/0378-4371(84)90096-7, URL http://www.sciencedirect.com/science/article/pii/0378437184900967
67. Sanchez JM (2010) Cluster expansion and the configurational theory of alloys. Phys Rev B 81:224,202, DOI 10.1103/PhysRevB.81.224202, URL http://link.aps.org/doi/10.1103/PhysRevB.81.224202

68. Settles B (2009) Active Learning Literature Survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison
69. Shanno DF (1970) Conditioning of quasi-Newton methods for function minimization. Mathematics of Computation 24(111):647–656
70. Stoloff NS, Sims CT, Hagel WC (1987) Superalloys II. Wiley
71. Taylor A, Floyd R (1953) Constitution of nickel-rich alloys of nickel-chromium-aluminum system. Institute of Metals - Journal 81:451–464
72. Taylor RH, Curtarolo S, Hart GL (2010) Ordered magnesium-lithium alloys: First-principles predictions. Physical Review B 81(2):024,112
73. Torn A, Zilinskas A (1987) Global Optimization. Springer
74. Van Der Walt S, Colbert SC, Varoquaux G (2011) The NumPy array: a structure for efficient numerical computation. Computing in Science &amp; Engineering 13(2):22–30
75. van de Walle A (2009) Multicomponent multisublattice alloys, nonconfigurational entropy and other additions to the Alloy Theoretic Automated Toolkit. Calphad 33(2):266–278, DOI 10.1016/j.calphad.2008.12.005, URL http://www.sciencedirect.com/science/article/pii/S0364591608001314
76. Walle A, Ceder G (2002) Automating first-principles phase diagram calculations. Journal of Phase Equilibria 23:348–359, DOI 10.1361/105497102770331596, URL http://dx.doi.org/10.1361/105497102770331596
77. van de Walle A, Asta M, Ceder G (2002) The alloy theoretic automated toolkit: A user guide. Calphad 26(4):539–553, DOI 10.1016/S0364-5916(02)80006-2, URL http://www.sciencedirect.com/science/article/pii/S0364591602800062

Recent Developments in Spectral Element Simulations of Moving-Domain Problems

Paul Fischer, Martin Schmitt, and Ananias Tomboulides

Abstract Presented here are recent developments in spectral element methods for simulations of incompressible and low-Mach-number flows in domains with moving boundaries. Features include PDE-based mesh motion, implicit treatment of fluid–structure interaction based on a Green's function decomposition, and an arbitrary Lagrangian-Eulerian formulation for low-Mach-number flows that includes an evolution equation for the background thermodynamic pressure. Several examples illustrate the basic principles introduced in the text.

1 Introduction

With advances in high-performance parallel computers, scalable iterative solvers, and high-order discretizations, much progress has been made toward direct numerical simulation (DNS) and large eddy simulation (LES) of transitional and turbulent flows in complex domains. Indeed, researchers now can consider spectral-element-based DNS for flow past wing sections at chord-Reynolds number $Re_c = 400,000$ [1]. DNS of the flow in internal combustion (IC) engines is close at hand, with significant advances recently presented in [2–4].

Since its introduction by Patera [5], several developments have made the spectral element method (SEM) a powerful tool for simulation of turbulent flows in complex geometries. Key advances include high-order operator splitting strategies that lead to decoupled linear symmetric positive definite subproblems at each timestep [6–10];

P. Fischer (✉)
Argonne National Laboratory and University of Illinois at Urbana-Champaign, Urbana, IL, USA
e-mail: fischerp@illinois.edu

M. Schmitt
ETH Zurich, Zurich, Switzerland

Bosch GmbH, Gasoline Systems, Schwieberdingen, Germany
e-mail: Martin.Schmitt2@de.bosch.com

A. Tomboulides
Argonne National Laboratory and Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: ananiast@auth.gr

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_7

fast multilevel preconditioners [11–13] coupled with scalable parallel coarse grid solvers [14–16]; stable formulations for the convective operator [17–19]; and high-performance implementations [20]. Here, we present recent developments in the SEM for simulations of incompressible and low-Mach number flows in domains with moving boundaries. Our interests are in turbulent flows having prescribed boundary motion, such as piston and valve motion in IC engines, and in fluid-structure interactions where the motion of the domain boundary is part of the solution that derives from dynamical constraints coupled with the Navier-Stokes equations.

The standard approach to efficient simulation of turbulent flow is to treat the nonlinear terms explicitly in time, which leaves a linear symmetric unsteady-Stokes operator to be solved, implicitly, at every timestep. (As discussed below, the Stokes problem is typically solved by using an additional time-splitting in order to decouple the pressure and velocity solves.) The justification for this semi-implicit approach to temporal discretization derives from the following. First, the viscous and incompressibility constraints are associated with fast time scales (infinite, in the case of incompressibility, as it derives from letting the speed of sound go to infinity), which warrant implicit treatment. Second, these terms are linear and symmetric, which make them amenable to robust iterative solution strategies such as preconditioned conjugate gradients. Third, explicit treatment of the convection operator avoids solution of a nonlinear nonsymmetric system and requires a mild timestep restriction of $\Delta t = O(|U|\Delta x)$ to ensure stability, where U and Δx are respectively characteristic sizes of the velocity and grid-spacing. Moreover, this timestep restriction is typically comparable to that required from an accuracy standpoint because the principal dynamics of turbulent flow are governed by first-order derivatives in space and time. The stability requirement $\Delta t = O(\Delta x)$ is thus generally not overly constraining.

Moving domains introduce new sources of nonlinearity and stiffness. In closed systems such as internal combustion engines, one must address the changes in thermodynamic pressure and, in the presence of combustion, changes in geometry on short timescales associated with the chemistry. Fluid-structure interaction (FSI) problems, where the solid part of the domain constitutes an additional unknown, introduce additional sources of stiffness associated with disparate timescales between the fluid and solid response. Here, we describe recent developments that address several of these moving-domain issues while retaining the computational efficiency demanded for turbulent flow simulations. The work describes novel developments in time-accurate low-Mach combustion for closed domains and in stable decoupled FSI solution strategies that are particularly appropriate for the response of rigid bodies subjected to forces generated by incompressible flows.

The article is organized as follows. Section 2 provides a review of the arbitrary Lagrangian-Eulerian (ALE) formulation based on the $\mathbb{P}_N - \mathbb{P}_{N-2}$ spectral element method for the incompressible Navier-Stokes equations, as developed by Ho and collaborators [21–23]. Section 3 describes an ALE formulation for low-Mach-number flows that allow compression and expansion of the domain volume.

Specifics of the SEM are provided in Sect. 4, and several schemes for efficient mesh-velocity updates are described in Sect. 5. Section 6 presents a decoupled-implicit formulation for fluid-structure systems with a few degrees of freedom. We give examples in Sect. 7 and a short conclusion in Sect. 8.

2 $\mathbb{P}_N - \mathbb{P}_{N-2}$ Navier-Stokes Formulation

We consider unsteady incompressible flow in a given computational domain $\Omega(t)$ governed by the Navier-Stokes equations,

$$\frac{\partial \mathbf{u}}{\partial t} = -\nabla p + \tfrac{1}{Re}\nabla \cdot (\nabla + \nabla^T)\mathbf{u} - \mathbf{u}\cdot\nabla\mathbf{u}, \qquad \nabla\cdot\mathbf{u} = 0, \tag{1}$$

subject to prescribed velocity conditions on the domain boundary, $\partial\Omega(t)$. Here, $\mathbf{u}(\mathbf{x},t) = (u_1, u_2 u_3)$ represents the fluid velocity components as a function of space, $\mathbf{x} = (x_1, x_2, x_3)$, and time, t; p is the pressure field; and $Re = L_0 U_0/\nu_0$ is the Reynolds number based on a characteristic length scale, L_0, velocity scale, U_0, and kinematic viscosity of the fluid, ν_0. We are interested in moving-geometry simulations where the motion of the domain boundary, $\partial\Omega(t)$, may be either prescribed or unknown, as is the case for fluid–structure interaction problems. Our moving-domain formulation is based on the ALE formulation for the spectral element method developed by Ho and collaborators [21–23]. We review those developments here to set the stage for subsequent sections.

To highlight the key aspects of the ALE formulation, we introduce the weighted residual formulation of (1): *Find* $(\mathbf{u},p) \in X_b^N(\Omega(t)) \times Y^N(\Omega(t))$ *such that*

$$\frac{d}{dt}(\mathbf{v},\mathbf{u}) = (\nabla\cdot\mathbf{v},p) - \frac{1}{Re}(\nabla\mathbf{v},\mathbf{s}) - (\mathbf{v},\mathbf{u}\cdot\nabla\mathbf{u}) + c(\mathbf{v},\mathbf{w},\mathbf{u}), \quad (\nabla\cdot\mathbf{u},q) = 0, \tag{2}$$

for all test functions $(\mathbf{v},q) \in X_0^N(\Omega(t)) \times Y^N(\Omega(t))$. Here, we use the compatible velocity-pressure spaces introduced by Maday and Patera [24]: $X^N(\Omega(t)) \subset H^1(\Omega(t))$ is the set of continuous Nth-order spectral element (SE) basis functions described in Sect. 4; X_b^N is the subset of X^N satisfying the Dirichlet conditions on $\partial\Omega(t)$; X_0^N is the subset of X^N satisfying homogeneous Dirichlet conditions on $\partial\Omega(t)$; Y^N is the space of discontinuous SE basis functions of degree N-2; and H^1 is the usual Sobolev space of functions that are square integrable on $\Omega(t)$, whose derivatives are also square integrable. Furthermore, in (2), we have introduced the $\mathscr{L}^2$ inner product, $(\mathbf{f},\mathbf{g}) := \int_{\Omega(t)} \mathbf{f}\cdot\mathbf{g}\, dV$ and the stress tensor $\mathbf{s}$ having components $s_{ij} := (\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i})$. A new term in (2) is the trilinear form involving the mesh velocity, $\mathbf{w}$,

$$c(\mathbf{v},\mathbf{w},\mathbf{u}) := \int_{\Omega(t)} \sum_{i=1}^{3}\sum_{j=1}^{3} v_i \frac{\partial w_j u_i}{\partial x_j}\, dV, \tag{3}$$

which derives from the Reynolds transport theorem when the time derivative is moved outside the bilinear form, $(\mathbf{v}, \mathbf{u}_t)$.

The advantage of (2) is that it greatly simplifies time differencing and avoids grid-to-grid interpolation as the domain evolves in time. With the time derivative outside the integral, each bilinear or trilinear form involves functions at a specific time, t^{n-j}, integrated over $\Omega(t^{n-j})$. Geometric deformation within elements is specified by a mesh velocity, $\mathbf{w} := \mathbf{x}_t$, that is essentially arbitrary provided that $\mathbf{w}$ is smooth and satisfies the kinematic condition

$$\mathbf{w}\cdot\hat{\mathbf{n}}|_{\partial\Omega} = \mathbf{u}\cdot\hat{\mathbf{n}}|_{\partial\Omega}, \tag{4}$$

where $\hat{\mathbf{n}}$ is the unit normal at the domain surface, $\partial\Omega(t)$.

Our temporal discretization is based on a semi-implicit formulation in which the time derivative at t^n is approximated with a kth-order backward difference formula (BDFk). Terms on the right-hand side of (2) are evaluated either implicitly at t^n or via kth-order extrapolation (EXTk). Specifically, we write

$$\textstyle\sum_{j=0}^{k} \frac{\beta_j}{\Delta t} (\mathbf{v}^{n-j}, \mathbf{u}^{n-j})_{n-j} = (\nabla\cdot\mathbf{v}^n, p^n)_n - \frac{1}{Re}(\nabla\mathbf{v}^n, \mathbf{s}^n)_n + \sum_{j=1}^{k}\alpha_j \widetilde{N}^{n-j} + O(\Delta t^k) \tag{5}$$

$$(q^n, \nabla\cdot\mathbf{u}^n)_n = 0. \tag{6}$$

The subscript on the inner products $(.,.)_{n-j}$ indicates integration over $\Omega(t^{n-j})$. The coefficients β_j and α_j are standard BDFk/EXTk coefficients (e.g., as in Table 1), and the approximations are accurate to $O(\Delta t^k)$, which is the global truncation error for this timestepping scheme. The term $\widetilde{N}^{n-j}$ accounts for all *nonlinear* contributions at time level t^{n-j}, including the mesh motion term (3). For any time level t^m we define

$$\widetilde{N}^m := c(\mathbf{v}^m,\mathbf{w}^m,\mathbf{u}^m)_m - (\mathbf{v}^m, \mathbf{u}^m\cdot\nabla\mathbf{u}^m)_m \tag{7}$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3}\int_{\Omega(t)} v_i^m \left[\frac{\partial w_j^m u_i^m}{\partial x_j^m} - u_j^m \frac{\partial u_i^m}{\partial x_j^m}\right] dV.$$

Table 1 BDFk/EXTk coefficients for uniform Δt

k	β_0	β_1	β_2	β_3	α_1	α_2	α_3
1	1	-1	0	0	1	0	0
2	$\frac{3}{2}$	$-\frac{4}{2}$	$\frac{1}{2}$	0	2	-1	0
3	$\frac{11}{6}$	$-\frac{18}{6}$	$\frac{9}{6}$	$-\frac{2}{6}$	3	-3	1

Moving to the left all terms in (5)–(6) that involve unknowns at t^n and neglecting the $O(\Delta t^k)$ terms, we obtain the update step for (2): *Find* $(\mathbf{u}^n, p^n) \in X_0^N(\Omega) \times Y^N(\Omega)$ *such that, for all* $(\mathbf{v}^n, q^n) \in X_0^N(\Omega^n) \times Y^N(\Omega^n)$,

$$\frac{\beta_0}{\Delta t}(\mathbf{v}^n, \mathbf{u}^n)_n + \frac{1}{Re}(\nabla \mathbf{v}^n, \mathbf{s}^n)_n - (\nabla \cdot \mathbf{v}^n, p^n)_n = r^n, \qquad (q^n, \nabla \cdot \mathbf{u}^n)_n = 0. \quad (8)$$

Here, the right-hand side is

$$r^n = \sum_{j=1}^{k} \left[\alpha_j \widetilde{N}^{n-j} - \frac{\beta_j}{\Delta t}(\mathbf{v}^{n-j}, \mathbf{u}^{n-j})_{n-j} \right]. \quad (9)$$

We note that the test functions $\mathbf{v}$ and q are functions of time as a result of the motion of $\Omega(t)$. In practice, however, all integrals are evaluated in a fixed reference frame and they are stationary basis functions in this frame, integrated against the time-evolving functions with the appropriate Jacobian. Specifically, for the spectral element method, $\Omega(t) = \bigcup_e \Omega^e(t)$, where each element is represented by a map $\mathbf{x}^e(\mathbf{r}, t)$, where $\mathbf{r} \in \hat{\Omega} := [-1, 1]^d$ and d is the number of space dimensions. Such a decomposition is illustrated in Fig. 1 for $d = 2$. The test functions and the underlying bases for the unknowns are taken as tensor product Lagrange interpolating polynomials in $\hat{\Omega}$. Thus, an inner product $\mathbf{I} := (v, u) = \int_\Omega v\,u\,dV$ in the two-dimensional case takes the form

$$\mathbf{I} = \int_{\Omega(t)} v\,u\,dV = \sum_{e=1}^{E} \int_{\Omega^e(t)} v\,u\,dV$$

$$= \sum_{e=1}^{E} \int_{-1}^{1} \int_{-1}^{1} v^e(r, s)\, u^e(r, s, t)\, \mathscr{J}^e(r, s, t)\, dr\, ds, \quad (10)$$

where the Jacobian $\mathscr{J}^e(r, s, t) = \left| \frac{\partial x_i^e}{\partial r_j} \right|$ is the determinant of the $d \times d$ matrix of the metric terms associated with the transformation $\mathbf{x}^e(\mathbf{r}, t)$ that maps $\hat{\Omega}$ to $\Omega^e(t)$.

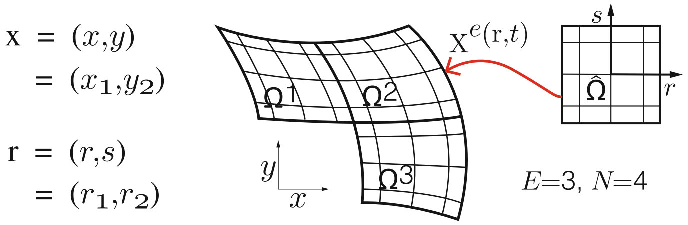


Fig. 1 Two-dimensional illustration of a spectral element domain decomposition

(Here, superscript e refers to element number and should not be confused with the temporal index m or n in (5)–(9).) Because the test functions are stationary in $\hat{\Omega}$ their time derivative following the material points is zero,

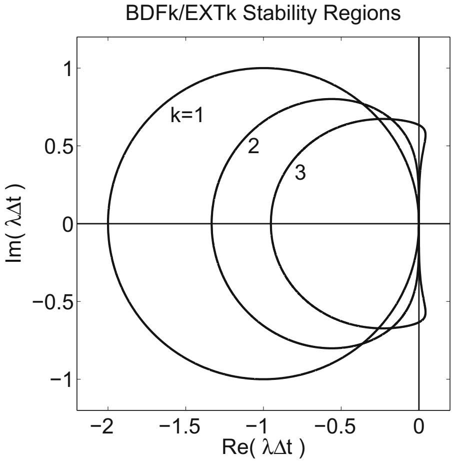
$$\frac{dv_i}{dt} = \frac{\partial v_i}{\partial t} + \mathbf{w} \cdot \nabla v_i \ = \ 0, \tag{11}$$

which is a critical component in the derivation of (2)–(3) because it allows one to substitute $-\mathbf{w} \cdot \nabla v_i$ for $\frac{\partial v_i}{\partial t}$ [21]. Spectral element bases are discussed further in Sect. 4 and in [25].

The timestepping strategy (8) has the advantage that all terms associated with the fast time scales (i.e., the pressure and second-order viscous diffusion terms) are *linear*, which makes an implicit treatment straightforward. Explicit treatment of the nonlinear terms results in a stability constraint on the step size that scales as $\Delta t = O(\Delta x/U)$, corresponding to the standard Courant condition. (The stability regions for BDF*k*/EXT*k* are shown in Fig. 2.) The ALE time advancement from step t^{n-1} to t^n is outlined in Algorithm 1.

In Step 4, one can solve the full Stokes problem using an Uzawa algorithm (e.g, [24, 26]). For large timesteps and highly viscous flows, Uzawa iteration is a reasonable choice. For high Reynolds-number flows, however, an approximate solution strategy via high-order algebraic splitting of the Stokes operator is more effective [7, 9–11, 27]. This splitting can be viewed as a single step in an iterative

Fig. 2 Stability regions for BDF*k*/EXT*k*



Algorithm 1

1. Compute contributions to the right-hand side of (8) from the geometry at t^{n-1}, and combine with values from preceding timesteps t^{n-j}.
2. Update the mesh position $\mathbf{x}^n \in \Omega(t^n)$ using BDF*k*/EXT*k* applied to $\mathbf{x}_t = \mathbf{w}$.
3. Generate geometric terms (per Sect. 4) for Ω^n required to evaluate the operators on the left of (8).
4. Solve the unsteady Stokes system (8) for $(\mathbf{u}^n, p^n)$.
5. Update interior values of $\mathbf{w}^n$ from prescribed boundary values (4).

process. With kth-order extrapolation of the pressure prior to splitting, however, one can realize kth-order accuracy in time without the need for iteration. We refer to [11] for further details and to Sect. 7.1 for temporal convergence results for both stationary and moving domain examples.

Save for the inertial terms associated with advection, (8) implicitly captures all the dynamics of the system including, most importantly, the unsteady components of the fluid inertia. The key point is that (8) is *linear* and thus admits superposition when satisfying dynamical constraints, such as addressed in Sect. 6, with *no need for nonlinear iteration.*

3 $\mathbb{P}_N - \mathbb{P}_N$ Low-Mach-Number Formulation

Many engineering systems feature flows where compressibility is not negligible. In internal combustion engines, for example, thermal dilation and especially compression from the piston motion result in significant density variations. In this section we address recent developments extending the SE-based low-Mach-number formulation of [28, 29] to support moving domains and in particular closed systems of variable volume.

For the numerical simulation of low-speed compressible reacting flows, the existence of acoustic pressure waves severely restricts explicit-integration timestep sizes because of the large discrepancy between the flow velocity and the speed of sound. When acoustic waves are not of interest, regular perturbation techniques can be used to decouple the waves from the governing equations [30–32]. This analysis leads to a decomposition of the pressure as

$$p(x,t) = p_0(t) + \epsilon p_1(x,t), \tag{12}$$

where the hydrodynamic pressure (p_1) is decoupled from the thermodynamic pressure (p_0), and ϵ is defined as γMa^2, where γ is the ratio of specific heat capacities and Ma is the Mach number. The resulting low-Mach-number governing equations for N_g-component reactive gaseous mixtures are the following.

Continuity

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \tag{13}$$

Momentum

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\cdot\nabla\mathbf{u}\right) = -\nabla p_1 + \nabla\cdot(\mu\mathbf{s}) \tag{14}$$

$$\mathbf{s} = \nabla\mathbf{u} + (\nabla\mathbf{u})^T - \frac{2}{3}(\nabla\cdot\mathbf{u})\,I \tag{15}$$

Energy

$$\rho c_p \left(\frac{\partial T}{\partial t} + \mathbf{u} \cdot \nabla T \right) = \nabla \cdot (\lambda \nabla T) - \sum_{i=1}^{N_g} h_i \dot{\omega}_i + \frac{\gamma - 1}{\gamma} \frac{dp_0}{dt} \tag{16}$$

$$c_p = \sum_{i=1}^{N_g} c_{p,i} Y_i \tag{17}$$

Species

$$\rho \left(\frac{\partial Y_i}{\partial t} + \mathbf{u} \cdot \nabla Y_i \right) = -\nabla \cdot (\rho Y_i \mathbf{V}_i) + \dot{\omega}_i \quad i = 1, \ldots, N_g \tag{18}$$

Ideal gas law

$$p_0 = \rho T / W \tag{19}$$

In (13)–(19), $h_i, \dot{\omega}_i, Y_i$, and $\mathbf{V}_i, W_i, c_{p,i}$ are the enthalpy, chemical production term, mass fraction, diffusion velocity, molecular weight, and heat capacity of species i, respectively; λ is the thermal conductivity; p_1 and p_0 are the so-called hydrodynamic and thermodynamic pressures, respectively; $W = \left(\sum_{i=1}^{N_g} Y_i / W_i \right)^{-1}$ is the mean molecular weight; c_p is the mixture heat capacity; and I is the identity matrix. The species diffusion velocities $\mathbf{V}_i$ are given by Fick's law

$$\mathbf{V}_i = -\left(D_i / X_i \right) \nabla X_i, \tag{20}$$

D_i and $X_i = Y_i W_i / W$ being the ith species mixture-averaged diffusivity and mole fraction, respectively. All quantities appearing in the equations above are already nondimensionalized by using reference values for $L_0, U_0, \rho_0, W_0, c_{p0}$ and T_0; in particular p_1 is nondimensionalized by using $\rho_0 U_0^2$ and p_0 by using $\rho_0 \mathscr{R} T_0 / W_0$, where $\mathscr{R}$ is the universal gas constant. The reaction rate constants for the calculation of the chemical source terms $\dot{\omega}_i$ in Eqs. (16) and (18) are assumed to follow an extended Arrhenius expression.

In the low-Mach-number formulation, Eq. (13) is replaced by Eq. (21), which is obtained by combining the continuity (13), energy (16), species (18), and state (19) equations. When the domain volume changes in time, the temporal variation of the thermodynamic pressure, p_0, is nonzero. The governing system for this background pressure is derived below, starting with the low-Mach relationship for the divergence,

$$\nabla \cdot \mathbf{u} = -\frac{1}{\rho}\frac{D\rho}{Dt} = \frac{1}{T}\frac{DT}{Dt} + \sum_{i=1}^{N_g} \frac{W}{W_i}\frac{DY_i}{Dt} - \frac{1}{p_0}\frac{dp_0}{dt}$$
$$= Q_T + \left(\frac{1}{c_p W}\frac{\gamma - 1}{\gamma} - 1\right)\frac{1}{p_0}\frac{dp_0}{dt}. \qquad (21)$$

Here, Q_T is the thermal divergence, which couples the flow field with the temperature and species,

$$Q_T = \frac{1}{\rho}\sum_{i=1}^{N_g} \frac{W}{W_i}\left(-\nabla \cdot \rho Y_i \mathbf{V}_i + \dot{\omega}_i\right) + \frac{1}{\rho c_p T}\left(\nabla \cdot (\lambda \nabla T) - \sum_{i=1}^{N_g} h_i \dot{\omega}_i\right). \qquad (22)$$

In (21), density is determined only by the thermodynamic state T, Y_i, and p_0 and not by the velocity, since acoustic waves are neglected. By contrast, incompressible formulations do not consider the effect of density variations because $\nabla \cdot \mathbf{u} \equiv 0$.

The background thermodynamic pressure, p_0, is obtained by integrating over the domain as follows:

$$\int_{\Omega} \frac{1}{p_0}\frac{dp_0}{dt}\,dv = \left(1 - \frac{1}{c_p W}\frac{\gamma - 1}{\gamma}\right)^{-1}\left(\int_{\Omega} Q_T\,dv - \int_{\partial\Omega} \mathbf{u} \cdot \mathbf{n}\,ds\right). \qquad (23)$$

Because p_0 is a function of time only, the integral on the left corresponds to multiplying the integrand by the domain volume.

Numerical Methodology Spatial discretization of (13)–(21) is based on the weighted residual formulation of the preceding section, save that pressure in this case is continuous and of the same order as the velocity. We consequently refer to this scheme as the $\mathbb{P}_N - \mathbb{P}_N$ method. The resultant system of ordinary differential equations (ODEs) is integrated in time with a high-order splitting scheme for low-Mach-number reactive flows [28]. The low-Mach-number formulation allows the thermochemistry subsystem to be decoupled from the hydrodynamic subsystem, which has the advantage that an appropriate stiff ODE solver can be used to integrate the fully coupled discretized energy and species equations, thus avoiding additional splitting errors.

For the thermochemistry subsystem, the spatially discretized energy, species, and thermodynamic pressure equations, (16)–(18) and (23), are integrated in time with a variable-step kth-order ($k = 1, \ldots, 5$) integrator, CVODE [33]. The density is removed from the equations by using the equation of state. The equations are solved implicitly with the exception of the convecting velocity fields, which are approximated by using high-order explicit extrapolation. The links between thermo- and hydrodynamic subsystems are the density and the divergence constraint (21), which account for the influence of density variations on the velocity field.

The solution of the hydrodynamic subsystem is based on a projection-type velocity correction scheme introduced by Orszag et al. [34]. As a first step, the velocities are updated with the nonlinear terms and a pressure Poisson equation is solved by using boundary conditions based on a third-order extrapolation of viscous contribution of the velocity. Once the hydrodynamic pressure, p_1, is known, the velocity is corrected in a second implicit viscous correction step based on standard Helmholtz equations (36). The low-Mach-number formulation yields kth-order accuracy in time (typ., $k = 3$) for all hydrodynamic variables in combination with minimal splitting errors as shown in [29] and [34]. As is the case for the $\mathbb{P}_N - \mathbb{P}_{N-2}$ formulation (8), this projection scheme amounts to solving, approximately, a linear Stokes problem at each timestep, with boundary conditions being applied at time t^n.

Arbitrary Lagrangian-Eulerian Formulation Extension of the $\mathbb{P}_N - \mathbb{P}_N$ formulation to the ALE framework follows essentially the same steps as for the incompressible $\mathbb{P}_N - \mathbb{P}_{N-2}$ method of Sect. 2. A detailed derivation of the ALE equations can be found in [35]. Sections 7.2 and 7.3 discuss validation of the code modifications for constant and variable thermodynamic pressures.

In addition to solving the ALE momentum equations (13)–(15) and the pressure Poisson equation, the ALE/low-Mach formulation requires the energy and species equations to be integrated together with the single ODE for the thermodynamic pressure (23). Similar to the momentum equation, the ALE form of temperature (energy) and species equations is derived by introducing the mesh velocity, $\mathbf{w}$, in the convective operator. The resulting weighted residual statement reads as follows: *Find* $T,\ Y_i \in X_b^N$ *such that*

$$\begin{aligned}\frac{d}{dt}(\psi, T) - (\psi, \nabla\cdot(\mathbf{w}T) - \mathbf{u}\cdot\nabla T) = \\ -(\nabla\psi, \lambda\nabla T) - \left(\psi, \sum_{i=1}^{N_g} h_i\dot{\omega}_i\right) + \frac{\gamma-1}{\gamma}\left(\psi, \frac{dp_0}{dt}\right)\end{aligned} \tag{24}$$

$$\begin{aligned}\frac{d}{dt}(\psi, Y_i) - (\psi, \nabla\cdot(\mathbf{w}Y_i) - \mathbf{u}\cdot\nabla Y_i) = \\ -(\nabla\psi, \rho D_i\nabla Y_i) + (\psi, \dot{\omega}_i) \quad \forall\,\psi \in X_0^N,\end{aligned} \tag{25}$$

where the ψs are interpreted to be a different set of test functions for each of the thermal/species equations. Here, the surface integrals have been omitted under the assumption that only homogeneous boundary conditions are considered.

In the absence of chemical reactions (i.e., of numerical stiffness) and when the thermodynamic pressure is constant, the ALE energy and species equations are integrated by using the same semi-implicit formulation as with the momentum. In this case, the semi-discrete form of the equations becomes

$$\begin{aligned}\frac{\beta_0}{\Delta t}(\psi, T^n)_n + (\nabla\psi, \lambda\nabla T^n)_n = \\ -\sum_{j=1}^{k}\frac{\beta_j}{\Delta t}\left(\psi, T^{n-j}\right)_{n-j} + \sum_{j=1}^{k}\alpha_j \tilde{N}_T^{n-j}\end{aligned} \tag{26}$$

$$\begin{aligned}\frac{\beta_0}{\Delta t}\left(\psi, Y_i^n\right)_n + \left(\nabla\psi, \rho D_i \nabla Y_i^n\right)_n = \\ -\sum_{j=1}^{k}\frac{\beta_j}{\Delta t}\left(\psi, Y_i^{n-j}\right)_{n-j} + \sum_{j=1}^{k}\alpha_j \tilde{N}_{Y_i}^{n-j},\end{aligned} \tag{27}$$

where

$$\tilde{N}_T^{n-j} = \left(\psi, [\nabla\cdot\mathbf{w}T - \mathbf{u}\cdot\nabla T]^{n-j}\right)_{n-j}$$

and

$$\tilde{N}_{Y_i}^{n-j} = \left(\psi, [\nabla\cdot\mathbf{w}Y_i - \mathbf{u}\cdot\nabla Y_i]^{n-j}\right)_{n-j}.$$

In the presence of chemical reactions and thermodynamic pressure variation, the ALE energy and species equations are integrated implicitly by using CVODE as follows.

$$\begin{aligned}\left(\psi, \frac{dT}{dt}\right)_n &= (\psi, [\tilde{\mathbf{w}} - \tilde{\mathbf{u}}]\cdot\nabla T^n)_n \\ &\quad - (\nabla\psi, \lambda\nabla T^n,)_n - \sum_{i=1}^{N_g}\left(\psi, h_i\dot{\omega}_i^n\right)_n + \frac{\gamma-1}{\gamma}\left(\psi, \frac{dp_0}{dt}\right)_n\end{aligned} \tag{28}$$

$$\begin{aligned}\left(\psi, \frac{dY_i}{dt}\right)_n &= \left(\psi, [\tilde{\mathbf{w}} - \tilde{\mathbf{u}}]\cdot\nabla Y_i^n\right)_n \\ &\quad - \left(\nabla\psi, \lambda\nabla Y_i^n\right)_n + \left(\psi, \dot{\omega}_i^n\right)_n\end{aligned} \tag{29}$$

and

$$\tilde{\mathbf{w}} - \tilde{\mathbf{u}} = \sum_{j=1}^{k}\alpha_j\mathbf{w}^{n-j} - \sum_{j=1}^{k}\alpha_j\mathbf{u}^{n-j}$$

The ALE formulation is thus implemented in the energy and species equations by replacing the fluid velocity $\mathbf{u}$ in the convective term with $(\mathbf{u} - \mathbf{w})$ and by updating the geometry $\Omega(t)$. We note that because CVODE uses adaptive timestepping, the mass matrix must be updated and inverted at intermediate time points in the interval $[t^{n-1}, t^n]$. Fortunately, as shown in the next section, the high-order quadrature of

Algorithm 2

1. Compute T, Y_i, and p_0 at t^n from (28)–(29) and (23) using CVODE with explicit updates of $\mathbf{x} \in \Omega(t)$.
2. Calculate Q_T^n from (22).
3. Update the mesh velocity and hydrodynamic subsystem (13)–(15) using Algorithm 1.

the spectral element method yields a diagonal mass matrix that allows this system to be advanced at low cost. We summarize the low-Mach ALE formulation in Algorithm 2.

4 Spectral Element Method

Here, we describe the spectral element bases, operator evaluation, and implementation of inhomogeneous boundary conditions that are central to our moving-domain simulations. A critical aspect of the SEM is that neither the global nor the *local* stiffness matrices are ever formed. Elliptic problems are solved iteratively and thus require only the action of matrix-vector multiplication. Preconditioning is based on either diagonal scaling or hybrid multigrid-Schwarz methods with local smoothing effected through the use of separable operators [11–13, 36]. Exclusive reliance on matrix-free forms is particularly attractive in an ALE context because the overhead to update the operators as the mesh evolves is effectively nil.

We illustrate the basic components by considering the scalar elliptic problem,

$$-\nabla \cdot \mu \nabla u \;+\; \gamma u = f, \quad u = g \text{ on } \partial\Omega_D, \quad \nabla u \cdot \hat{\mathbf{n}} \;=\; 0 \text{ on } \partial\Omega \backslash \partial\Omega_D, \tag{30}$$

with Dirichlet conditions imposed on $\partial\Omega_D$ and Neumann conditions on the remainder of the boundary, $\partial\Omega\backslash\partial\Omega_D$. The coefficients and data satisfy $\mu > 0$, $\gamma \geq 0$, $f \in \mathscr{L}^2(\Omega)$, and $g \in C^0(\partial\Omega_D)$. This boundary value problem arises in many contexts in our Navier-Stokes solution process. With $\gamma = \beta_0/\Delta t$ and ν a constant, it is representative of the implicit subproblem for the velocity components in (5). With $\gamma = 0$ we have a variable-coefficient Poisson problem that arises in the pressure substep for the low-Mach formulation and in the lifting operators for the mesh velocity that will be introduced at the end of this section.

The discrete variational formulation of (30) is as follows: *Find* $u(\mathbf{x})$ *in* X_b^N *such that*

$$(\nabla v,\, \mu \nabla u) \;+\; (v,\, \gamma u) = (v, f) \qquad \forall\; v \;\in\; X_0^N, \tag{31}$$

where, as in the Navier-Stokes case, X_b^N (X_0^N) denotes the space of functions in X^N that satisfy $u = g$ ($u = 0$) on $\partial\Omega_D$. We symmetrize (31) by moving the boundary data to the right-hand side. If u_b is any known function in X_b^N, the reformulated system is as follows: *Find* $u_0(\mathbf{x})$ *in* X_0^N *such that*

$$(\nabla v,\, \mu\nabla u_0) \;+\; (v,\, \gamma u_0) = (v,f) \;-\; (\nabla v,\, \mu\nabla u_b) \;-\; (v,\, \gamma u_b) \qquad \forall\; v \;\in\; X_0^N, \tag{32}$$

with $u \;:=\; u_0 \,+\, u_b$.

We formally introduce a *global representation* of $u(\mathbf{x})$, which is never used in practice but which affords compact representation of the global system matrices. Let any $u \;\in\; X^N$ be represented in terms of a Lagrange (nodal) interpolating basis,

$$u(\mathbf{x}) = \sum_{\hat{j}=1}^{\bar{n}} u_{\hat{j}}\phi_{\hat{j}}(\mathbf{x}), \tag{33}$$

with basis functions $\phi_{\hat{j}}(\mathbf{x})$ that are continuous on Ω. The number of coefficients, $\bar{n}$, corresponds to all basis functions in X^N. Ordering the coefficients with boundary nodes numbered last yields n interior nodes such that $X_0^N \;=\; \text{span}\{\phi_{\hat{j}}\}_1^n$. Let I_n be the $n \times n$ identity matrix and $R \;=\; [I_n \;\; O]$ be an $n \times \bar{n}$ restriction matrix whose last $(\bar{n} - n)$ columns are empty. For any function $u(\mathbf{x}) \;\in\; X^N$ we will denote the set of $\bar{n}$ basis coefficients by $\underline{\bar{u}}$ and the set of n interior coefficients by $\underline{u}$. Note that $\underline{u} \;=\; R\underline{\bar{u}}$ always holds, whereas $\underline{\bar{u}}_0 \;=\; R^T\underline{u}_0$ holds only for functions $u_0 \in X_0^N$.

We define the stiffness $\bar{A}$ and mass $\bar{B}$ matrices having entries

$$\bar{A}_{ij} := (\nabla\phi_i,\, \mu\nabla\phi_j), \quad \bar{B}_{ij} \;:=\; (\phi_i,\, \phi_j), \quad i,j \in \{1,\ldots,\bar{n}\}^2. \tag{34}$$

The systems governing the interior coefficients of u_0 are the $n \times n$ restricted stiffness and mass matrices, $A = R\bar{A}R^T$ and $B = R\bar{B}R^T$, respectively. A is invertible if $n < \bar{n}$. We refer to $\bar{A}$ as the Neumann operator because it is the stiffness matrix that would result if there were no Dirichlet boundary conditions. It has a null space of dimension one, corresponding to the constant function.[1]

With the preceding definitions, the discrete equivalent of (32) is

$$\underline{v}^T A\, \underline{u}_0 \;+\; \gamma \underline{v}^T B\, \underline{u}_0 = \underline{v}^T R \left[\bar{B}\underline{\bar{f}} \;-\; \bar{A}\, \underline{\bar{u}}_b \;-\; \gamma \bar{B}\, \underline{\bar{u}}_b \right]. \tag{35}$$

Here, we have exploited the fact that u_0 and v are in X_0^N, and for illustration we have made the simplifying assumptions that γ is constant and that $f \;\in\; X^N$. Neither of these assumptions is binding. Full variability, including jumps in μ, γ, and f across element boundaries, can be handled in the SEM.

[1] We remark that $\bar{A}$ governs the pressure in certain Navier-Stokes formulations when the system is closed. A pressure with zero mean is readily computed iteratively by projecting the constant mode out of the right-hand side and out of the pressure with each iteration.

Because (35) holds for all $\underline{v} \in \mathbb{R}^n$, the linear system for the unknown interior basis coefficients is

$$H\,\underline{u}_0 = R\left[\bar{B}\underline{f} - \bar{H}\,\underline{\bar{u}}_b\right], \tag{36}$$

with $\bar{H} := \bar{A} + \gamma\bar{B}$ and $H := R\bar{H}R^T$. The full solution to (30) then is given by (33) plus

$$\underline{\bar{u}} = R^T\underline{u}_0 + \underline{\bar{u}}_b. \tag{37}$$

For the case $f = 0$, we recognize in (36)–(37) the energy-minimizing projection,

$$\underline{\bar{u}} = \underline{\bar{u}}_b - R^T\left(R\bar{H}R^T\right)^{-1}R\bar{H}\,\underline{\bar{u}}_b, \tag{38}$$

which extends the trace of u_b into the interior of Ω in a smooth way provided that γ is also smooth.

Spectral Element Bases In the SEM, the global bases ϕ_j are never formed. Rather, all operations are evaluated locally within each of E nonoverlapping hexahedral (curvilinear brick) elements whose union forms the domain $\Omega = \bigcup_{e=1}^{E}\Omega^e$. Functions in X^N are represented as tensor-product polynomials in the reference element, $\hat{\Omega} := [-1,1]^d$, whose image is mapped isoparametrically to each of the elements, as illustrated for the case $d = 2$ in Fig. 1. As an example, a scalar field $u(\mathbf{r})$ on Ω^e in three dimensions would be represented in terms of local basis coefficients u^e_{ijk} as

$$u^e(\mathbf{r}) = \sum_{k=0}^{N}\sum_{j=0}^{N}\sum_{i=0}^{N} h_i(r)\,h_j(s)\,h_k(t)\,u^e_{ijk}. \tag{39}$$

Here, $\mathbf{r} = [r,s,t] = [r_1,r_2,r_3] \in \hat{\Omega}$ are the computational coordinates,[2] and $h_i(\xi)$ are Nth-order Lagrange polynomials having nodes at the Gauss-Lobatto-Legendre (GLL) quadrature points, $\xi_j \in [-1,1]$. This choice of nodes provides a stable basis and allows the use of pointwise quadrature, resulting in significant savings in operator evaluation. Typical discretizations involve $E = 10^2$–10^7 elements of order $N = 8 - 16$ (corresponding to 512-4,096 points per element). Vectorization and cache efficiency derive from the local lexicographical ordering within each element and from the fact that the action of discrete operators, which nominally have $O(EN^6)$ nonzeros, can be evaluated in only $O(EN^4)$ work and $O(EN^3)$ storage through the use of tensor-product-sum factorization [25, 37].

[2] In this section, we occasionally use "t" to represent the third coordinate in the reference domain $\hat{\Omega}$. It should not be confused with time because there is no temporal variation in the current context.

The geometry, $\mathbf{x}^e(\mathbf{r})$, takes exactly the same form as (39), and derivatives are evaluated by using the chain rule. For example, the pth component of the gradient of u at the GLL node $\boldsymbol{\xi}_{ijk} := (\xi_i, \xi_j, \xi_k)$ is computed as

$$\left.\frac{\partial u}{\partial x_p}\right|_{\boldsymbol{\xi}_{ijk}} = \left.\frac{\partial r_1}{\partial x_p}\right|_{\boldsymbol{\xi}_{ijk}} \sum_{i'=0}^{N} \hat{D}_{ii'} u_{i'jk} + \left.\frac{\partial r_2}{\partial x_p}\right|_{\boldsymbol{\xi}_{ijk}} \sum_{j'=0}^{N} \hat{D}_{jj'} u_{ij'k} + \left.\frac{\partial r_3}{\partial x_p}\right|_{\boldsymbol{\xi}_{ijk}} \sum_{k'=0}^{N} \hat{D}_{kk'} u_{ijk'},$$

where $\hat{D}$ is the one-dimensional derivative matrix on $[-1, 1]$. $\hat{D}_{ij} = \left.\frac{dh_j}{dr}\right|_{\xi_i}$. We note that if the metric terms $\frac{\partial r_q}{\partial x_p}$ are precomputed, then the work to evaluate all components of the gradient, $\frac{\partial u}{\partial x_p}$, is $(6N + 15)EN^3 \approx (6N + 15)\bar{n}$, and the number of memory accesses is $O(\bar{n})$. The work to compute the metrics $\frac{\partial r_q}{\partial x_p^e}$ is similarly $O(N\bar{n})$. Using $\hat{D}$, one evaluates the 3×3 matrix $\frac{\partial x_q^e}{\partial r_p}$, then inverts this matrix pointwise in $O(N^3)$ operations to obtain $(F_{pq}^e)_{\xi_{ijk}} := \left.\frac{\partial r_q}{\partial x_p^e}\right|_{\xi_{ijk}}$. If $\underline{u}^e$ is the lexicographically ordered set of basis coefficients on element Ω^e, its gradient can be compactly expressed as

$$\underline{w}_p^e = \sum_{q=1}^{3} F_{pq}^e D_q \underline{u}^e, \quad p = 1,\ 2,\ \text{or}\ 3, \tag{40}$$

where $D_1 = I \otimes I \otimes \hat{D}$, $D_2 = I \otimes \hat{D} \otimes I$, $D_3 = \hat{D} \otimes I \otimes I$, and, for each p, q and e, F_{pq}^e is a diagonal matrix.

The high order of the SEM coupled with the use of GLL-based Lagrangian interpolants allows the integrals in (34) to be accurately approximated by using pointwise quadrature. In particular, the mass matrix becomes diagonal. For a single element one has

$$\begin{aligned} B_{\hat{\imath}\hat{\imath}'}^e &:= \int_{\Omega^e} \phi_{\hat{\imath}}\, \phi_{\hat{\imath}'}\, d\mathbf{x} = \int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1} \left[h_i(r)h_j(s)h_k(t)\right] \left[h_{i'}(r)h_{j'}(s)h_{k'}(t)\right] \mathscr{J}^e\, dr\, ds\, dt \\ &\approx \sum_{i''j''k''} \rho_{i''}\rho_{j''}\rho_{k''} \left[h_i(\xi_{i''})h_j(\xi_{j''})h_k(\xi_{k''})\right] \left[h_{i'}(\xi_{i''})h_{j'}(\xi_{j''})h_{k'}(\xi_{k''})\right] \mathscr{J}_{i''j''k''}^e \\ &= \rho_i\rho_j\rho_k \mathscr{J}_{ijk}^e\, \delta_{ii'}\delta_{jj'}\delta_{kk'}, \end{aligned} \tag{41}$$

where $\mathscr{J}^e = \left|\frac{\partial x_p^e}{\partial r_q}\right|$ is the pointwise Jacobian associated with the mapping $\mathbf{x}^e(\mathbf{r})$, ρ_j is the quadrature weight corresponding to the GLL point ξ_j, and $\delta_{ii'}$ is the Kronecker delta. For compactness, we have also introduced the lexicographical ordering $\hat{\imath} := i + (N + 1)(j - 1) + (N + 1)^2(k - 1)$. The same map takes the trial function (i', j', k') to $\hat{\imath}'$. The tensor-product form of the local mass matrix is

$B^e = J^e(\hat{B} \otimes \hat{B} \otimes \hat{B})$, where $\hat{B}$ =diag(ρ_k) is the 1D mass matrix containing the GLL quadrature weights and J^e is the diagonal matrix of Jacobian values at the quadrature points.

Combining the mass matrix with the gradient operator yields the local stiffness matrix as typically applied in the SEM, namely,

$$A^e = \sum_{p=1}^{3} \sum_{q=1}^{3} D_p^T \left(\mu\, G_{pq}^e\right) D_q, \qquad G_{pq}^e \;:=\; B^e \sum_{q'=1}^{3} F_{q'p}^e F_{q'q}^e. \tag{42}$$

We note that $G_{pq}^e = G_{qp}^e$ is a symmetric tensor field that amounts to six diagonal matrice of size $(N+1)^3$ for each element Ω^e. Likewise, the variable diffusivity μ is understood to be a diagonal matrix evaluated at each gridpoint, ξ_{ijk}^e. We emphasize that for the general curvilinear element case A^e is completely *full*, with $(N+1)^6$ nonzeros, which makes it prohibitive to form for $N > 3$. However, the factored form (42) is *sparse*, with only $6(N+1)^3$ nonzeros for all the geometric factors G_{pq}^e plus $O(N^2)$ for derivative matrices (and an additional $(N+1)^3$ if μ is variable). The total storage for the general factored stiffness matrix is $\sim 7n_l$, where $n_l = E(N+1)^3$ is the total number of gridpoints in the domain. Moreover, the total work per matrix-vector product is only $\sim 12Nn_l$, and this work is effectively cast as highly vectorizable matrix-matrix products [20, 25, 38].

To complete the problem statement, we need to assemble the local stiffness and mass matrices, B^e and A^e, and apply the boundary conditions, both of which imply restrictions on the nodal values u_{ijk}^e and v_{ijk}^e. For any $u(\mathbf{x}) \in X^N$ we can associate a single nodal value u_g for each unique $\mathbf{x}_g \in \Omega$. where $g \in \{1,\ldots,\bar{n}\}$ is a global index. Let $g = g_{ijk}^e$ be an integer that maps any $\mathbf{x}_{ijk}^e$ to $\mathbf{x}_g$; let $l = i + (N+1)(j-1) + (N+1)^2(k-1) + (N+1)^3(e-1)$ represent a lexicographical ordering of the local nodal values; and let $m = E(N+1)^3$ be the total number of local nodes. We define Q^T as the $\bar{n} \times m$ Boolean gather-scatter matrix whose lth column is $\underline{\hat{e}}_{g(l)}$, where $g(l)$ is the local-to-global pointer and $\underline{\hat{e}}_g$ is the gth column of the $\bar{n} \times \bar{n}$ identity matrix. For any $u \in X^N$ we have the global-to-local map $\underline{u}_L = Q\underline{u}$, where $\underline{u}_L = \{\underline{u}^e\}_{e=1}^E$ is the collection of *local* basis coefficients. With these definitions, the discrete bilinear form for the Laplacian becomes

$$(\nabla v, \mu \nabla u) = \sum_{e=1}^{E} (\underline{v}^e)^T A^e \underline{u}^e = \underline{v}_L^T A_L\, \underline{u}_L = (Q\underline{v})^T A_L\, Q\underline{u} = \underline{v}^T Q^T A_L\, Q\underline{u}, \; = \underline{v}^T \bar{A} \underline{u}.$$

Here A_L = block-diag$\{A^e\}$ is termed the *unassembled* stiffness matrix, and $\bar{A} = Q^T A_L Q$ is the *assembled* stiffness matrix. To obtain the mass matrix, we consider the inner product,

$$(v,u) = \sum_{e=1}^{E} (\underline{v}^e)^T B^e \underline{u}^e \;=\; \underline{v}_L^T B_L\, \underline{u}_L \;=\; (Q\underline{v})^T B_L\, Q\underline{u} \;=\; \underline{v}^T Q^T B_L\, Q\underline{u}, \;=\; \underline{v}^T \bar{B} \underline{u}.$$

Here, $B_L = \text{block-diag}\{B^e\}$ and $\bar{B} = Q^T B_L Q$ are, respectively, the diagonal unassembled and assembled mass matrices comprising local mass matrices, B^e.

We close this section on basis functions by defining elements of the pressure space. For the $\mathbb{P}_N - \mathbb{P}_N$ (low-Mach) formulation described in Sect. 3, we take $Y^N = X^N$. That is, the pressure is continuous and represented by basis functions having the form (39). For the $\mathbb{P}_N - \mathbb{P}_{N-2}$ formulation of Maday and Patera [24], the elements of Y^N have the tensor-product form of (39) except that the index ranges from 0 to N-2 and the nodal points are chosen to be the Gauss-Legendre quadrature points rather than the GLL points. Furthermore, interelement continuity is not enforced on either the pressure, p, or the corresponding test function, q. Element-to-element interaction for the pressure derives from the fact that the velocity $\mathbf{u}$ and test functions $\mathbf{v}$ are in $X^N \subset H^1$. We refer to [11, 24, 25] for additional detail concerning the SEM bases and implementation of the $\mathbb{P}_N - \mathbb{P}_{N-2}$ formulation.

5 Mesh Motion

Mesh displacement is computed by integrating the ODE $\mathbf{w} = \dot{\mathbf{x}}$ in time, where the mesh velocity $\mathbf{w}$ is subject to the kinematic constraint (4). The main idea is to smoothly blend the boundary data into the domain interior. The original SEM-ALE formulation of Ho [21] used an elasticity solver in order to lift the mesh-velocity boundary data to the domain interior. This approach has proven robust for many complex motions, including free-surface applications. It is expensive, however, with the mesh solve costing as much as or more than the velocity/pressure solve.

We have found in several instances that simpler strategies offer significant cost savings and can generate adequate blending functions. For example, for a tensor-product domain with a free surface located at height $z = H(x, y)$ and no motion on the floor at $z = 0$, one can define the vertical mesh velocity satisfying (4),

$$w_z(x, y, z) = \frac{z}{H(x, y)} \frac{\mathbf{u}(x, y, H) \cdot \hat{\mathbf{n}}}{\hat{\mathbf{z}}} \cdot \hat{\mathbf{n}}, \tag{43}$$

where $\mathbf{u}$ is the fluid velocity, $\hat{\mathbf{n}}$ is the unit normal at the surface, and $\mathbf{z}$ is the unit vector in the z direction. This approach has been used in free-surface Orr-Sommerfeld examples [39].

For more complex domains, we typically solve Laplace's equation (i.e., (30) with $\gamma = f = 0$) in order to blend the surface velocities to the interior, relying on the maximum principle to give a bounded interpolant. Fluid dynamics applications often require high-resolution meshes near walls in order to resolve boundary-layer turbulence. If unconstrained, mesh deformation can compromise the quality of these critical boundary-layer elements. The deformation can be mitigated, however, by increasing the diffusivity near the walls so that the mesh velocity tends to match that of the nearby object. The bulk of the mesh deformation is effectively pushed into the far field, where elements are larger and thus better able to absorb significant

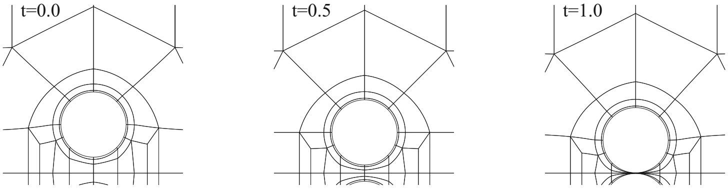


Fig. 3 Two-cylinder mesh deformation resulting from variable-diffusivity solver for mesh velocity

deformation. We usually set $\mu(\mathbf{x}) = 1 + \alpha e^{-\delta^2}$ with $\alpha = 9$ and $\delta := d/\Delta$ the distance to the wall normalized by a chosen length scale, Δ. In the absence of any other scale information, we set Δ equal to the average thickness of the first layer of spectral elements in contact with the given object. To compute d, we use a Euclidian graph-based approximation to the true distance function. A naïve computation of the distance function begins by initializing d to a large number at each gridpoint, setting $d = 0$ on boundary nodes, and then iterating, with each point i assigning d_i to be $\min(d_i, d_j + d_{ij})$ for all points j connected to i, where d_{ij} is the Euclidian distance between i and j. The iteration proceeds until no distances are updated. The idea of using variable diffusivity has been explored by other authors in finite-element contexts where the coefficient is based on local element volumes (e.g., [40, 41]) and can also be applied to the elasticity equations.

Figure 3 shows a close-up of an ALE spectral element mesh for a pair of unit-diameter cylinders moving toward each other until the gap is .03. Here, $\Delta = 0.1$; and a new diffusivity function, μ_{new}, is computed every 100 timesteps based on an updated distance function. In order to make the function smooth in time, the diffusivity is blended with preceding values by using a weighted update, $\mu^n = 0.95\mu^{n-1} + .05\mu_{\text{new}}$. (With a more efficient distance function, one could simply update the diffusivity at every step instead of using a weighted update.) Jacobi-preconditioned conjugate gradient (CG) iteration is used to solve for the mesh velocity. When coupled with projection in time [42], only a few iterations per step are required in order to reduce the CG residual to 10^{-5}. Figure 3 shows clearly that this procedure preserves element shapes near the cylinders except in the gap region where the near-wall elements must yield to the cylinder motion. By tuning the parameters one can ensure that compression in the gap is evenly distributed so that the centermost elements are not squeezed to zero thickness before the near-wall elements yield. We note that because the diffusivity is based on the geometry, there is little hysteresis in the mesh deformation, which is not necessarily true if the mesh diffusivity is based on element sizes.

We remark that if the geometric motion is prescribed, one can solve for the mesh position at a few time points, optimize the mesh at these points (while retaining the base topology), and then use a spline to generate the mesh velocity at all instances in time. Such a strategy would yield optimal meshes that vary smoothly in time and that incur low overhead for mesh motion. The base solutions can be generated in a separate off-line calculation, for example, with the PDE-based approach just described.

6 Fluid–Structure Interaction

Here, we consider systems in which the boundary motion is determined dynamically through interactions with the external flow field, rather than prescribed. A critical feature of these problems is that the resulting system can be extremely stiff. Indeed, because of incompressibility, the pressure responds instantaneously to acceleration of boundaries, with the net effect that the system has added apparent mass arising from the Navier-Stokes equations.

The stiffness associated with fluid-structure interaction (FSI) problems is well known and has been the topic of much recent activity. Several strategies have been pursued to develop robust and fast methods. A particularly robust approach is to use a monolithic scheme with nonlinear iteration to solve for all fluid and solid variables at each step. A comprehensive overview of this strategy is provided by Hron and Turek [43]. Another strategy is to couple independent fluid and structural codes, which offers the potential for using the state of the art from each of the disciplines (e.g., using a structural code with support for contact problems, nonlinear material response, and anisotropic materials). Decoupled methods generally either are explicit or rely on subiterations at each step to improve stability. Gerbeau et al. [44] analyze the stability of several coupling strategies, including subiteration approaches, and identify added mass as one of the principal sources of instability. In a subsequent paper [45], Gerbeau and coworkers identify that the added-mass effect constitutes a linear phenomenon and suggest a coupled, but linear, FSI solution strategy to keep the work low while retaining good stability properties. Farhat et al. [46] demonstrate that a fully explicit subiteration-free strategy using staggered fluid/structure updates can be robust even in the presence of strong added-mass effects for examples having catastrophic (i.e., rapid) structural response.

Recently, a set of schemes with implicit treatment of the added-mass effect have been developed by Banks and coworkers that allow for a decoupled approach without subiteration [47–49]. The authors consider incompressible flows interacting with elastic solids [48] and structural shells [49], as well as FSI for light rigid bodies in compressible flow [47]. The key idea of these papers is to identify the added-mass tensor from a characteristic analysis of the fluid-structure interaction. For the incompressible flow cases, they further introduce a new set of mixed (Robin) boundary conditions for the velocity and pressure, as has been considered by other authors (see, e.g., [50] for an extensive review).

We consider an extension of these ideas to the case of light rigid bodies for incompressible flow. The scheme is fully implicit and exploits the linearity of the unsteady Stokes problem (8). The approach of [47] for rigid-body responses is based on a characteristics analysis associated with compressible flow. The authors identify the interface stress with the difference in velocity between the fluid and the structure. Consideration of such a difference is sensible in the compressible case because it is a measure of the temporal response of the fluid to the motion of the structure. In the case of a rigid solid and an incompressible fluid, however, there is no compliance, and the response is instantaneous. Nonetheless, the added-mass effect is a linear phenomena associated with the acceleration of the object that ultimately manifests as a linear function of the unknown velocity at time t^n. Here, we introduce a Green's function approach to identifying the added mass in the incompressible case and incorporating its effect into the implicit Stokes update step (8) through superposition.[3]

We illustrate the procedure with the example of flow past a cylinder of mass m that is allowed to oscillate in the y direction, subject to a restoring force $F_\kappa = -\kappa\,\eta$, where the positive spring coefficient κ may be a function of the displacement η. In addition to providing a relatively simple model, this problem is of interest in its own right and continues to be a topic of analysis [51, 52]. From Newton's third law, the cylinder motion is governed by

$$m\ddot{\eta} = F_{\text{net}} \;=\; F_f \;+\; F_\kappa. \tag{44}$$

The challenge of (44) is that the fluid forces F_f are strongly dependent on the acceleration of the object, $\ddot{\eta}$, particularly as the mass, m, tends toward zero. In this limit we must have $F_f \;\equiv\; -F_\kappa$ or suffer unbounded acceleration. For this reason, we seek an implicit coupling between (44) and the ALE formulation (8).

We begin with a BDFk/EXTk temporal discretization of (44),

$$\frac{m}{\Delta t}\sum_{j=0}^{k}\beta_j\,\dot{\eta}^{n-j} = F_f^n \;+\; \tilde{F}_\kappa^n, \tag{45}$$

where kth-order extrapolation is used to compute the restoring force,

$$\tilde{F}_\kappa^n = \sum_{j=1}^{k}\alpha_j F_\kappa^{n-j} \;=\; F_\kappa^n \;+\; O(\Delta t^k) \tag{46}$$

Note that $\dot{\eta}^n\,\hat{\mathbf{y}}$, the product of the unknown cylinder velocity at t^n with the unit normal in the y direction, corresponds to the boundary condition on the cylinder surface for $\mathbf{u}^n$ in (8).

We next break F_f^n into two contributions: $F_f^n = F_s + \alpha F_g$, where F_s is the standard fluid lift force that would result from advancing (8) with a given cylinder velocity, $\dot{\eta}_s$, whose value is at our discretion and whose choice is discussed shortly. We denote the solution of this system as $(\mathbf{u}_s,\,p_s)$.

The second part of the force, F_g, is the lift that results from the Green's function pair $(\mathbf{u}_g,\,p_g)$ satisfying the following: *Find* $(\mathbf{u}_g,\,p_g) \in X_1^N \times Y^N$ *such that*

$$\frac{\beta_0}{\Delta t}(\mathbf{v},\mathbf{u}_g)_n \;+\; \frac{1}{Re}(\nabla\mathbf{v},\mathbf{s}_g)_n - (\nabla\cdot\mathbf{v},p_g)_n = 0 \qquad (q,\nabla\cdot\mathbf{u}_g)_n \;=\; 0 \tag{47}$$

[3] We remark that Patera's original SEM paper [5] used a similar Green's function approach to enforce the divergence-free constraint at domain boundaries.

for all $(\mathbf{v}, q) \in X_0^N \times Y^N$, where X_1^N is the subset of X^N that vanishes on $\partial\Omega_D$ save for the cylinder surface, where $\mathbf{u}_g = (0, 1, 0)$. From the velocity pressure pair $(\mathbf{u}_g, p_g)$ we compute the lift F_g. Note that we do not actually solve the unsteady Stokes problem (47), but rather its time-split surrogate consistent with that used to advance (8).

For either of the formulations described in the preceding sections the implicit substep used to update $(\mathbf{u}^n, p^n)$ is *linear*, and superposition may be used to satisfy any number of constraints. The key idea is thus to set

$$\mathbf{u}^n = \mathbf{u}_s + \alpha \mathbf{u}_g, \quad p^n = p_s + \alpha p_g, \quad F^n = F_s + \alpha F_g, \quad \dot{\eta}^n = \dot{\eta}_s + \alpha, \tag{48}$$

where α is chosen to satisfy (45) exactly. Because both sides of (45) are linear in α, one has directly

$$\alpha = \frac{F_s + \tilde{F}_\kappa^n - \frac{m}{\Delta t}\left(\beta_0 \dot{\eta}_s + \sum_{j=1}^{n} \beta_j \dot{\eta}^{n-j}\right)}{\frac{m}{\Delta t}\beta_0 - F_g}. \tag{49}$$

We make several remarks concerning this procedure. First, the case $m = 0$ presents no difficulty because F_g is never zero. In fact, F_g is negative (the restoring force is opposite the applied velocity perturbation), so (49) can never suffer from a vanishing denominator. Second, $(\mathbf{u}_s, p_s, F_s)$ results from the standard Navier-Stokes update. Most of the expense is in iterative solution of the pressure, which can be minimized if the apparent acceleration of the cylinder is zero, that is, if $\dot{\eta}_s := -(\sum_{j=1}^{k} \beta_j \dot{\eta}^{n-j})/\beta_0$. The variation in $\dot{\eta}^n$ is made up by the contribution from the Green's function, whose cost is independent of α. For computation of both the s and g variables, significant cost savings are realized by using initial guesses that are projections onto the space of prior solutions [42]. We remark further that F_g represents the influence of the added mass. From the denominator of (49) we see that the effective added mass is

$$m_a = -\Delta t F_g / \beta_0.$$

The only time dependence for $(\mathbf{u}_g, p_g)$ arises from the fact that the domain is time varying. Otherwise, one could compute $(\mathbf{u}_g, p_g)$ once in a preprocessing step and reuse it for all time provided that β_0 and Δt are invariant. We use such an approach for rotating cylinder cases where the geometry is indeed invariant.

We note that explicit computation of $\tilde{F}_\kappa^n$, which readily admits incorporation of fully nonlinear responses (e.g., [52]), is a potential source of instability. Under standard conditions, however, the Courant restriction on the fluid velocity update will suffice to ensure that explicit treatment of the mass-spring system will be stable. Consider the case where the spring is sufficiently stiff such that stability is a concern. The dominant eigenvalue in this case is $\lambda_\kappa := \pm i\sqrt{\kappa/m_v}$, where $m_v = m + m_a$ is the nonzero virtual mass that includes the added mass. Figure 2 shows that the BDFk/EXTk stability region for $k = 3$ includes a portion of the imaginary axis and that this system will be stable when $|\lambda_\kappa \Delta t| < 0.6$. For the same timestepper, explicit treatment of advection imposes a stability constraint of the form $\Delta t \lambda_{CFL} \leq 0.6$, where, for the SEM,

$$\lambda_{CFL} \approx 1.5 \max_i \left| \frac{u_i}{\Delta x_i} \right|, \tag{50}$$

with u_i and Δx_i representing characteristic velocities and grid spacing at gridpoint $\mathbf{x}_i$. (See Fig. 3.5.2 in [25].) If the cylinder is oscillating in the stiff-spring limit with amplitude η_0, then the velocity scale is $|u_i| \approx \eta_0 \sqrt{\kappa/m_v}$, and we have

$$\Delta t \leq \frac{0.6}{\max(1, 1.5\frac{\eta_0}{\Delta x_i})} \sqrt{\frac{m_v}{\kappa}}. \tag{51}$$

The Courant condition will hold under the assumption that the displacement is larger than the characteristic grid spacing (i.e., $\eta_0/\Delta x_i > 1$). However, if the spring is so stiff that translational motion is suppressed ($\eta_0 < \Delta x_i$), then the Courant condition due to spring motion will not come into play, and the stiffness associated with a large spring constant could restrict Δt.

Extension of the Green's function approach to more structural degrees of freedom is straightforward. For each DoF, one generates a solution pair $(\mathbf{u}_g, p_g)$, $g = 1, \ldots, N_{DoF}$, each of which leads to a nontrivial force or torque on each and every object. One obtains an $N_{DoF} \times N_{DoF}$ matrix corresponding to (49) whose solution results in an implicit solution to all the dynamical constraints. For a few DoFs, solution of this system is not a challenge. However, the cost of solving N_{DoF} systems for the independent Green's functions can become prohibitive if N_{DoF} becomes too large. Another extension is to use the Green's function approach to remove the stiffest contributions to an otherwise explicitly coupled strategy. In particular, for compressible solids the mean compression mode (i.e., the volumetric change) induces long-range accelerations in the fluid. It is straightforward to compute the associated added mass by solving for the Greens function associated with the mean compression mode and to add a multiple of this solution to obtain the requisite force balance, as done in (48)–(49). We are currently investigating this idea, to be discussed in a future article, in the context of coupling Nek5000 with a large nonlinear structures code.

7 Results

Here, we consider several examples that illustrate the techniques introduced in the preceding sections. These methods have been implemented Nek5000, which is an open source spectral element code for fluid, thermal, and combustion simulations that scales to over a million processors [53].

7.1 Temporal-Spatial Accuracy

We illustrate the spatial and temporal convergence of the baseline $\mathbb{P}_N - \mathbb{P}_N$ and $\mathbb{P}_N - \mathbb{P}_{N-2}$ discretizations using the BDFk/EXTk schemes outlined in the text.

We consider the family of exact eigenfunctions for the incompressible Stokes and Navier-Stokes equations derived by Walsh [54], which are generalizations of Taylor-Green vortices in the periodic domain $\Omega = [0, 2\pi]^2$. For all integer pairs (m, n) satisfying $\lambda = -(m^2 + n^2)$, families of eigenfunctions can be formed by defining streamfunctions that are linear combinations of the functions

$$\cos(mx)\cos(ny),\ \sin(mx)\cos(ny),\ \cos(mx)\sin(ny),\ \sin(mx)\sin(ny).$$

With the eigenfunction $\mathbf{u}^0 := (-\psi_y, \psi_x)$ as an initial condition, a solution to the Navier-Stokes equations is $\mathbf{u} = e^{\nu\lambda t}\mathbf{u}^0(\mathbf{x})$. Figure 4 shows the vorticity for a case proposed by Walsh, with $\psi = (1/4)\cos(3x)\sin(4y) - (1/5)\cos(5y) - (1/5)\sin(5x)$. The analytical solution is stable only for modest Reynolds numbers. Interesting long-time solutions can be realized, however, by adding a relatively high-speed mean flow $\bar{\mathbf{u}}$, in which case the exact solution is

$$\tilde{\mathbf{u}}(\mathbf{x}, t) = \bar{\mathbf{u}} + e^{\nu\lambda t}\mathbf{u}^0[\mathbf{x} - \bar{\mathbf{u}}t], \tag{52}$$

where the brackets imply that the argument is modulo 2π in x and y. By varying $\bar{\mathbf{u}}$, one can advect the solution a significant number of characteristic lengths before the eigensolution decays.

We typically run this case with periodic boundary conditions, but that is not as strong of a test as having Dirichlet conditions, which are a well known source of difficulty in time advancement of the incompressible Navier-Stokes equations [27, 34]. In the present case, since we have an exact solution as a function of space and time we can run the Dirichlet case with the solution prescribed on all four sides of the domain. Starting with the initial condition of Fig. 4 (left), we take $\nu = .01$

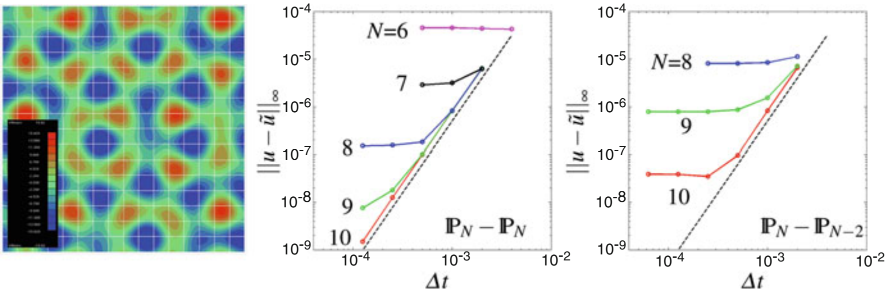


Fig. 4 Eddy solution results at $Re = 100$: (left) vorticity at $t = 0$ for the initial condition (52), (center) maximum pointwise error at time $t = 2\pi$ as a function Δt for $\mathbb{P}_N - \mathbb{P}_N$ with $N = 6$–10 and (right) for $\mathbb{P}_N - \mathbb{P}_{N-2}$ with $N = 8$–10. The dashed curve is $500\Delta t^3$

and $\bar{\mathbf{u}} = (1, .3)$ and evolve the solution to a final time $T = 2\pi$. In that time, the peak amplitude of the perturbation velocity, $\mathbf{u} - \bar{\mathbf{u}}$, decays from 2.0 to 0.46. The mesh consists of a 16×16 array of square spectral elements.

The right two panels in Fig. 4 show the maximum pointwise error in the x-component of the velocity for $\mathbb{P}_N - \mathbb{P}_N$ (center) and $\mathbb{P}_N - \mathbb{P}_{N-2}$ (right) as a function of Δt for several values of N. The general trend is that the error is dominated by spatial error for sufficiently small values of Δt and becomes dominated by temporal error as Δt is increased until the CFL condition is violated, at which point the solution is unstable. Both discretizations demonstrate $O(\Delta t^3)$ accuracy for the velocity and both show exponential convergence in space. For small Δt, increasing the polynomial order by just 1 yields more than an order-of-magnitude reduction in error until the curve hits the temporal-error threshold. Notice that the relatively poor performance of $\mathbb{P}_N - \mathbb{P}_{N-2}$ may be explained by lack of resolution for the pressure. Based on this argument, one would expect the $N = 10$ error for $\mathbb{P}_N - \mathbb{P}_{N-2}$ to be about the same as $N = 8$ for $\mathbb{P}_N - \mathbb{P}_N$. Indeed, the $N = 10$ $\mathbb{P}_N - \mathbb{P}_{N-2}$ result is bracketed by $N = 8$ and 9 for $\mathbb{P}_N - \mathbb{P}_N$. We note that for this case the resolution of the pressure is a gating issue because its maximum wavenumber is essentially twice that of the velocity, as must be the case given that the pressure is the only term that can cancel the quadratic product involving the velocity eigenfunctions.

We next use the Walsh example to test our ALE formulations. Once again we have inhomogeneous Dirichlet conditions on all of $\partial\Omega$ corresponding to $\tilde{\mathbf{u}}$ (52). We prescribe the mesh velocity, and for these tests we also lift the kinematic constraint (4) since there is no need for the boundary to be a material surface. We take an initial configuration $(x^0, y^0) \in \Omega^0 = [0, 7]^2$ and evolve this with the prescribed mesh velocity

$$\dot{x} = \omega \cos(\omega t)\ \sin(\pi y^0/7), \tag{53}$$

$$\dot{y} = \omega \cos(\omega t/2)\ \sin(\pi x^0/7)(2y - 1), \tag{54}$$

with $\omega = 5$. Configurations of the domain at two time points are shown in Fig. 5, from which one can see that this is not a volume-preserving transformation. Of course it does not need to be because the known boundary data corresponds to a divergence-free field at each point in space and time. The rightmost panel in Fig. 5 shows that third-order accuracy is once again attained, albeit with a larger error than for the nonmoving case of Fig. 4. Somewhat surprisingly, the mesh motion leads to a greater increase in temporal error than the increase in spatial error that one might expect from the deformation of the elements. This increased temporal error results from the rapid mesh motion combined with the relatively high spatial wavenumber of the solution, which gives rise to rapid fluctuations in $\mathbf{u}_t$.

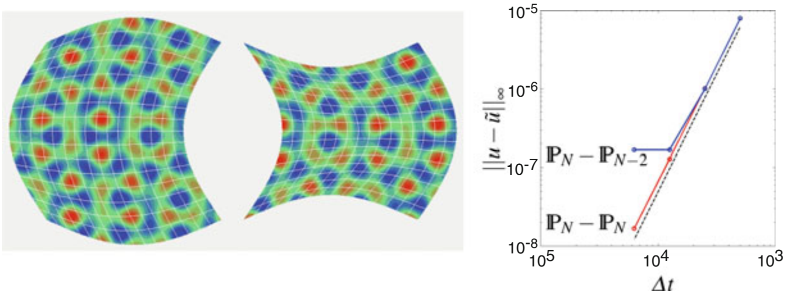


Fig. 5 Eddy solution results at $Re = 100$ for moving domain case: (left and center) vorticity and domain configurations at two timepoints, (right) maximum pointwise error vs. Δt at time $t = 7$ for $\mathbb{P}_N - \mathbb{P}_N$ and $\mathbb{P}_N - \mathbb{P}_{N-2}$ with $N = 10$. The dashed curve is $50000\Delta t^3$

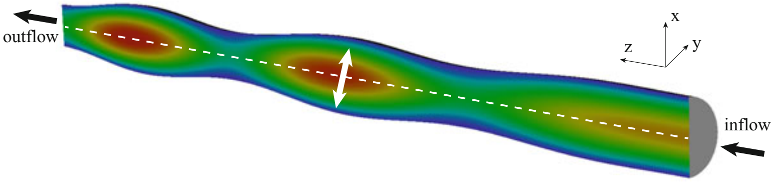


Fig. 6 Distribution of velocity magnitude for the $\mathbb{P}_N - \mathbb{P}_N$ approach on a vertical slice at t = 50

7.2 Constant Pressure Example

The discretization of the convective term including the mesh velocity used in the newly implemented ALE method in the $\mathbb{P}_N - \mathbb{P}_N$ formulation is identical to the one used in the $\mathbb{P}_N - \mathbb{P}_{N-2}$ approach by Ho and Patera, which was extensively validated in [21] and [23]. The accuracy of this scheme was also assessed in [55] using an analytic solution in an expanding mesh setup.

For the 3D case, verification of the $\mathbb{P}_N - \mathbb{P}_N$ implementation begins with tube setup of Fig. 6. The moving mesh in this case generates a peristaltic pumping that strongly influences the temporal and spatial evolution of the velocity field. The pipe has a base radius $R = 1/2$ and length $L = 16$. The prescribed mesh velocity is

$$w_x = -W\frac{x}{R}\cos(kz - \omega t), \quad w_y = -W\frac{y}{R}\cos(kz - \omega t), \quad w_z = 0, \qquad (55)$$

where $W := A/\omega$ is the velocity amplitude and $A := 0.1\ \tanh(0.2z)\ \tanh(0.2t)$ is the amplitude of the displacement. The prescribed wavenumber is $k = \pi/3$, and the frequency is $\omega = 1$. The Reynolds-number is always below 200 so that the flow remains laminar. At the inflow a steady parabolic velocity profile with a maximum axial velocity $u_z = 1$ at the cylinder center is imposed while at the outflow zero-Neumann boundary conditions are used. At the pipe walls the velocity is set equal to the mesh velocity in order to prevent a flow across the walls. The numerical setup including the mesh is given in the example `peris` of the Nek5000 package.

Fig. 7 Instantaneous and averaged axial velocity for $\mathbb{P}_N - \mathbb{P}_N$ and $\mathbb{P}_N - \mathbb{P}_{N-2}$ at $t = 50$

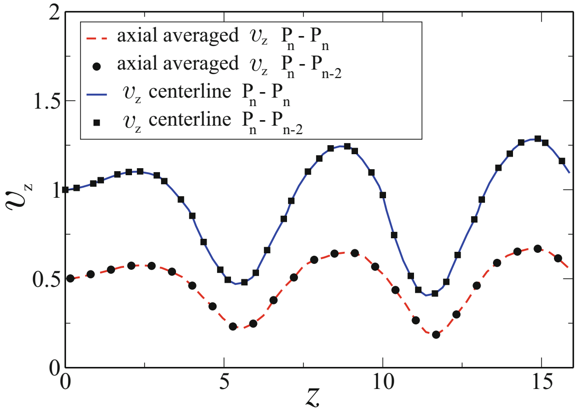


Figure 6 shows the velocity magnitude $|\mathbf{u}|$ distribution on an axial slice through the pipe. The highest flow velocities can be observed in regions with larger pipe diameters. The velocity magnitude of the $\mathbb{P}_N - \mathbb{P}_N$ ALE formulation compared with the $\mathbb{P}_N - \mathbb{P}_{N-2}$ results at $t = 50$ in Fig. 7. The dashed line and the circle markers represent the averaged axial velocity magnitude versus the channel length, while the solid line and the square markers indicate the velocity magnitude along the centerline (marked by the dashed line in Fig. 6). The mean and the instantaneous velocity magnitudes show excellent agreement between the two formulations.

The implementation of the ALE approach in the temperature equation is verified by simulating in the same setup using non-isothermal conditions at the inflow. Initial and boundary conditions for the velocity field and mesh movement are identical to the preceding flow case. The temperature at the walls is fixed to T/Tref =1, where Tref = 300 K. At the inflow a parabolic temperature profile is imposed with a maximum temperature of T/Tref = 1.125 in the pipe center, while zero-Neumann boundary conditions are used at the outflow boundary. A homogeneous N_2/O_2 mixture ($Y_{O_2} = 0.21$, $Y_{N_2} = 0.79$) flows into the channel at a constant pressure of 1 atm. The temperature difference between pipe wall and inflow is chosen low enough to limit its influence on the flow field, because in the $\mathbb{P}_N - \mathbb{P}_{N-2}$ formulation the incompressible Navier-Stokes equations are solved, whereas the $\mathbb{P}_N - \mathbb{P}_N$ approach is based on the low-Mach-number formulation.

The computed temperature fields are shown at $t = 8$ in Fig. 8 (left). The decreasing temperatures in the flow direction are due to the cooler pipe walls. The local temperature peaks in the thicker pipe segments are due to the larger distance from the cylinder wall. The temperature distributions show excellent agreement using the two formulations, as also seen in the instantaneous centerline profiles of Fig. 8 (right).

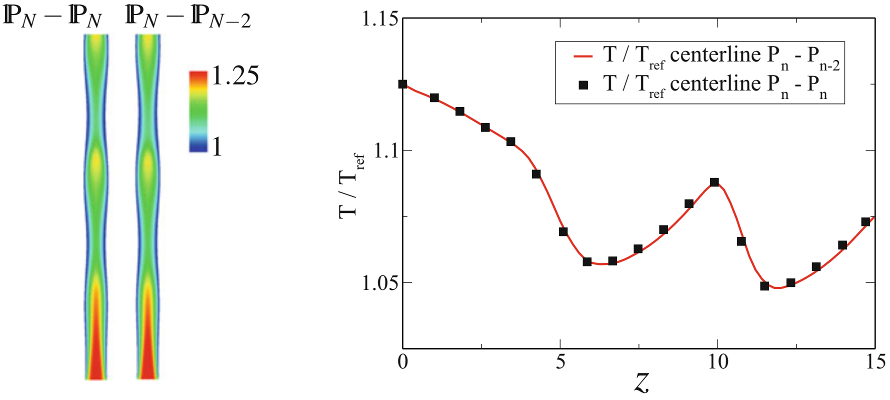


Fig. 8 Comparison of temperature for the $\mathbb{P}_N - \mathbb{P}_{N-2}$ and $\mathbb{P}_N - \mathbb{P}_N$ ALE formulations at $t = 8$: (left) centerplane distribution at $t = 8$ and (right) instantaneous centerline temperatures

7.3 Varying Pressure Examples

The implementation of the variable thermodynamic pressure is validated by comparison with a zero-dimensional CHEMKIN [56] simulation of isentropic compression. A two-dimensional setup with a constant width of 75 mm and an initial height of 90 mm is compressed until a height of 15 mm is reached, resulting in a compression ratio of 6. The piston speed is 200 rpm. Homogeneous conditions for temperature (T = 819.45 K), pressure (p = 1 atm), and composition (YN2 = 0.7288, YO2 = 0.1937 and YCH4 = 0.0775) are used at BDC. Zero-velocity boundary conditions are employed at the liner and the cylinder head, and the piston velocity is imposed at the piston. Zero-flux conditions are imposed for the temperature and species boundaries at all walls. For the homogeneous adiabatic CHEMKIN calculation the same geometry and initial conditions are considered. In both cases, the chemical reactions are calculated based on a reduced mechanism for CH4 combustion with 21 species and 87 reactions.

In Fig. 9, the computed temperature and pressure time-histories are compared with the 0-D CHEMKIN calculation. At time $t = 0$ the piston is at BDC and at $t = 4$ at TDC. The continuously increasing temperature during compression results in autoignition at a nondimensional time $t = 3$. At TDC the temperature and pressure are 2731 K and 20 atm, respectively. The plots show nearly identical evolutions of the temperature and pressure profiles. The minimal offset in the autoignition timing lies within the uncertainty of numerical settings such as the chosen timesteps or imposed tolerances.

We next consider an example of fully turbulent compression from the direct numerical simulations (DNS) presented in [2, 4]. The initial condition at bottom dead center (BDC) (180^o CA) were derived by a precursor DNS of the intake stroke simulating the mixing of a unburnt $\lambda = 2$ H2/air mixture at 500 K in the

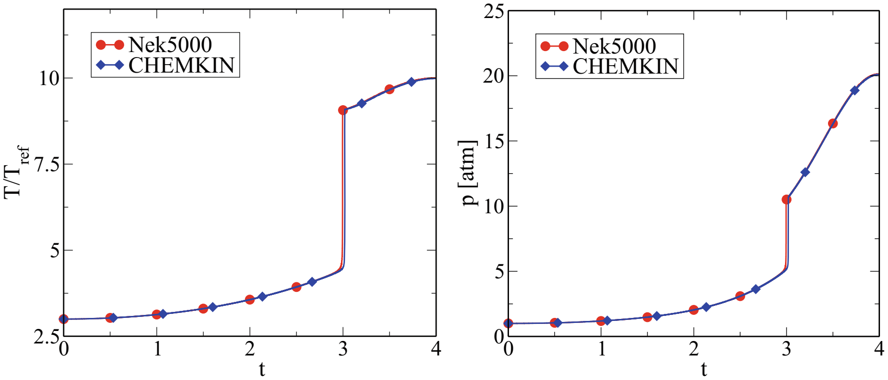


Fig. 9 Comparison of the temperature (left) and pressure (right) evolution during compression between CHEMKIN and Nek5000

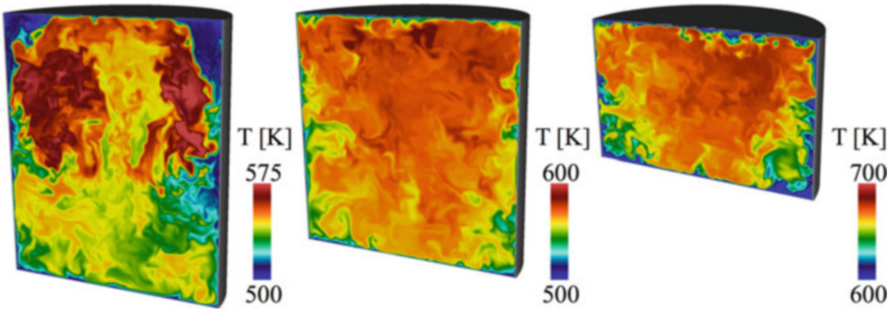


Fig. 10 Centerplane temperature distributions during a compression stroke for an engine-like flow configuration at 180^o, 225^o, and 270^oCA

intake channel with a burnt $\lambda = 2$ H2/air mixture at 900 K in the cylinder. During compression the wall temperature is fixed to 500 K, and the Reynolds number based on cylinder diameter and maximum piston velocity is $Re = 2,927$.The temperature rise resulting from compression is evident in Fig. 10, which shows the temperature distributions at 180^o, 225^o, and 270^oCA. The relatively cool region at the bottom of the cylinder results from the piston scouring cold fluid from the walls and the relatively hot regions in the upper part of the cylinder at 180^oCA are related to hot EGR gases entrained into the core of the ring vortex generated during the intake stroke. As demonstrated in [2, 4], the final temperature distribution is not strongly dependent on the initial thermal distribution; one obtains essentially the same distribution at 270^oCA even when the initial distribution at 180^oCA is uniform. A detailed analysis of the flow and temperature field evolutions during compression can be found in [3].

7.4 Dynamic Response

We illustrate the implicit fluid–structure interaction formulation by considering the case of flow past a cylinder of radius R_0 and mass $m = \rho_c \pi R_0^2$ that is allowed to oscillate in the y-direction subject to a spring constant $\kappa = (2\pi f_b)^2$. Here, the density of the cylinder is ρ_c; the characteristic length scale is the cylinder diameter $D_0 = 2R_0$; and the time scale is the convective time, $\tau := D_0/U_0$, where U_0 is the inflow velocity. We assume the fluid density $\rho = 1$ and define the Reynolds number $Re = U_0 D_0/\nu_0$. We consider $Re = 100, f_b = 0.167$, and $\rho_c = 0$ and 10. Several authors have studied the $\rho_c = 10$ case under these conditions and found the displacement amplitude to be $\eta_{\max}$ in interval 0.49 to 0.503 [52, 57, 58].

The cylinder is centered at $(x, y) = (0, 0)$, and the domain consists of 218 spectral elements of order $N = 14$ with inflow conditions $(u, v) = (1, 0)$ at $x = -13.75$, homogeneous Neumann (outflow) conditions at $x = 38.75$, and periodic boundary conditions at $y = \pm 25$. The timestep is $\Delta t = .005$. A close-up of the mesh and the vorticity at the peak vertical displacement is shown in Fig. 11a. Time traces of the displacement for $\rho_c = 0$ and 10 are shown in Fig. 11b. These cases were started with an initial condition corresponding to a fully developed von Karman street at $Re = 100$. The asymptotic amplitude and frequencies were found by a nonlinear least-squares fit (over a longer time than shown in the figure) to be $A = .0242$ and $\omega = 1.075$ for $\rho_c = 0$ and $A = .505$ and $\omega = 1.040$ for $\rho_c = 10$. For these cases, the added mass from (50) is $m_a \approx 1.234$ times the displaced mass, which is slightly greater than the unit value predicted by potential theory for flow past a cylinder. This increase is explained by the fact that the unsteady Stokes subproblem, which includes the Δt time constant, entrains additional mass due to viscous effects. With a reduction in Δt and viscosity, (47) with (50) predicts the potential flow result to within five significant digits.

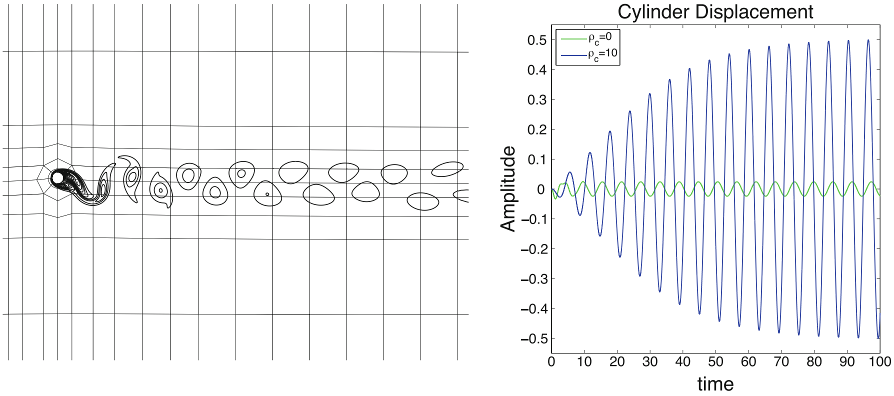


Fig. 11 Sprung cylinder example: (**a**) vorticity and part of the domain showing the spectral element boundaries at peak displacement; (**b**) amplitude and frequency as a function of cylinder mass

8 Conclusions

We have described recent advances in the SEM that target efficient simulation of turbulent flows in moving domains. A new ALE-based low-Mach formulation has been introduced that allows simulation of turbulence in closed domains such as IC engine cylinders. Several examples attest to the fidelity of this approach when compared with the baseline $\mathbb{P}_N - \mathbb{P}_{N-2}$ formulation [21, 23, 24], with analytical solutions in two dimensions, and with the zero-dimensional results of CHEMKIN [56]. Strategies for efficient mesh motion have been described, including the use of variable-coefficient Laplace solvers with projection in time to yield low-cost extension of boundary data into the domain interiors with controlled mesh quality. A decoupled, iteration-free, implicit solution strategy for fluid–structure systems with a few degrees of freedom has also been presented that exploits the underlying linearity of the governing processes to allow superpositions of solutions. These developments set the stage for several forthcoming turbulence simulations of relevance to the transportation and energy sectors and for future FSI simulations in which the structural code is essentially a black-box routine.

Acknowledgements This material was based upon work supported by U.S. Department of Energy, Office of Science, the Office of Advanced Scientific Computing Research, under Contract DE-AC02-06CH11357 . An award of computer time was provided by the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program. This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility. The work of the second author was carried out at the Aerothermochemistry and Combustion Systems Laboratory, LAV-ETH Zurich.

References

1. S. Hosseini, R. Vinuesa1, P. Schlatter, A. Hanifi, D. Henningson, Int. J. of Heat and Fluid Flow (submitted)
2. M. Schmitt, Direct numerical simulations in engine-like geometries. Ph.D. thesis, ETH Zurich (2014). Zurich, CH
3. M. Schmitt, K. Boulouchos, Int. J. of Engine Res. p. 1468087415619289 (2015)
4. M. Schmitt, C. Frouzakis, Y. Wright, A. Tomboulides, K. Boulouchos, Int. J. of Engine Res. **17**(1), 63 (2016)
5. A. Patera, J. Comput. Phys. **54**, 468 (1984)
6. S. Orszag, M. Israeli, M. Deville, J. Sci. Comp. **1**, 75 (1986)
7. Y. Maday, A. Patera, E. Rønquist, J. Sci. Comput. **5**, 263 (1990)
8. A. Tomboulides, M. Israeli, G. Karniadakis, J. Sci. Comput. **4**, 291 (1989)
9. J. Perot, J. Comp. Phys. **108**, 51 (1993)
10. W. Couzy, Spectral element discretization of the unsteady Navier-Stokes equations and its iterative solution on parallel computers. Ph.D. thesis, Swiss Federal Institute of Technology-Lausanne (1995). Thesis nr. 1380
11. P. Fischer, J. Comput. Phys. **133**, 84 (1997)
12. P. Fischer, J. Lottes, in *Domain Decomposition Methods in Science and Engineering Series*, ed. by R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Widlund, J. Xu (Springer, Berlin, 2004)

13. J.W. Lottes, P.F. Fischer, J. Sci. Comput. **24**, 45 (2005)
14. H. Tufo, P. Fischer, J. Parallel Distrib. Comput. **61**, 151 (2001)
15. P. Fischer, J. Lottes, W. Pointer, A. Siegel, J. Phys. Conf. Series **125**, 012076 (2008)
16. J. Lottes, Independent quality measures for symmetric AMG components. Tech. Rep. ANL/MCS-P1820-0111, Argonne National Laboratory, Argonne, IL, USA (2011)
17. J. Boyd, J. Comput. Phys. **143**, 283 (1998)
18. P. Fischer, J. Mullen, Comptes rendus de l'Académie des sciences, Série I- Analyse numérique **332**, 265 (2001)
19. J. Malm, P. Schlatter, P. Fischer, D. Henningson, J. Sci. Comp. **57**, 254 (2013)
20. H. Tufo, P. Fischer, in *Proc. of the ACM/IEEE SC99 Conf. on High Performance Networking and Computing, Gordon Bell Prize* (IEEE Computer Soc., CDROM, 1999)
21. L. Ho, A Legendre spectral element method for simulation of incompressible unsteady viscous free-surface flows. Ph.D. thesis, Massachusetts Institute of Technology (1989). Cambridge, MA.
22. L. Ho, Y. Maday, A. Patera, E. Rønquist, Comput. Methods Appl. Mech. Engrg. **80**, 65 (1990)
23. L. Ho, A. Patera, Comput. Methods Appl. Mech. Engng. **80**, 355 (1990)
24. Y. Maday, A. Patera, in *State-of-the-Art Surveys in Computational Mechanics*, ed. by A. Noor, J. Oden (ASME, New York, 1989), pp. 71–143
25. M. Deville, P. Fischer, E. Mund, *High-Order Methods for Incompressible Fluid Flow* (Cambridge University Press, Cambridge, 2002)
26. P. Fischer, A. Patera, J. Comput. Phys. **92**, 380 (1991)
27. J. Guermond, P. Minev, J. Shen, Comput. Methods Appl. Mech. Engrg. **195**, 6011 (2006)
28. A.G. Tomboulides, J.C.Y. Lee, S.A. Orszag, J. Sci. Comp. **12**, 139 (1997)
29. A. Tomboulides, S. Orszag, J. Comput. Phys. **146**(691–706) (1998)
30. B.T. Chu, X. Kovasznay, J. Fluid Mech. **3**(5), 494 (1958)
31. R.G. Rehm, H.R. Baum, J. Res. Nat. Bur. Stand. **83**(3), 97 (1978)
32. A. Majda, J. Sethian, Combust. Sci. Tech. **42(3–4)**, 185 (1985)
33. G. Byrne, A. Hindmarsh, Int. J. High Perform. Comput. Appl. **13**, 354 (1999)
34. S. Orszag, M. Israeli, M. Deville, J. Sci. Comp. **1**, 75 (1986)
35. J. Donea, A. Huerta, J.P. Ponthot, A. Rodriguez-Ferran, Encyclopedia of computational mechanics **DOI: 10.1002/0470091355.ecm009**, 1:14 (2004)
36. P. Fischer, N. Miller, H. Tufo, in *Parallel Solution of Partial Differential Equations*, ed. by P. Bjørstad, M. Luskin (Springer, Berlin, 2000), pp. 158–180
37. S. Orszag, J. Comput. Phys. **37**, 70 (1980)
38. P. Fischer, Spectral element solution of the navier-stokes equations on high performance distributed-memory parallel processors. Ph.D. thesis, Massachusetts Institute of Technology (1989). Cambridge, MA.
39. D. Giannakis, P. Fischer, R. Rosner, J. Comput. Phys. **228**, 1188 (2009)
40. A. Masud, T.J.R. Hughes, Comput. Methods Appl. Mech. Engrg. **146**, 91 (1997)
41. H. Kanchi, A. Masud, Int. J. Numer. Methods Fluids **54**, 923 (2007)
42. P. Fischer, Comput. Methods Appl. Mech. Engrg. **163**, 193 (1998)
43. J. Hron, S. Turek, *A Monolithic FEM/Multigrid Solver for an ALE Formulation of Fluid-Structure Interaction with Applications in Biomechanics*, *Lecture Notes in Computational Science and Engineering*, vol. 53 (Springer, 2010)
44. J.F. Gerbeau, F. Nobile, P. Causin, Comput. Methods Appl. Mech. Engrg. **194**, 4506 (2005)
45. M. Fernandez, J. Gerbeau, C. Grandmont, Comptes rendus de l'Académie des sciences, Série I- Analyse numérique **342**, 279 (2006)
46. C. Farhat, A. Rallu, K. Wang, T. Belytschko, Int. J. Numer. Methods Eng. **84**, 73 (2010)
47. J. Banks, W.D. Henshaw, B. Sjögreen, J. Comput. Phys. **231(17)**, 5854 (2013)
48. J. Banks, W.D. Henshaw, D.W. Schwendeman, J. Comput. Phys. **269**, 108 (2014)
49. J. Banks, W.D. Henshaw, D.W. Schwendeman, J. Comput. Phys. **268**, 399 (2014)
50. Fernández, M. Landajuela, M. Vidrascu, J. Comput. Phys. **297**, 156 (2015)
51. P. Bearman, J. Fluids and Structures **27**, 648 (2010)

52. R. Tumkur, R. Calderer, A. Masud, A. Pearlstein, L. Bergman, A. Vakakis, J. Fluids and Structures **40**, 214 (2013)
53. P. Fischer, in *22nd AIAA Computational Fluid Dynamics Conference, AIAA Aviation* (AIAA 2015-3049, 2015)
54. O. Walsh, in *The NSE II-Theory and Numerical Methods*, ed. by J. Heywood, K. Masuda, R. Rautmann, V. Solonikkov (Springer, 1992), pp. 306–309
55. T. Bjontegaard, E.M. Rønquist, Comput. Methods Appl. Mech Engng. **197(51)**, 4763–4773 (2008)
56. Kee, R.J., F.M. Rupley, J.A. Miller, M.E. Coltrin, J.F. Grcar, E. Meeks, H.K. Moffat, A.E. Lutz, G. DixonLewis, M.D. Smooke, J. Warnatz, G.H. Evans, R.S. Larson, R.E. Mitchell, L.R. Petzold, W.C. Reynolds, M. Caracotsios, W.E. Stewart, P. Glarborg, C. Wang, , O. Adigun, CHEMKIN collection, Release 3.6. Tech. rep., Reaction Design, Inc., San Diego, CA (2000)
57. T.K. Prasanth, S. Mittal, J. Comput. Phys. **594**, 463 (2008)
58. R. Tumkur, P. Fischer, L. Bergman, A. Vakakis, A. Pearlstein, submitted (2015)

Eight Great Reasons to Do Mathematics

Chris Budd

Abstract In 2012 the UK Government identified eight great technologies which would act as a focus for future scientific research and funding. Other governments have produced similar lists. These vary from Big Data, through Agri-Science to Energy and its Storage. Mathematics lies at the heart of all of these technologies and acts to unify them all. In this paper I will review all of these technologies and look at the math behind each of them. In particular I will look in some detail at the mathematical issues involved in Big Data and energy. Overall I will aim to show that whilst it is very important that abstract mathematics is supported for its own right, the eight great technologies really do offer excellent opportunities for exciting new mathematical research and applications.

1 Introduction

Lets face it, at the moment we still do have a problem with the image of mathematics, which is perceived, widely, to be useless and irrelevant to the modern world. Of course this is very far from the truth, as every (applied) mathematician knows. Indeed mathematics lies at the heart of nearly all of modern technology, as well as much of art and popular culture.

There are spectacular examples of the role played by mathematics and by mathematicians in the developments in technology over the last 150 years. Perhaps the best of these is the discovery of electromagnetic waves by purely mathematical reasoning by Maxwell. It is very hard to think of any modern technology, whether it is a TV, a mobile phone, a SatNav device, a computer or a microwave cooker, which doesn't completely rely on Maxwells fundamental discoveries. There are numerous other examples. Everyone now uses Google to search the Internet, and the algorithm behind this, developed by Brin and Page, relies on finding eigenvectors of (very large) matrices. The Internet itself only works because of a deep understanding of the mathematical behavior of networks and the heavy use of probability theory

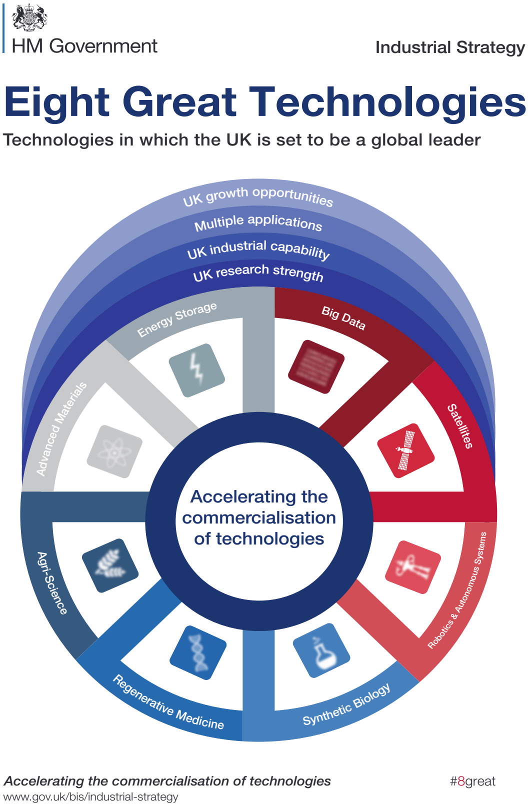
C. Budd (✉)
Department of Mathematics, University of Bath, Bath BA2 7AY, UK
e-mail: mascjb@bath.ac.uk

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_8

in ensuring that information is transmitted reliably over it. Modern medicine has been revolutionized by the use of medical imaging technology, which relies on the mathematical theory of inverse problems, and also on the graphical presentation of medical statistics (to policy makers), which was pioneered by Florence Nightingale. Mathematics is hugely important in the computer graphics and games industry. We all rely on mathematics to keep our financial (and other transactions) secure, and our view of the change in the climate over the next 100 years is, of course, a view informed by careful mathematical models.

So where are we heading next? Mathematics, by its very nature, is boundless in its applications, and should, of course, be pursued as an abstract study for its own sake, and in this way will drive future technology, rather than be driven by it. (Maxwells work is a perfect example of this). However, in the UK, HM Government has itself identified a list of Eight Great Technologies which it sees as the future technologies in which the UK will be a world leader. These were launched in 2012 (although some have been added since) in a speech by the former minister for science The Rt. Hon David Willetts MP. This speech has led to an HM Government Industrial Strategy report and a flurry of activity on many websites. More information on the eight great technologies is given in the report [1]. See also the government publication illustrated in Fig. 1.

Fig. 1 The original government publication on the Eight Great Technologies



HM Government

Industrial Strategy

Eight Great Technologies

Technologies in which the UK is set to be a global leader

UK growth opportunities
Multiple applications
UK industrial capability
UK research strength
Energy Storage
Big Data
Satellites
Robotics & Autonomous Systems
Synthetic Biology
Regenerative Medicine
Agri-Science
Advanced Materials
Accelerating the commercialisation of technologies

Accelerating the commercialisation of technologies
www.gov.uk/bis/industrial-strategy

#8great

It was noticeable that in his speech, in the report, and in the activity it generated, that the role of mathematics was only mentioned briefly, in the context of Big Data, which we will look at presently. This is a symptom of the issues regarding the perception of mathematics that I have highlight above. However, as I will attempt to show in this article, mathematics lies at the heart off all of the eight great technologies and links them all together. Indeed I would argue that they are eight great mathematical technologies. Furthermore, I would expect that the process of getting mathematicians engaged with them, and addressing the huge challenges that they bring, will lead to many breakthroughs in pure mathematics. So, lets go for it. They are truly eight great, but certainly not the only, reasons for doing maths.

2 The Eight Great Technologies

In the original speech in 2012 the Eight Great Technologies were identified as being

1. Big Data
2. Satellites and space
3. Robotics and autonomous systems
4. Synthetic biology
5. Regenerative medicine
6. Agricultural science
7. Advanced materials
8. Energy and its storage.

More recently, quantum based technology has been added to this list, and it is likely to grow further, but I will confine myself to the original list for the purposes of this article. Indeed I cannot in this article do justice to all of these original technologies, so my intention is to say a little about the mathematics in all of them, a bit more about the mathematics associated with Agri-Science and Advanced Materials and, to give some detail about the mathematics behind Big Data and Energy.

The UK list was identified by the *Policy Exchange Think Tank* and the *Technology Strategy Board* in collaboration with research scientists and members of the research funding bodies. A technology made it on the list if:

- It represented an important area of scientific advance
- There was already some existing capacity for it in the UK
- It was likely that new commercial technologies would arise from it
- There was some popular support for it

In an era of austerity and cuts (or at least no increase) in science funding in the UK, the technologies offered the promise of an immediate £600M and then up to £1.5 Bn of new *capital investment*. This is on top of £4.6 Bn baseline science research funding. So it was a substantial commitment of funds.

It is interesting that at about the same time that the eight great technologies were launched in the UK a similar list of ten ‘National Science Challenges’ was launched in May 2013 in New Zealand, with a similar promise of targeted funding. The New Zealand list is interesting both for its similarities, and its differences from, the UK list. For example it is much more health and environment focused. It is as follows [2]

1. Ageing
2. Birth and childhood health
3. Diabetes and cancer
4. Nutrition
5. National Bio-diversity
6. Agriculture/land and water quality
7. Marine resource sustainability
8. Antarctica
9. New technologies
10. Natural disasters

Similar lists, which overlap considerably in content with those from the UK and New Zealand have been compiled in other countries’ government publications as well as in the popular media (such as the MIT Technology Review or the Scientific American).

3 The First Great Technology: Big Data

One of the biggest challenges that we all face is the challenge of big data and we will look at this first, and in some detail. This was rightly put at the top of HM governments list (although it did not appear in the New Zealand list) and it is my firm belief that Big Data impacts hugely on all of the eight great technologies. The reason is simple. We live in the information age, and most of what we do is hugely influenced by our access to massive amounts of data, whether this is through the Internet, on our computers, or on our mobile phones. About 100 years ago when we were transmitting information by Morse code, the transmission rate was 2 bytes per second. This improved with the use of the teleprinter to 10 bytes per second, and then with the modem to 1 kilobyte per second. In contrast, with modern data we are looking at transmission rates of over *1 gigabyte per second*. Similarly early computers (such as the one I used to do my PhD!) had about 1 kilobyte of random access memory (RAM) (with more data having to be stored, unreliably, on magnetic or even paper tape). Whereas a modern lap top has several gigabytes of RAM and up to 1 terabyte of memory. Access to such a large amount of data leads in turn to large technological and ethical problems. Mathematics can help us with the former, and we should all be aware of the latter. So, what does the ‘Challenge of Big Data’ mean? According to a recent UK report [1] it is:

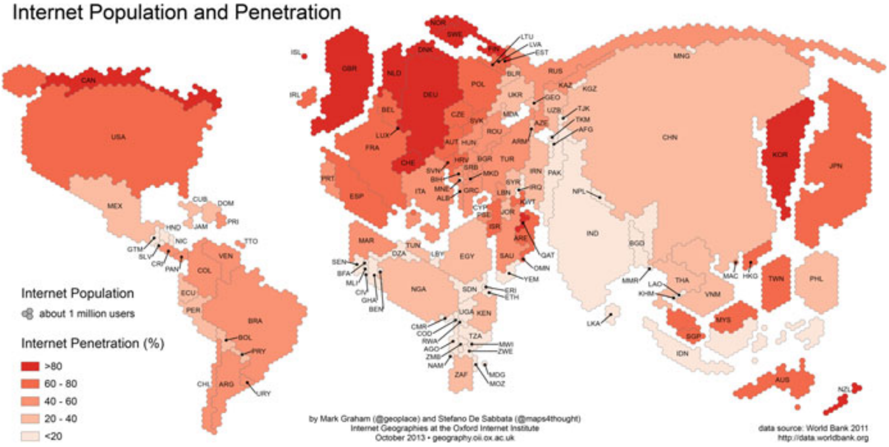
The collection, handling, assurance, curation, analysis and use of:

- Large amounts of existing data using existing methods and technology
- Existing data using new methods and technology
- New data using new methods and technology

The challenge of dealing with such data is always to derive value from large signals, where the useful data may be buried in an avalanche of noise.

3.1 Where Does Big Data Come From?

Perhaps the leading source of current Big Data comes from the *Internet*. According to a recent estimate, about 10^{21} bytes (a zettabyte) of information are added to the Internet every year, much of which is graphical in content. The 'internet penetration' in both the UK and Canada (see Fig. 2) is over 80%, and in all but a few countries is over 20 %. This wealth of data leads in turn to the huge mathematical problems of how we identify, search and organize this information. A major source of this data comes from the ever growing content on Social Media websites. For example, *Facebook* was launched in 2004. It now has 2 Billion registered users (about 1/4 of the world's population!) of what 1.5 Billion are active. Around 2.5 Billion pieces of content (around 500 terabytes of information) are added every day to Facebook sites, with most of this data stored as pictures. The search engine *Google* is estimated to somewhere around 1–15 exabytes (10^{15} bytes) of data (which it searches by using an algorithm based on finding eigenvectors of very large matrices.) Another source of Big Data comes from mobile and smart phones. There are now more mobile phones than people in the world, with the potential for 2.5×10^{19} possible



by Mark Graham (@geoplace) and Stefano De Sabbata (@maps4thought)
Internet Geographies at the Oxford Internet Institute
October 2013 • geography.oii.ox.ac.uk

data source: World Bank 2011
http://data.worldbank.org

Fig. 2 The huge penetration of the Internet

simultaneous conversations. The forthcoming plans for a 5G network will operate on millimetre wavelengths at frequencies as around 70GHz (and are already being piloted in my home town of Bristol, UK). This will offer data rates at 1 gigabyte per second offered simultaneously to tens of workers on the same office floor and with several hundreds of thousands of simultaneous connections to be supported for massive sensor deployments. Such sensors can provide constant monitoring of, say, our state of health, with significant ethical implications. Indeed the future is rapidly approaching (such as the *Internet of things*) in which our devices simply communicate with each other (for example the cooker talks to the dishwasher and also to the supermarket every time a meal is prepared) with little or no human interference.

As well as the devices above, significant amounts of data, of significant interest to the social sciences, comes from the way that we use them and the information that it gives about our lifestyles. Again there are significant ethical issues here. Every time that we make a purchase with Amazon, use our bank on-line, switch on an electrical device, or simply use a mobile phone or write an email, we are creating data which contains information which can in principle be analysed. For example our shopping habits can be determined, or our location tracked and recorded. Mathematics can be used at all stages of this, but we must never lose sight of the moral dimension in so doing.

3.2 The Nature of Big Data

In one sense, Big Data has been the subject of mathematical investigation for at least 100 years. Any mathematical model described by a partial differential equation with an infinite number of degrees of freedom, naturally leads to a source of a large amount of data. A classical example of this is meteorology, in which the current meteorological models (typically based on extensions of the Navier-Stokes equations) are solved on super computers with discretisations with 10^9 degrees of freedom informed (in a typical 6 hour forecast window) by 10^6 observations of the state of the atmosphere and oceans. Similar large data sets arise in climate models, geophysics and astronomy. However, the data in these problems, whilst very large, is also well structured and well understood (with known levels of uncertainty), as befits its origins in the physical sciences for which we have good and well understood mathematical models. The real challenges of understanding and dealing with Big Data do not come from these data sets, however large they may be. In contrast the real difficulties arise from data which has its origins (as described above), in the biological sciences, the social sciences and in particular in people based activity. Such data is **Challenging** in that it is: garbled, partial, unreliable, complex, soft, fast arriving, and (of course) big. It is also **Novel** and very different from much of the data arising from physical models in that it is: heterogeneous, qualitative, relational, and partial.

3.3 What Questions Do We Want to Ask of Big Data?

The novel aspects of Big Data lead in turn to challenges in how we deal with it, indeed how we visualise it, make speculations from it, model it, understand it, experiment on those systems which generate it, and ultimately how we might control those systems. The mathematical and scientific challenges behind these questions are as varied as they are important, and the very scale of big data makes automation necessary and this, in turn, necessarily relies on mathematical algorithms.

As examples of such questions we can include the following:

- Ranking information from vast networks in web browsers such as Google
- Identifying consumer preferences, loyalty or even sentiment and making personalised recommendations
- Modelling uncertainties in health trends for individual patients
- Monitoring health in real time (especially in the environment that 5G will lead to)
- Using smart data gathered from energy usage to optimise the way that the energy is then supplied to consumers.

3.4 The Mathematics of Big Data

It is fair, I think, to say, that many of the future advances in modern mathematics (together with theoretical computer science) will either be stimulated by the applications of Big Data or driven by the need to understand Big Data. Of course many existing mathematical techniques (some of which until recently were considered as ‘pure mathematics’) are now finding significant applications in our understanding of Big Data. A key example of this is the mathematics of network theory. This describes objects, described by nodes, and the connections between them, described by edges. Network theory explains the connections between the objects (often formulated through an adjacency matrix), allows us to search the network for connections between the data (by finding structures in the adjacency matrix), and can describe (via differential equations) the movement of information around the network itself. As an example the nodes could be computers or website on the computes, and the edges, connections between the computers or links between the websites. The nodes can be people and the connections to their friends on Facebook or Twitter, or they could be mobile hand sets and the link a conversation or simply a close proximity which might lead to interference. This issue is particularly important, as with 7 billion people in the world, there are a potential of 2×10^{19} conversations over a mobile phone network, each of which must not interfere with any other. Indeed, managing the mobile phone network (which is of course also hugely used to download data) is a significant and growing application of the theory of graph colouring which until recently was regarded as firmly in the domain of pure mathematics. Other examples of networks which lead to big data include: **social networks**: Friendship, sexual partners, Facebook and other social media, **organisational networks**: Management, crime syndicates, Eurovision, **technological networks**: World-wide-web, Internet, the power grid, electronic circuits, **information networks**: DNA, Protein-Protein interactions, citations, word-of-mouth, myths and rumours, **transport networks**: Airlines, food logistics, underground and overground rail systems, **ecological networks**: Food chains, diseases and infection mechanisms. For many more examples see the review article by Newman [3].

Network theory can be used to address more of the many questions related to Big Data as described above. Specifically network theory based algorithms can be used to segment data and find clusterings in data. Such information is vital in data mining and pattern recognition, and is especially important to the retail industry, segmenting graphs (which can include images) into meaningful communities, finding friendship groupings, investigating the organisation of the brain, and even finding Eurovision voting patterns. These voting patterns are illustrated in Fig. 3. A careful analysis of the network illustrated by this figure shows that the rumours of voting blocks really are true! [4]

Such analysis can also help with the very significant problem encountered in many applications of linking databases with different levels of granularity in space and time Equally important is the question of how connected the network is, and what is the shortest length ℓ of a path through the network. This is essential for efficient routing in the Internet, interpretation of logistic data, speed of word of mouth communications and marketing. Network theory is also essential in searching for influential nodes in huge networks (of huge importance to search engines), and in finding the resilience of a network which can be used to break a terrorist organisation, or to stop an epidemic.

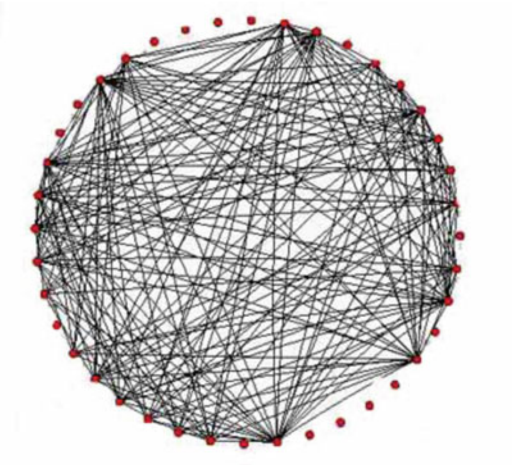


Fig. 3 A network showing who voted for who in the Eurovision Song Contest. Can you spot any patterns?

Of course, network theory, whilst important, is just one of a variety of **mathematical techniques used to study Big Data**. As much of Big Data takes the form of **images**, mathematical algorithms which classify, interpret, analyse and compress images are extremely important in all Big Data studies. Linear signal processing, and related statistical methods have long been used to analyse and interpret images. But there has recently been a significant growth in novel mathematical algorithms, drawing on ideas in 'pure mathematics'. Some of these algorithms, particularly those for image segmentation or denoising, are based on the analysis of nonlinear partial differential equations, leading to some powerful and unexpected applications of such areas of analysis as the p-Laplacian [6]. Algebraic topology plays a very useful role in classifying images, and in particular *persistent homology* [7] can be used to find means of classifying objects in and image which do not depend upon the orientation of the object and is a method for computing topological features of an object at different spatial scales. Cohomology and tropical geometry, in particular combinatorial skeleta allow for a different for of object classification. Finally, techniques from category theory can be used to 'parse' an image to see how the various components fit together, and also (in the context of machine learning) to allow for machines to 'perceive' what the objects are in an image and to make 'reasoned' decisions about it.

A recent and exciting development in the mathematical analysis of Big Data, due to Emmanuel Candès, Justin Romberg, Terence Tao and David Donoho [5] is the area of *compressed sensing*. Traditional signal processing has used Fourier or wavelet based methods to represent data, and compression is then achieved by a suitable truncation of this representation. In contrast, compressed sensing aims to exploit sparsity in the data and to achieve compression by direct sampling. (One mechanism for doing this is to use more 'blocky' representations of figures using piecewise constant representations, achievable through techniques using L_1 or TVD optimisation of the figures.) Compressed sensing is finding very important applications including in the representation of large data sets arising in medical applications.

Big Data is of course also a significant driver for advances in **computer science**, and the development of novel computing algorithms. These include encrypted computation, (which relies heavily on results in number theory), quantum annealing and quantum algorithmics. This is only a short list. Other areas of mathematics and computer science which have found applications in the study of Big Data include: segmentation clustering, optimal and dynamic sampling, uncertainty modelling and generalised error bounds, trend tracking and novelty detection, context awareness, integration of multi-scale models, real-time forecasting, data integrity and provenance methods, visualization methods, data compression and visualisation, dimension reduction, machine learning, logic and reasoning, and optimisation and decision.

Essentially, watch this space! I am confident that we will see great advances in pure, applied and computational maths arising from these challenges.

3.5 The UK Response

The UK government has responded positively to the importance of funding mathematically focused research into Big Data. To this end the Engineering and Physical Sciences Research Council (EPSRC) (which is the rough equivalent of NSERC or NSF) has committed around £40 Million to found the *Alan Turing Institute* (ATI). This is a collaboration between the founding partner universities of Oxford, Cambridge, Edinburgh, Warwick and University College London (UCL) together with non-academic partners including GCHQ (the UK equivalent of the NSA) with Andrew Blake from Microsoft Research as the first director. The site of the ATI will be the British Library close to St. Pancras station in North London. See the website https://turing.ac.uk/ for more details. According to this website

> The work of the Alan Turing Institute will enable knowledge and predictions to be extracted from large-scale and diverse digital data. It will bring together the best people, organisations and technologies in data science for the development of foundational theory, methodologies and algorithms. These will inform scientific and technological discoveries, create new business opportunities, accelerate solutions to global challenges, inform policy-making, and improve the environment, health and infrastructure of the world in an Age of Algorithms.

I expect to see similar developments in many other countries in the near future.

4 Satellites and Space, Robotics and Autonomous Systems, Synthetic Biology, Regenerative Medicine

The methods for working with Big Data have natural applications in the **second** great technology: *satellite and space technology*. In fact one of the big early success stories in the data revolution was the use of error correcting codes in the 1970's to transmit the images from distant planets back to the Earth without error. Such codes are usually based on finite function fields, again an area previously thought of as pure mathematics. With satellites playing possibly the major role in transmitting more and more information, the need for evermore sophisticated mathematical algorithms to keep this information accurate and secure, will continue to drive mathematical developments in algebra and discrete systems. The mathematical theory of systems of symplectic Hamiltonian ordinary differential equation, posed on Lie Groups, is also finding a major application in helping to understand and control the dynamics of satellite systems. As such calculations have to be done over long time periods, the numerical methods to approximate the solution of such systems have to be very carefully designed. There is currently a great interest in *geometric integration methods* which combine numerical analysis with differential geometry to make such calculations accurate and reliable [8]. Similar numerical methods are also used to simulate the movement and control of the *robotic systems* which form the **third** of the great technologies. Other applications of mathematics to robotics include Bayesian machine learning algorithms (which also can make use of category theory), pattern recognition techniques (which link again to Big Data), neural networks and computer vision. There is also a natural between Big Data and the **fourth** great technology of *genomics and synthetic biology*. In particular, this technology relies on understanding how genes and proteins interact and this can be studied by using gene/protein networks, where the edges of the network describe allele combinations that control specific phenotypes. The **fifth** great technology of regenerative medicine involves those aspects of tissue engineering and molecular biology which deals with the "process of replacing, engineering or regenerating human cells, tissues or organs to restore or establish normal function" [10]. This involves mathematical modelling, and especially the development of novel flexible materials (see also the seventh great technology). Of course mathematics has many other applications to medicine including medical statistics (with huge relevance to Big Data), modelling, and curing, cancer and in the various inverse problems arising in medical imaging.

5 The Sixth Great Technology: Agri-Science

Food and beverage processing is the world's largest manufacturing industry and a recent UN Forecast has stated that if the population continues to rise at its present rate, then the world food output must increase by 70% by 2050. Achieving this output is significant challenge to agriculture and to the science behind agriculture. Mathematics plays an important part of this, with many existing, and potential, applications in agri-science and food technology. The fundamental process of growing (including irrigation), freezing, cold storing, cooking, making, eating and even digesting food are all areas in which the application of areas of mathematics such as thermodynamics, (non-Newtonian) fluid mechanics, and partial differential equation theory, can make a very big difference to food safety and production. As an example the partial differential equation

$$H_t = \kappa \nabla^2 T + \mu_0 Q e^{-x/d} \tag{1}$$

where H is the enthalphy of a food product, x the distance into the food, T its temperature, Q the strength of a microwave field and μ_0 the dielectric permittivity of the food, can be used to predict the temperature of a moist food stuff when heated in a microwave oven. This in turn can be used to predict the safety of the microwave cooking process. Mathematics can even be used to simulate the production of such an iconic Canadian product as Maple Syrup, see [12], in which partial differential equations are used to model the movement of sap in a Maple Tree. Similarly, the logistics of feeding a growing world population, which requires food to be packaged, transported and disposed of safely and efficiently, requires the mathematics of optimization and operational research (and network theory and indeed Big Data again). Another major challenge to agri-science is the future of the bee population. This is in a state of significant decline, and if the bees were to vanish then much of modern agriculture would not be possible. Mathematics can help in the challenge of saving the bees by providing a technology by which bee populations in a hive can be monitored in a non invasive or harmful manner. This is achieved by using tomographic X-Ray imaging, where the shadows cast by X-rays are used to look inside the beehive. Diagnostic radioentomology is a technique, developed primarily by Mark Greco, that takes entomological studies and combines them with medical diagnostic methodologies [13]. This imaging technique is sufficiently sensitive to resolve individual bees and to see how they are responding to changes in their environment. This technology for resolving the bees is similar to that used earlier in the CAT scanners used in medical imaging before MRI scanners replaced them. However there are important differences. Most notably, the X-ray dosage for bees has to be very low indeed to avoid injuring them, also bees have a splendid habit of moving around whilst they are being scanned. This reduction in dosage, and also an improvement in the image quality (at the expense of greater computational complexity but with increased safety for the bees) can be achieved by using the compressed sensing techniques described above. An image of the bees can be seen in Fig. 4.

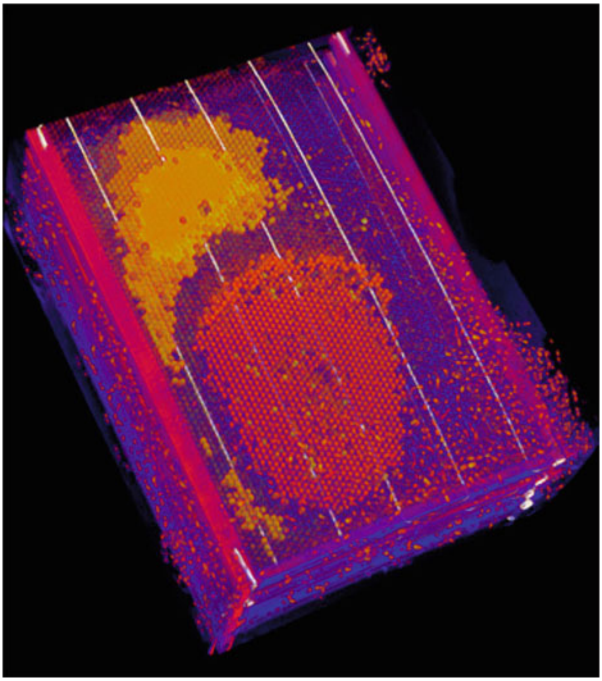


Fig. 4 An image taken from the gallery in [14], showing a tomographic reconstruction of a beehive, in which the bees are seen as red dots and honey as yellow

6 The Seventh Great Technology: Advanced Materials

We all rely on materials, some of which are natural like wood and stone, and others are manufactured such as steel, glass and concrete. However, with modern technology, we can now design and manufacture meta-materials with a wide variety of prescribed mechanical, electrical, thermal and other properties. Such modern advanced materials are often composites of different materials with very different properties, which are combined in a complex manner. The resulting behavior of the composite material often then emerges from the way that these different properties interact, in a manner which is often very different from the sum of the different parts. Some examples of such modern materials are the photonic crystals which are used to transmit light with almost zero loss, the complex composites used in aircraft wings, liquid crystals which are used in many displays, and perhaps most intriguingly, the possibility of materials which, in a manner inspired by Harry Potter, confer invisibility on the user [9]. The mathematics needed to design and study such materials is particularly rich and challenging. At its heart is multi-scale analysis and homogenization, although it also uses ideas from complexity theory, advanced theory from the calculus of variations [11] and (again) network analysis. This mathematics can be used to study materials as ancient as rock, or as modern as carbon fibres. It is no coincidence that major growth in modern applied mathematics is in exactly these areas, and I anticipate that we will see even more in the future (Fig. 5).

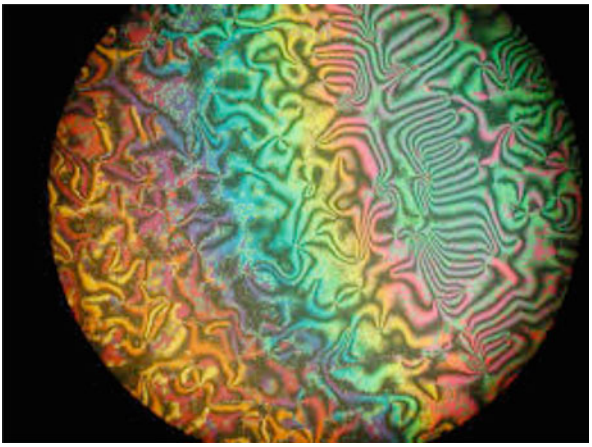


Fig. 5 Modelling the complex patterns in liquid crystals requires advanced mathematical methods from the calculus of variations

7 The Eighth Great Technology: Energy and Its Storage

One of the more interesting jobs I have had was as the CEGB research fellow, which was a joint position between the University of Oxford and the (old) Central Electricity Generating Board (now privatized and split up into many different companies). This job gave me an a good appreciation of the issues faced by the power generating industry and the importance of mathematics at many levels of this industry. The job of the electricity supply industry is to supply electricity reliably to us, regardless of the demand. Or more simply put, to keep the lights on. This is not always easy, as too much demand and an insufficient supply could potentially result in a power cut. As an example, in the World Cup on the 4th July 1990, England was playing in the semi-final against West Germany. This was an exciting and emotional match which England finally lost on penalties. At the end of the match the power demand surged by 2.8 GW (about 1 million electric kettles), or 11 % of the total network demand. Due to good planning by the UK National Grid operators, the lights stayed on in this case. But it was a relatively close run thing. In contrast, in 2003, following an isolated failure, there was a cascade of control system failures of the NE US power grid. In this cascade, when one line was shut down to deal with the failure, this lead to an overload of other nearby lines, which then also shut down. More and more lines then shut down, leading to a large system failure in which the lights *did* go out. The resulting US NE Coast Blackout is estimated to have cost 5 Billion Dollars and can be seen in Fig. 6.

The annual consumption of electricity in the UK is 300 TWh, and this electrical power is supplied over a complex network starting, usually, with power being generated at a power station. This is then transmitted over a high voltage network, before being reduced in voltage and distributed to commercial, industrial and residential consumers. To ensure that the lights always stay on, the planners need to solve a large number of nonlinear differential-algebraic equations, described on (another!) complex network (with 30 million nodes representing different households, industries and other users of electricity), to work out how much

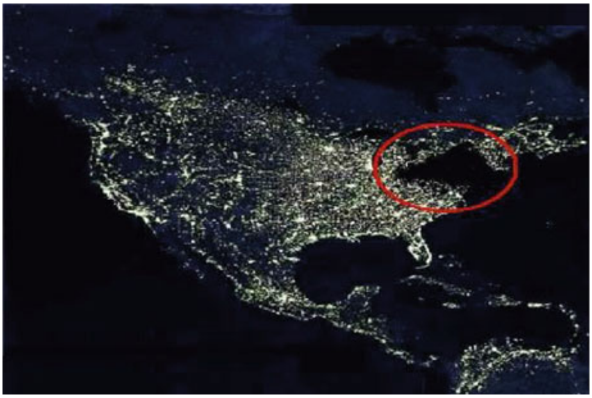


Fig. 6 The day the lights did go out in NE America in 2003. The circle shows the blackout region. Image taken from [17]

electricity can be generated, distributed and stored. However this is not easy as electricity must be consumed as soon as it is purchased, it cannot be stored in large quantities and the user has a very low tolerance to interruptions in the supply. These challenges are going to increase significantly in the future with a greater emphasis on low Carbon generation, a much more distributed supply network (with a significant increase in lower power generation from renewable sources such as solar and wind often at a domestic level), the increase in the use of electric vehicles, an increase in local electricity storage, and the advent of the SMART Grid in which users both have much greater control over their energy demands and also supply much more information to the Grid company [16] (another example of a Big Data problem). These are all challenges which mathematicians are well placed to address.

As an example, in an AC supply network, the steady state form of the power flow equations typically take the form of a large number of quadratic equations with complex coefficients (yes quadratic equations really are useful!) derived from Kirchoffs laws, defined over a network in which the nodes are the power stations and the users (eg. households) and the links are the electricity cables and transformers connecting them together. Typically in such a network the AC voltage at a node j is represented by a (complex) *voltage* V_j with an associated *phase* δ_j and current I_j. The AC power at this node is then given by

$$P_j + iQ_j = V_j I_j^*. \tag{2}$$

Here P_j and Q_j are respectively the real and the reactive power. Typically these are known at each node and represent the local demand or supply of power at the node. (Both of these aspects could involve issues associated with energy storage). The nodes at j and k are typically connected by bus bars with complex *admittance*

$$Y_{jk} = G_{jk} + iB_{jk}. \tag{3}$$

In a power cable, losses are very low and we expect G_{jk} to be close to zero. The current flow I_{jk} through the bus is then given by Ohm's law so that

$$I_{jk} = Y_{jk}\left(V_j - V_k\right). \tag{4}$$

Over all of the nodes we must satisfy the power demand given above. If δ_{ij} is the *phase difference* between the voltage at nodes j and k then we can apply Kirchoff's laws to give the following parametrised set of quadratic equations for the *steady state* [15]

$$\sum_{k=1}^{N} |V_j||V_k| \left(G_{jk}\sin\left(\delta_{jk}\right) - B_{jk}\cos\left(\delta_{jk}\right)\right) = Q_j. \tag{5}$$

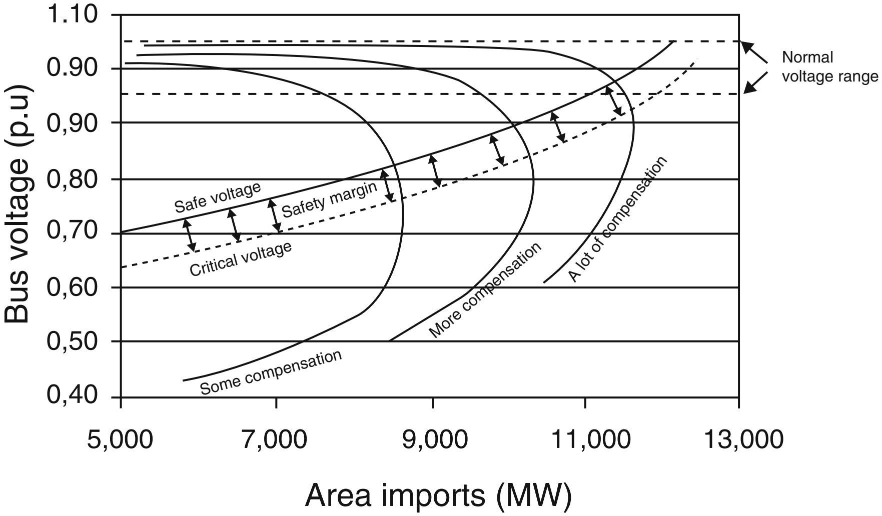


Fig. 7 A typical nose curve showing the voltage response to different levels of power demand with (from left to right) increasing levels of available compensating power supplied to the grid

and

$$\sum_{k=1}^{N} |V_j||V_k| \left(G_{jk} \cos\left(\delta_{jk}\right) + B_{jk} \sin\left(\delta_{jk}\right)\right) = P_j. \tag{6}$$

Here N is the number of nodes, which may be very large (of the order of many millions). However, we expect to see a degree of sparsity in this system due to a limit to the number of nodes connected to each other by bus conductors, which can help with the analysis. Companies managing the grid have to solve such systems rapidly to cope with changing patterns for supply and demand. As with all parametrised quadratic equations we expect to see the number of solutions of the system changing as the parameters vary. In particular, steady solutions can vanish at *fold bifurcations* as the demanded load increase. An example of such a fold bifurcation, usually called a *nose curve*, is given in Fig. 7. If a fold bifurcation happens in practice, and the demand exceeds the maximum permitted value, then the voltage in the network *collapses to zero* and we have a national power cut [15]. Power cuts due to a voltage collapse have occurred in both Italy and Sweden. A lot of interesting mathematics including advanced (hierarchical) linear algebra, nonlinear systems theory coupled to networks, and algebraic geometry, needs to be developed in order to predict these bifurcation points and to identify well in advance whether a power failure is likely to occur and how this can be (efficiently) avoided by bringing more power stations on-line. This is all made much more complicated by constraints arising from different pricing policies for networks and the vagueries of demand due to individual preferences [16].

Equally important in the modern era of a rapidly evolving network, informed by huge amounts of data, are problems associated with understanding the *dynamics of the grid*. This is of special significance when considering the question of energy storage, explicitly mentioned in the Eight Great Technologies. Electricity is often stored in batteries and hence as DC. This needs to be converted to AC through an *inverter* and then matched (in phase) to the overall network. Inverters have complex dynamics (linked to a phase locked loop) and can introduce instabilities into the overall grid, and hence the possibility of problems with the electricity supply. Other dynamic terms need to be added to the network to account for the behaviour of the turbines in power stations in response to load, the empirically observed behaviour of electrical devices such as motors and cities, the rapidly changing output of renewable energy generation, the smart grid and (in the near future) for large scale electrical vehicle charging. Any dynamical system set up to model the system will have to take into account for the many different time-scales that the grid operates under, from the near instantaneous times of electricity transport, to the times needed to operate switches, the daily cycle of human activity, through to seasonal trends in electrical usage. To do such a study effectively will naturally lead to much new mathematics including the analysis of (bifurcations in) multi-scale and/or non-smooth stochastic dynamical systems on large networks, a subject close to my heart, and currently in its infancy. See [18] for a review of the current state of the art in this emerging area.

8 Conclusions

I hope that I have whetted your appetite. The eight great technologies certainly present enormous opportunities for mathematicians in the next 50 years and beyond. I encourage all mathematicians regardless of which country they live in, to rise to the challenges presented by these **eight great reasons to do mathematics**.

But, the moral of this article, is that whilst lots of new mathematics is needed for all the eight great technologies, to keep the lights on you must be especially good at solving quadratic equations!

References

1. D. Willetts, *Eight Great Technologies*, (2013), Policy Exchange. http://www.policyexchange.org.uk/images/publications/eight%20great%20technologies.pdf
2. New Zealand Ministry of business, innovation and employment, *National Science Challenges*, (February, 2016) http://www.mbie.govt.nz/info-services/science-innovation/national-science-challenges/about
3. M. Newman, *The structure and function of complex networks*, (2003), SIAM Review **45**, 167–256

4. D. Fenn, O. Suleman, J. Efstathiou, N.F. Johnson, *How does Europe Make Its Mind Up? Connections, cliques, and compatibility between countries in the Eurovision Song Contest*, (2006), Physica A: Statistical Mechanics and its Applications, **360**, 576–598.
5. E.J. Candes, J.K. Romberg, T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, (2006), Communications on Pure and Applied Mathematics **59**, 1207–1223.
6. A. Kuijper, *Image analysis using p-Laplacian and geometrical pdes*, (2007), Proc. Appl. Math. Mech., **7**, 1011201–1011202.
7. H. Edelsbrunner, *Persistent homology in image processing*, (2013), in: Graph-based representations in pattern recognition, Proceedings of the 9th IAPR-TC-15 International Workshop, GbRPR 2013, Vienna, Austria, May 15–17, 2013., Springer.
8. E. Hairer, G. Wanner, C. Lubich, *Geometric Numerical Integration Structure-Preserving Algorithms for Ordinary Differential Equations*, (2006), Springer Series in Computational Mathematics **31**, ISBN: 978-3-540-30663-4.
9. A. Greenleaf, Y. Kuryalev, M. Lassas, G. Uhlman, *Inverse problems and invisibility*, (2009), Bulletin AMS **46**, 55–97.
10. C. Mason and P. Dunnill, *A brief definition of regenerative medicine*. Regenerative Medicine, (2008), **3**, 1–5
11. H. Kusumaatmaja, A. Majumdar, *Free energy pathways of a multistable liquid crystal device*. (2015), Soft Matter, **11**, 4809–4817.
12. I. Graf, M. Ceseri, J.M. Stockie, *Multiscale model of a freeze–thaw process for tree sap exudation*, (2015), Journal of the Royal Society INterface, **12**, DOI: 10.1098/rsif.2015.0665.
13. M. Greco, *Imaging Techniques for Improved Bee Management*, (2010). Agroscope Liebefeld-Posieux, ALP, CH-3003 Berne.
14. M. Greco, *Diagnostic Radioentomology*, (2013) http://www.radioentomology.com
15. Y.V. Makarov, D.J. Hill and I.A. Hiskens, *Properties of quadratic equations and their application to power system analysis*, (2000), International Journal of Electrical Power and Energy Systems, **22**, 313–323.
16. V. Hamidi, F. Li, F. Robinson, *Demand response in the UK's domestic sector*, (2009), Electric Power Systems Research, **79**, 1722–1726.
17. E. Vielmetti, *Northeast Halloween Snowstorm. Power Outage Maps*, (2011) https://datacenterpro.wordpress.com/2011/10/31/northeast-halloween-snowstorm-power-outage-maps/
18. M. di Bernardo, C. Budd, A. Champneys and P. Kowalczyk, *Piecewise-smooth dynamical systems: theory and applications*, (2009), Springer

Part VI
Finance and Systemic Risk

Calculation of Exposure Profiles and Sensitivities of Options under the Heston and the Heston Hull-White Models

Q. Feng and C.W. Oosterlee

Abstract Credit Valuation Adjustment (CVA) has become an important field as its calculation is required in Basel III, issued in 2010, in the wake of the credit crisis. *Exposure*, which is defined as the potential future loss on a financial contract due to a default event, is one of the key elements for calculating CVA. This paper provides a backward dynamics framework for assessing exposure profiles of European, Bermudan and barrier options under the Heston and Heston Hull-White asset dynamics. We discuss the potential of the *Stochastic Grid Bundling Method* (SGBM), which is based on the techniques of *simulation*, *regression* and *bundling* (Jain and Oosterlee, Applied Mathematics and Computation, 269:412–431, 2015). By SGBM we can relatively easily compute the Potential Future Exposure (PFE) and sensitivities over the whole time horizon. Assuming independence between the default event and exposure profiles, we give here examples of calculating exposure, CVA and sensitivities for Bermudan and barrier options.

1 Introduction

In the wake of the credit crisis, regulators put more strict capital requirements to cover losses caused by default events. A recent capital charge was introduced in Basel III, i.e. the Credit Value Adjustment (CVA). CVA is the difference between the risk-free contract value and the contract value that accounts the possibility of a counterparty's default [16]. It can be computed as the integral over the time horizon as the expectation of the discounted losses on a default event, multiplied by the probability of default at that moment and the percentage of loss given default [30]. The computational complexity of CVA arises from the uncertainties of the losses

Q. Feng (✉)
Center for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands
e-mail: qian@cwi.nl

C.W. Oosterlee
CWI and Delft University of Technology, Delft, The Netherlands
e-mail: c.w.oosterlee@cwi.nl

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_9

of a default event and the likelihood of the counterparty's default in the future. An unstable dependence structure between the counterparty's default probability and the corresponding losses in the future may exist, which makes the computation of CVA complicated [5, 16].

Credit *exposure* is defined as potential future losses without any recovery. Exposure evolves over time as the market moves with volatility, and typically cannot be expressed in closed form. Before the appearance in Basel II [2], concepts as expected exposure (EE) and potential future exposure (PFE) had emerged and were commonly used as the representative metrics for credit exposure [16]. EE represents the *average* expected loss in the future, while PFE can manifest the *worst* exposure given a certain confidence level. These two quantities illustrate the loss from both a pricing and risk management perspective [16], respectively. In order to get these metrics of future losses, in practice, the exposure profile needs to be computed for a large number of scenarios on a set of time steps. This is one of the involved parts in computing CVA.

A general Monte Carlo (MC) framework is formulated by Pykhtin and Zhu [30] for the computation of exposure profiles for over-the-counter (OTC) derivative products. There are three basic components: (1) Monte Carlo path generation for a series of simulation dates under some underlying dynamics; (2) valuation of mark-to-market (MtM) values of the contract for each realization at each simulation date, by some numerical method; (3) calculation of exposure for each simulation at each simulation date.

Calculation of exposure profiles asks for efficient numerical methods, as the computational demand grows rapidly w.r.t. the number of MC paths. Different numerical methods have been combined with the MC forward paths to handle the computational demand of exposure, such as the Finite Difference Monte Carlo Method [12] or the Monte Carlo COS method[1] [31]. Computational complexity increases for CVA of a whole portfolio, as there are then multiple financial derivatives in the exposed portfolio. Inclusion of various market factors in the asset dynamics, such as stochastic asset volatility and stochastic interest rates, further increases the computational effort.

We will use the Stochastic Grid Bundling Method (SGBM) for the efficient and flexible computation of exposure. The SGBM technique was proposed for pricing multiple-asset Bermudan derivative contracts under Black-Scholes dynamics in [20]. In the present work we extend SGBM to computing exposure values of options under a stochastic volatility asset equity model with stochastic interest rates. We show the impact of adding stochastic volatility and stochastic interest rates on the metrics of future losses (i.e. CVA, EE, PFE). A stochastic volatility may explain the implied volatility surface observed in the derivatives market (such as the volatility smile) [18], and uncertainty in the interest rate may give a significant contribution to the price, especially of long-term financial derivatives [25]. The

[1] The COS method is an option pricing method for European/Bermudan options based on the Fourier-cosine series developed first by F.Fang and C.W. Oosterlee.

hybrid model chosen to model these stochastic quantities is the Heston Hull-White model [17]. We will also study the impact of stochastic interest rate and stochastic volatility, respectively, under the Black-Scholes Hull-White model and the Heston model.

SGBM is based on *simulation*, *regression*, and *bundling* [20], and the method is very suitable for the computation of exposure profiles. The idea of using simulation and regression for pricing options with early exercise has been used by Carriere [9], Tsitsiklis and Van Roy [33], and Longstaff and Schwartz [27]. There are several recent modifications and comparisons of pricing techniques with regression, such as the work by Broadie and Cao [7], by Broadie et al. [8] and by Stentoft [32]. SGBM distinguishes itself from other regression-based simulation methods in the following ways. First of all, a bundling technique is employed to ensure an accurate *local* calculation of the exposures on each path. Secondly, the conditional expectations of basis functions used for regression in SGBM are analytic expressions when the underlying framework is affine or can be approximated by an affine model. They are used for the calculation of the continuation values. Thirdly, compared to the popular Longstaff-Schwartz (LS) method that uses the 'in-the-money' paths, SGBM uses the information of all paths and assigns exposure values to each path at each monitoring date. These features ensure the accuracy of computing exposure values on each path, which is in particular important for PFE. Furthermore, sensitivities of the EE can be calculated accurately with little extra effort.

The flexibility of SGBM is demonstrated by placing the computation of exposure profiles, for different option types under different asset dynamics, in a general unifying framework based on backward recursion. The options considered include European, Bermudan and barrier options. The reminder of the paper is structured as follows: Sect. 2 provides the mathematical framework for CVA and exposure, discusses the affine diffusion models for the underlying, and the backward dynamics for calculation of the exposure of options, and their exposure sensitivities. In Sect. 3, we present the SGBM algorithm in detail. In Sect. 4 the choice of basis functions and the derivation of the discounted moments is presented, as well as a simple bundling technique that ensures the accuracy of the local, bundle-wise, regression. In Sect. 5, numerical results are presented to show the convergence and efficiency of the method, and the impact of the stochastic interest rate and stochastic volatility on the exposure metrics is studied in Sect. 5.4.

2 CVA and Exposure

CVA is the price of counterparty-credit risk. It is based on an expected value (the expected exposure) which is computed under the risk-neutral measure. There has been debate on the computation of PFE regarding whether to compute it under the real-world or the risk-neutral measure. It is argued that PFE should be computed based on simulations under the real-world measure, reflecting the future developments in the market realistically, from a risk management perspective [22].

In this paper we will focus on the computation of CVA, and we will compute EE and PFE under the risk-neutral measure as well. However, the numerical techniques in this paper can be also be used for computing PFE under the real-world measure, which is a next stage of our work.

The default probability will also be measured under the risk-neutral measure in this paper. The implied default probability of the counterparty typically is retrieved from market prices of CDS (credit default swap) or corporate bonds issued by this counterparty. Notice that the implied default probability under the risk-neutral measure in general is different from that inferred from historical data under the real-world measure, and of the two the former is typically higher than the latter [5].

2.1 Mathematical Formulation

Assuming a market without friction. Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a complete probability space on a finite time horizon $[0, T]$ including all required quantities, where Ω is the sample space, $\mathscr{F}$ is the sigma algebra of events at time T, and $\mathbb{P}$ is the probability measure. We assume the existence of a risk-neutral probability measure $\mathbb{Q}$, equivalent to $\mathbb{P}$, under which the current value of a financial asset is equal to its expected discounted payoff in the future. The uncertainty of the market includes a set of influencing factors, such as the (log-)stock price and its volatility, and the short rate. These quantities can all be expressed by an n-dimensional Markov process $(\mathbf{Y}_t)_{t\in[0,T]}$, $\mathbf{Y}_t = [Y_t^1, Y_t^2, \dots, Y_t^n]$ in some space $U \subset \mathbb{R}^n$. The natural filtration $(\mathscr{F}_t)_{t\in[0,T]}$ on the probability space is the sigma algebra associated to $\mathbf{Y}_t$, and $\mathscr{F}_t$ includes all information about the market up to time t. We further suppose the existence of a risk-free asset, $B(t) = \exp\left(\int_0^t r_u du\right)$, where $r_t = r(\mathbf{Y}_t)$ is the short rate at time t. The associated stochastic discounting factor in the period $[t, s]$ is defined as $D(t, s) := \exp\left(-\int_t^s r_u du\right)$. The value of a default-free zero coupon bond (ZCB) at time t with maturity T is given by $p(t, T) := \mathbb{E}^{\mathbb{Q}}\left[D(t, T)\middle|\mathscr{F}_t\right]$.

We will study the exposure for investors towards option writers. Particularly, we will compute exposure profiles of OTC Bermudan, European and barrier options, for which the contract values of the options at time t are only determined by the variable $\mathbf{Y}_t$, i.e. the option values can be regarded as functions $V(t, \mathbf{Y}_t) : [0, T] \times U \to \mathbb{R}$. The exposure can also be measured in terms of the *replacement costs* for a derivative contract, i.e. the amount to replace the contract at current market rates [16]. Without transaction costs, the exposure of options in a default event is the loss defined by the replacement costs without any recovery. We assume that the exposure to the writer immediately becomes zero when the option is terminated, exercised or knocked out. Hence, exposure can be expressed by:

$$E(t, \mathbf{Y}_t) = \begin{cases} 0, & \text{if the option is terminated, knocked out or exercised,} \\ V(t, \mathbf{Y}_t), & \text{if the option is alive.} \end{cases} \tag{1}$$

In addition, the *discounted exposure* is defined by $E^*(t, \mathbf{Y}_t) := D(0, t) \cdot E(t, \mathbf{Y}_t)$.

The likelihood of default of the counterparty is another important quantity in the calculation of the CVA. We will utilize the intensity (the so-called reduced form) model, of which the construction has been widely studied. Some work on initial intensity models was presented by Jarrow and Turnbull [21], Madan and Unal [28], Duffie and Singleton [13]. Lando [23] presented the term structure of defaultable bonds with the assumption of independence between the risk-free interest rate and the default intensity. A detailed discussion of intensity modeling of default risk can be found in the books by Bielecki and Rutkowski [3], Lando [23] and Brigo and Mercurio [6].

We also discuss the intensity model briefly here. Let $h_t := h(\mathbf{Y}_t)$ be the $\mathscr{F}_t$-intensity of a jump process and $\tau_d > 0$ be the first jump time of this process. We construct a right continuous process $H_t = \mathbb{1}(\tau_d \le t)$, where $\mathbb{1}(\cdot)$ is the indicator function. The natural filtration generated by $\mathscr{H}_t$ is given by $\mathscr{H}_t := \sigma(H_s)_{s\in[0,t]}$. The enlarged filtration $\mathscr{G}_t = \mathscr{H}_t \vee \mathscr{F}_t$ thus includes all information of default events and market quantities up to time t. The *survival probability* under the risk-free measure $\mathbb{Q}$ at time t can be expressed by an intensity function:

$$PS(t) = \mathbb{Q}\left(\tau_d > t \middle| \mathscr{G}_t\right) = \mathbb{E}^{\mathbb{Q}}\left[\mathbb{1}\left(\tau_d > t\right) \middle| \mathscr{G}_t\right] = \exp\left(-\int_0^t h_s ds\right), \tag{2}$$

where intensity h_t defines the default probability on a small interval dt when $\tau_d > t$.

By definition, CVA materializes the expected loss in the future, which can be expressed by:

$$\begin{aligned}\text{CVA} := \mathbb{E}^{\mathbb{Q}}\left[\text{LGD} \cdot E^*(\tau_d, \mathbf{Y}_{\tau_d}) \middle| \mathscr{G}_0\right] &= \int_0^T \mathbb{E}^{\mathbb{Q}}\left[\text{LGD} \cdot E^*(t, \mathbf{Y}_t) \cdot d\left(-PS(t)\right) \middle| \mathscr{G}_0\right] \\ &= \int_0^T \mathbb{E}\left[\text{LGD} \cdot E^*(t, \mathbf{Y}_t) \cdot h_t \cdot \exp\left(-\int_0^t h_s ds\right) \middle| \mathscr{F}_0\right] dt,\end{aligned} \tag{3}$$

where LGD is the loss given default (as a percentage), and the details of derivation of the third equality can be found in [23, p.117].

There are three key elements in the calculation of CVA: the loss given default, the discounted exposure and the survival/default probability of the counterparty. In a real-life situation these three elements are typically not independent. Wrong-way risk (WWR) incurs when the exposure is adversely correlated with the credit quality of the counterparty, which may significantly increase CVA [16]. When assuming independence, the calculation formula of CVA is given by:

$$\text{CVA} = \text{LGD} \int_0^T \mathbb{E}^{\mathbb{Q}}\left[E^*(t, \mathbf{Y}_t) \middle| \mathscr{F}_0\right] d\left(-PS(t)\right), \tag{4}$$

where LGD is assumed to be a fixed ratio based on market information, and the marginal survival probability $PS(t)$ can be obtained via the implied survival probability curve on the CDS market [6].

The well-known quantities of the exposure distribution, EE and PFE, are important for risk management [16]. The mathematical formulas for the EE and PFE quantities are given by:

$$\mathrm{EE}(t) := \mathbb{E}^{\mathbb{Q}}\left[\mathrm{E}(t, \mathbf{Y}_t)\middle|\mathscr{F}_0\right], \tag{5}$$

$$\mathrm{PFE}^{\alpha}(t) := \inf\left\{x\middle|\mathbb{Q}\left\{\mathrm{E}(t, \mathbf{Y}_t) < x\middle|\mathscr{F}_0\right\} > \alpha\right\}, \tag{6}$$

where α is the confidence level. For calculating PFE, the confidence level $\alpha = 97.5\%$ is commonly used to measure the 'worst' losses [16]. Both quantities are deterministic functions in the period $[0, T]$.

2.2 Affine Diffusion Models

For the asset price processes under study here, we will benefit from the affine diffusion (AD) class of Markov stochastic processes $(\mathbf{Y}_t)_{t\in[0,T]}$, which can be expressed by the general form,

$$d\mathbf{Y}_t = \mu\left(\mathbf{Y}_t\right)dt + \sigma(\mathbf{Y}_t)d\widetilde{\mathbf{W}}_t, \tag{7}$$

where $\widetilde{\mathbf{W}}_t$ is an $\mathscr{F}_t$-measurable column vector of independent Wiener processes under measure $\mathbb{Q}$ in $\mathbb{R}^n$, the drift term $\mu\left(\mathbf{Y}_t\right) : U \to \mathbb{R}^n$, and the volatility term $\sigma(\mathbf{Y}_t) : U \to \mathbb{R}^{n\times n}$. In the AD class it is assumed that the drift term, the covariance $\left(\sigma(\mathbf{Y}_t)\sigma(\mathbf{Y}_t)^T\right)$ and the interest rate are of the affine form, i.e.

$$\begin{aligned}\mu(\mathbf{Y}_t) &= a_0 + a_1\mathbf{Y}_t, \text{ for any } (a_0, a_1) \in \mathbb{R}^n \times \mathbb{R}^{n\times n},\\ \left(\sigma(\mathbf{Y}_t)\sigma(\mathbf{Y}_t)^T\right)_{ij} &= (c_0)_{ij} + (c_1)_{ij}^T\mathbf{Y}_t, \text{ with } (c_0, c_1) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n\times n},\\ r(\mathbf{Y}_t) &= r_0 + r_1^T\mathbf{Y}_t, \text{ for } (r_0, r_1) \in \mathbb{R} \times \mathbb{R}^n.\end{aligned} \tag{8}$$

With this type of model, it can be shown that the discounted characteristic function (dChF) is of the following form:

$$\begin{aligned}\Phi(\mathbf{u}, \mathbf{Y}_t, t, T) &= \mathbb{E}\left[\exp\left(-\int_t^T r_u du + i\mathbf{u}^T\mathbf{Y}_T\right)\middle|\mathscr{F}_t\right]\\ &= \exp\left(A(\mathbf{u}, \tau) + \mathbf{B}^T(\mathbf{u}, \tau)\mathbf{Y}_t\right),\end{aligned} \tag{9}$$

with time lag $\tau = T - t$. The coefficients satisfy the ODE system [13, 17]

$$\begin{aligned}\frac{d}{d\tau}A(\mathbf{u},\tau) &= -r_0 + \mathbf{B}^T(\mathbf{u},\tau)a_0 + \frac{1}{2}\mathbf{B}^T(\mathbf{u},\tau)c_0\mathbf{B}(\mathbf{u},\tau), \quad A(\mathbf{u},0)=0,\\ \frac{d}{d\tau}\mathbf{B}(\mathbf{u},\tau) &= -r_1 + a_1^T\mathbf{B}(\mathbf{u},\tau) + \frac{1}{2}\mathbf{B}^T(\mathbf{u},\tau)c_1\mathbf{B}(\mathbf{u},\tau), \quad \mathbf{B}(\mathbf{u},0)=i\mathbf{u}.\end{aligned} \tag{10}$$

The dChF facilitates the calculation of the discounted moments in Sect. 4.1, which is one of the key components within the SGBM algorithm.

Based on this general expression for affine models, we will discuss several hybrid models.

2.2.1 Black-Scholes Hull-White Model and Heston Model

The famous Black-Scholes option pricing partial differential equation (PDE) [4] is based on the assumptions that the asset price follows a geometric Brownian motion with constant volatility and constant interest rate. We first relax the assumption of constant interest rate by a stochastic instantaneous short-rate r_t. In practice, interest rates vary over time and by tenor T, as observed in the zero coupon bond curves in the market [6]. The instantaneous forward rate at time t for a maturity $T > t$ is defined by:

$$f(t,T) := -\frac{\partial \log p(t,T)}{\partial T}. \tag{11}$$

The characterization of the term structure of interest rates is well-known from Vasicek [34], Cox, Ingersoll, and Ross [11], and Hull and White [19]. In this paper, we will also employ the *Black-Scholes Hull-White* hybrid (BSHW) model. Under risk-neutral measure $\mathbb{Q}$, the dynamics of the model $\mathbf{Y}_t = [x_t, r_t]^T$ are given by the following SDEs [6]:

$$\begin{aligned} dx_t &= \left(r - \frac{1}{2}\sigma^2\right)dt + \sigma dW_t^x,\\ dr_t &= \lambda(\theta(t) - r_t)dt + \eta dW_t^r, \end{aligned} \tag{12}$$

where $x_t = \log(S_t)$ represents the log-asset variable; the two correlated Wiener processes (W_t^x, W_t^r) are defined by $W_t^x = \widetilde{W}_t^{(1)}$ and $W_t^r = \rho_{x,r}\widetilde{W}_t^{(1)} + \sqrt{1-\rho_{x,r}^2}\widetilde{W}_t^{(2)}$, where $\widetilde{W}_t^{(1)}$ and $\widetilde{W}_t^{(2)}$ are two independent standard Wiener processes under measure $\mathbb{Q}$ and $|\rho_{x,r}| < 1$ is the instantaneous correlation parameter between the asset price and the short rate process; positive parameters σ and η denote the volatility of equity and interest rate, respectively; the drift term $\theta(t)$ is a deterministic function chosen to fit the term structure observed in the market, which must satisfy:

$$\theta(t) = f(0,t) + \frac{1}{\lambda}\frac{\partial}{\partial t}f(0,t) + \frac{\eta^2}{2\lambda^2}(1 - \exp(-2\lambda t)). \tag{13}$$

Another way of extending the Black-Scholes model is to define the variance as a diffusion process, like in the stochastic volatility model developed by Heston [18]. With state variable $\mathbf{Y}_t = [x_t, v_t]^T$, the Heston model is given by:

$$\begin{aligned} dx_t &= \left(r - \frac{1}{2}v_t\right)dt + \sqrt{v_t}dW_t^x, \\ dv_t &= \kappa(\bar{v} - v_t)dt + \gamma\sqrt{v_t}dW_t^v, \end{aligned} \tag{14}$$

where r is a constant interest rate; the two correlated Wiener processes (W_t^x, W_t^v) are defined by $W_t^x = \widetilde{W}_t^{(1)}$ and $W_t^v = \rho_{x,v}\widetilde{W}_t^{(1)} + \sqrt{1-\rho_{x,v}^2}\widetilde{W}_t^{(3)}$, where $\widetilde{W}_t^{(1)}$ and $\widetilde{W}_t^{(3)}$ are two independent standard Wiener processes under measure $\mathbb{Q}$ and $|\rho_{x,v}| < 1$ is the instantaneous correlation parameter between the asset price and the variance process; the constant positive parameters κ, $\bar{v}$, γ determine the reverting speed, the reverting level and vol-of-vol parameters, respectively. The associated PDE can be found in [18, p. 329].

2.2.2 Heston Hull-White Model and H1HW Model

Consider a state vector including all these stochastic quantities, i.e. $\mathbf{Y}_t = [x_t, v_t, r_t]^T$. The corresponding model can be defined by adding a HW interest rate process to the Heston stochastic volatility dynamics, as presented in [17]. The hybrid model of the equity, stochastic Heston asset volatility and stochastic interest rate is represented by the following SDEs:

$$\begin{aligned} dx_t &= \left(r_t - \frac{1}{2}v_t\right)dt + \sqrt{v_t}dW_t^x, \\ dv_t &= \kappa(\bar{v} - v_t)dt + \gamma\sqrt{v_t}dW_t^v, \\ dr_t &= \lambda(\theta(t) - r_t)dt + \eta dW_t^r, \end{aligned} \tag{15}$$

where the correlated Wiener processes (W_t^x, W_t^v, W_t^r) are defined by $W_t^x = \widetilde{W}_t^{(1)}$, $W_t^v = \rho_{x,v}\widetilde{W}_t^{(1)} + \sqrt{1-\rho_{x,v}^2}\widetilde{W}_t^{(2)}$, $W_t^r = \rho_{x,r}\widetilde{W}_t^{(1)} - \frac{\rho_{x,v}\rho_{x,r}}{\sqrt{1-\rho_{x,v}^2}}\widetilde{W}_t^{(2)} + \sqrt{\frac{1-\rho_{x,v}^2-\rho_{x,r}^2}{1-\rho_{x,v}^2}}\widetilde{W}_t^{(3)}$, in which $\widetilde{W}_t^{(1)}$, $\widetilde{W}_t^{(2)}$ and $\widetilde{W}_t^{(3)}$ are three independent standard Wiener processes under the risk-neutral measure $\mathbb{Q}$, and $\rho_{x,v}$ and $\rho_{x,r}$ are correlation parameters that satisfy $\rho_{x,v}^2 + \rho_{x,r}^2 < 1$; the parameters λ, $\theta(t)$, η are as in (12), and κ, $\bar{v}$ and γ are as in (14); the initial values satisfy $r_0 > 0$ and $v_0 > 0$.

The Heston Hull-White (HHW) SDE system in (15) is not affine. Conditioned on information at time t, the symmetric covariance matrix at time $s > t$ is given by:

$$\sigma\left(\mathbf{Y}_s\right)\sigma\left(\mathbf{Y}_s\right)^T=\begin{pmatrix} v_s & \rho_{x,v}v_s & \sqrt{v_s}\eta\rho_{x,r}\\ * & \gamma^2 v_s & 0\\ * & * & \eta^2\end{pmatrix}. \tag{16}$$

where the term $\sqrt{v_s}$ is not linear. Grzelak and Oosterlee in [17] approximated the covariance matrix in (16) by

$$\sigma\left(\mathbf{Y}_s\right)\sigma\left(\mathbf{Y}_s\right)^T\approx\hat{\sigma}\left(\mathbf{Y}_s\right)\hat{\sigma}\left(\mathbf{Y}_s\right)^T=\begin{pmatrix} v_s & \rho_{x,v}v_s & \mathbb{E}\left[\sqrt{v_s}\middle|v_t\right]\eta\rho_{x,r}\\ * & \gamma^2 v_s & 0\\ * & * & \eta^2\end{pmatrix}, \tag{17}$$

where the term $\sqrt{v}_s$ is approximated by its conditional expectation $\mathbb{E}\left[\sqrt{v_s}\middle|v_t\right]$, for which an analytic formula is given by:

$$\mathbb{E}\left[\sqrt{v_s}\middle|v_t\right]=\sqrt{2c(\tau_1)}e^{-\frac{\bar{\lambda}(\tau_1,v_t)}{2}}\sum_{k=0}^{\infty}\frac{1}{k!}\left(\frac{\bar{\lambda}(\tau_1,v_t)}{2}\right)^k\frac{\Gamma\left(\frac{1+d}{2}+k\right)}{\Gamma\left(\frac{d}{2}+k\right)}, \tag{18}$$

with $\tau_1:=s-t$, and

$$c(\tau_1)=\frac{1}{4\kappa}\gamma^2(1-e^{-\kappa\tau_1}),\quad d=\frac{4\kappa\bar{v}}{\gamma^2},\quad \bar{\lambda}(\tau_1,v_t)=\frac{4\kappa v_t e^{-\kappa\tau_1}}{\gamma^2(1-e^{-\kappa\tau_1})}. \tag{19}$$

This affine approximation of the HHW model with covariance (17) is called the H1HW model, and details can be found in [17]. In this paper, we further make an approximation of the calculation in (18), as presented in Appendix 3.

2.3 Pricing European, Bermudan and Barrier Options

We will study the CVA, EE and PFE of several types of options to show the flexibility of SGBM. We present the backward valuation dynamics framework for European, Bermudan and barrier options in this section. Let the collection of equally-spaced discrete monitoring dates be:

$$\mathscr{T}=\{0=t_0<t_1<\cdots<t_M=T,\Delta t=t_{m+1}-t_m\}.$$

The options will be valued at so-called monitoring dates to determine the exposure profiles. The received payoff from immediate exercise of the option at time t_m is given by

$$g(S_m) := \max\left(\omega(S_m - K), 0\right), \quad \text{with} \begin{cases} \omega = 1, & \text{for a call;} \\ \omega = -1, & \text{for a put,} \end{cases} \tag{20}$$

where K is the strike value and S_m is the underlying asset variable at time t_m.

The *continuation value* of the option at time t_m can be expressed by the conditional expectation of the discounted option value at time t_{m+1}. As we have assumed the Markov property of the process $\mathbf{Y}_m$, we replace the filtration $\mathscr{F}_m$ in the conditional expectation, i.e. the continuation values of the option will be written as a function of the state variable $\mathbf{Y}_m$, i.e.

$$c(t_m, \mathbf{Y}_m) := \mathbb{E}^{\mathbb{Q}}\left[D\left(t_m, t_{m+1}\right) \cdot V(t_{m+1}, \mathbf{Y}_{m+1}) \middle| \mathbf{Y}_m\right], \tag{21}$$

where $\mathbf{Y}_m$ is the state variable at time t_m, and $V(t_{m+1}, \mathbf{Y}_{m+1})$ is the option value at time t_{m+1}.

2.3.1 Bermudan Options

Bermudan options can be exercised at a series of time points before expiry date T. Denote the set of early-exercise dates by $\mathscr{T}_{\mathrm{E}}$. We will take a small step size Δt when simulating the market variables to enhance the accuracy of the CVA calculation, and we assume that the Bermudan option can only be exercised at some of these dates, i.e. $\mathscr{T}_{\mathrm{E}} \subset \mathscr{T}$.

We also assume that the option holder makes the exercise strategy aiming for the 'optimal' profit, and the option holder is not influenced by the credit quality of the option writer when making the decision. We further denote the optimal stopping time by τ_B, which is the optimal time to exercise the option under the assumptions. It should maximize the expected payoff at time $t = 0$, i.e.

$$V^{\mathrm{Berm}}(t_0, \mathbf{Y}_0) = \max_{\tau_B \in \mathscr{T}_{\mathrm{E}}} \mathbb{E}\left[D(0, \tau_B) \cdot g(S_{\tau_B}) \middle| \mathbf{Y}_0\right]. \tag{22}$$

The essential idea of pricing Bermudan options by simulation is to determine the optimal exercise strategy for each path. At each exercise date, the option holder compares the received payoff from immediate exercise with the expected payoff from continuation of the option to determine the optimal exercise strategy. The dynamics of pricing Bermudan options in backward induction derived by the Snell envelope [14, 27] can be expressed by:

$$V^{\mathrm{Berm}}(t_m, \mathbf{Y}_m) = \begin{cases} g(S_M) & \text{for } t_M = T, \\ \max\left\{c(t_m, \mathbf{Y}_m), g(S_m)\right\}, & \text{for } t_m \in \mathscr{T}_{\mathrm{E}}, \\ c(t_m, \mathbf{Y}_m), & \text{for } t_m \in \mathscr{T} - \mathscr{T}_{\mathrm{E}}. \end{cases} \tag{23}$$

2.3.2 European Options

Similar to pricing Bermudan options, the exposure profile of a European option can be determined based on simulation. The European option value at time T equals the received payoff $V^{\text{Euro}}(t_M, \mathbf{Y}_M) = g(S_M)$; at time points $t_m < T$, the value of the European option is equal to the discounted conditional expected payoff, i.e.,

$$V^{\text{Euro}}(t_m, \mathbf{Y}_m) := \mathbb{E}\left[D(t_m, t_M) \cdot g(S_M) \middle| \mathbf{Y}_m\right], \tag{24}$$

where $g(S_M)$ is the received payoff at time $t_M = T$. By the tower property of expectations, it can be calculated in a backward iteration as:

$$\begin{aligned} V^{\text{Euro}}(t_m, \mathbf{Y}_m) &= \mathbb{E}\left[D(t_m, t_{m+1}) \cdot \mathbb{E}\left[D(t_{m+1}, t_M) \cdot g(S_M) \middle| \mathbf{Y}_{m+1}\right] \middle| \mathbf{Y}_m\right] \\ &= \mathbb{E}\left[D(t_m, t_{m+1}) \cdot V^{\text{Euro}}(t_{m+1}, \mathbf{Y}_{m+1}) \middle| \mathbf{Y}_m\right] = c(t_m, \mathbf{Y}_m). \end{aligned} \tag{25}$$

2.3.3 Barrier Options

Barrier options become active/knocked out when the underlying asset reaches a predetermined level, i.e. the *barrier* level. There are four main types of barrier options: up-and-out, down-and-out, up-and-in, down-and-in options. Here we focus on the *down-and-out* barrier options. A down-and-out barrier option is active initially and gets knocked out (looses its value except for some rebate value) when the underlying hits the barrier; otherwise if the option is not knocked out during its lifetime, the holder will receive the payoff value at the expiry date T. The backward pricing dynamics of the down-and-out barrier options are thus given by [14],

$$V^{\text{barr}}(t_m, \mathbf{Y}_m) = \begin{cases} g(S_m) \cdot \mathbb{1}_{\{S_m > L\}} + r_b \cdot \mathbb{1}_{\{S_m \le L\}}, & \text{for } t_M = T, \\ c(t_m, \mathbf{Y}_m) \cdot \mathbb{1}_{\{S_m > L\}} + r_b \cdot \mathbb{1}_{\{S_m \le L\}}, & \text{for } t_m < T, \end{cases} \tag{26}$$

where $\mathbb{1}(\cdot)$ is the indicator function, L is the barrier level and r_b is the rebate value.

3 The Stochastic Grid Bundling Method (SGBM)

Monte Carlo simulation plays a primary role in computing CVA, i.e. generating N independent scenarios for each monitoring date $\mathscr{T}$. We denote the realization of the state vector $\mathbf{Y}_m$ on the i-th path at time t_m by $\hat{\mathbf{y}}_m(i)$, $i = 1, \ldots, N$. After finishing the calculation of the exposure profile on the generated stochastic grid, the CVA, assuming independence of exposures and defaults, can be computed by the following discrete formula:

$$\text{CVA} \approx \text{LGD} \sum_{m=0}^{M-1} \frac{1}{N} \sum_{i=1}^{N} \left(\exp\left(-\sum_{k=0}^{m-1} r(\hat{\mathbf{y}}_k(i)) \Delta t \right) \cdot E(t_m, \hat{\mathbf{y}}_m(i)) \right)$$
$$\cdot \Big(PS(t_m) - PS(t_{m+1}) \Big). \tag{27}$$

Similarly, the value at time t_m of the EE and PFE functions can be approximated by:

$$\text{EE}(t_m) \approx \frac{1}{N} \sum_{i=1}^{N} E(t_m, \hat{\mathbf{y}}_m(i)), \tag{28}$$

$$\text{PFE}(t_m) \approx \text{quantile}(E(t_m, \hat{\mathbf{y}}_m(i)), 97.5\%), \tag{29}$$

where the confidence level is set to $\alpha = 97.5\%$.

At expiry date $t_M = T$, the option values on each path can be computed immediately by the received payoff values. The key problem is to calculate the continuation values on each path in the backward algorithm at each monitoring time $t_m < T, m = 0, 1, \ldots, M-1$. SGBM combines regression and bundling techniques to compute these expected values.

3.1 Calculation of the Continuation Values

At time $t_m < T$, the generated paths are clustered into some non-overlapping bundles with as a criterion that the realizations $\hat{\mathbf{y}}_m(i)$ on paths within the same bundle should share similar values. The indices of the paths in the j-th bundle are in a set $\mathscr{B}_m^j$, $j = 1, 2, \ldots, J$, where J is the number of bundles. The realizations $\hat{\mathbf{y}}_m(i)$ of the state vector $\mathbf{Y}_m$ within the j-th bundle form a bounded domain $\mathbf{I}_m^j \subset \mathbb{R}^n$, when $m = 1, 2, \ldots, M-1$, given by

$$\mathbf{I}_m^j = \prod_{l=1}^{n} \left[\max_{i \in \mathscr{B}_m^{j-1}} \left(\hat{y}_m^{(l)}(i) \right), \max_{i \in \mathscr{B}_m^{j}} \left(\hat{y}_m^{(l)}(i) \right) \right], \tag{30}$$

where $\hat{y}_m^{(l)}(i)$ represents the l-th dimension of the realization $\hat{\mathbf{y}}(i)$, and $j = 2, 3, \ldots, J$. When $j = 1$ we define the realized domain $\mathbf{I}_m^1 = \prod_{l=1}^{n} \left[\min_{i \in \mathscr{B}_m^1} \left(\hat{y}_m^{(l)}(i) \right), \max_{i \in \mathscr{B}_m^1} \left(\hat{y}_m^{(l)}(i) \right) \right]$.

These subdomains $\{\mathbf{I}_m^j\}_{j=1}^J$ are disjoint. At the same time, the corresponding realizations $\hat{\mathbf{y}}_{m+1}(i)$ of the state vector $\mathbf{Y}_{m+1}$ within the j-th bundle also form a bounded domain in $\mathbb{R}^n$, i.e.

$$\mathbf{U}_{m+1}^j = \prod_{l=1}^{n} \left[\min_{i \in \mathscr{B}_m^j} \left(\hat{y}_{m+1}^{(l)}(i) \right), \max_{i \in \mathscr{B}_m^j} \left(\hat{y}_{m+1}^{(l)}(i) \right) \right], \tag{31}$$

where $\hat{y}^{(l)}_{m+1}(i)$ represents the l-th dimension of the realization $\hat{\mathbf{y}}_{m+1}(i)$. These domains typically overlap.

We assume that the option function $V(t_{m+1},\cdot)$ is an element of the L^2 space on the finite domain $\mathbf{U}^j_{m+1}$, i.e. it is square-integrable over $\mathbf{U}^j_{m+1}$ with some measure. Suppose that we have the values of this option function w.r.t. the realizations $\hat{\mathbf{y}}_{m+1}(i)$ on all paths, denoted by $\hat{v}_{m+1}(i)$, $i = 1, 2, \ldots, N$. Given the set of points $\{(\hat{\mathbf{y}}_{m+1}(i), \hat{v}_{m+1}(i))\}_{i=1}^N$, $i \in \mathscr{B}^j_m$, a commonly used approximation of the option function is a constructed function that is the 'best fit' for the data set in least squares sense. With a set of some basis functions $\{\phi_k\}_{k=1}^H$ in L_2, the option function can be approximated on $\mathbf{U}^j_{m+1}$ by a linear combination of the basis functions:

$$V(t_{m+1}, \mathbf{Y}_{m+1}) \approx Z_1(t_{m+1}, \mathbf{Y}_{m+1}) := \sum_{k=1}^{H} \beta^j_m(k)\phi_k(\mathbf{Y}_{m+1}), \tag{32}$$

where H is the number of basis functions, and $\beta^j_m(k)$ are the constant coefficients at time t_m of the k-th basis function ϕ_k within the j-th bundle $\mathscr{B}^j_m$, determined by regression:

$$\underset{\beta^j_m(k)\in\mathbb{R}, k=1,\ldots,H}{\arg\min} \sum_{i\in\mathscr{B}^j_m} \left(\hat{v}_{m+1}(i) - \sum_{k=1}^{H} \beta^j_m(k)\phi_k(\hat{\mathbf{y}}_{m+1}(i)) \right)^2, \tag{33}$$

of which the solution is denoted by $\{\hat{\beta}^j_m(k)\}_{k=1}^H$. Within the j-th bundle, the approximation of the option function on $\mathbf{U}^j_{m+1}$ is thus given by:

$$V(t_{m+1}, \mathbf{Y}_{m+1}) \approx Z_2(t_{m+1}, \mathbf{Y}_{m+1}) := \sum_{k=1}^{H} \hat{\beta}^j_m(k)\phi_k(\mathbf{Y}_{m+1}). \tag{34}$$

Hence the continuation function on the bounded domain $\mathbf{I}^j_m$ can be approximated by a linear combination of the conditional expected discounted basis functions defined by:

$$\begin{aligned} c_2(t_m, \mathbf{Y}_m) &:= \mathbb{E}^{\mathbb{Q}}\left[D(t_m, t_{m+1}) \cdot Z_2(t_{m+1}, \mathbf{Y}_{m+1}) \middle| \mathbf{Y}_m \right] \\ &= \sum_{k=1}^{H} \hat{\beta}^j_m(k)\psi_k(\mathbf{Y}_m, \Delta t), \end{aligned} \tag{35}$$

where the conditional expectation of the k-th discounted basis function is given by

$$\psi_k(\mathbf{Y}_m, \Delta t) := \mathbb{E}^{\mathbb{Q}}\left[D(t_m, t_{m+1}) \cdot \phi_k(\mathbf{Y}_{m+1}) \middle| \mathbf{Y}_m \right]. \tag{36}$$

We will approximate the 'real' continuation function $c(t_m, \cdot)$ given in equation (21) by the function $c_2(t_m, \cdot)$ defined in equation (35) on the bounded domain $\mathbf{I}_m^j$. When analytic formulas of the functions $\{\psi_k\}_{k=1}^H$ defined in (36) are available, the continuation value w.r.t. realization $\hat{\mathbf{y}}_m(i)$ on the i-th path within the j-th bundle can be easily computed by:

$$c(t_m, \hat{\mathbf{y}}_m(i)) \approx c_2(t_m, \hat{\mathbf{y}}_m(i)) = \sum_{k=1}^{H} \hat{\beta}_m^j(k)\psi_k(\hat{\mathbf{y}}_m(i), \Delta t). \tag{37}$$

In addition, we will show that the error of approximation of the continuation function at time t_m is bounded by the error of approximation of the option function at time t_{m+1} in Sect. 3.4.

3.2 Backward Algorithm

From Sect. 3.1, it is clear that the continuation values on each path at time t_m can be calculated in backward fashion as long as the option values at these paths at time t_{m+1} are available. In this section, we will present the backward algorithm of the SGBM for computing exposures of options, first for Bermudan options.

Initializing: At time $t_M = T$, the option values $\{\hat{v}_M(i)\}_{i=1}^N$ on all paths can be calculated from the received payoff.

Backward iteration: At time $t_m < T$, $m = M-1, M-2, \ldots, 1$,

- Step I: apply a bundling technique to cluster all paths into non-overlapping bundles, indexed by $\mathscr{B}_m^j, j = 1, 2, \ldots, J$.
- Step II: within the j-th bundle, $j = 1, 2, \ldots, J$, utilize the regression technique to calculate the continuation values at time t_m by:
  - Step (i): approximate coefficients $\{\hat{\beta}_m^j(k)\}_{k=1}^H$ within the j-th bundle by formula (33);
  - Step (ii): calculate continuation values on each path by formula (35) using the approximated coefficients obtained in Step (i).
- Step III: determine option values $\{\hat{v}_m(i)\}_{i=1}^N$ on all paths at time t_m by formula (23) using the approximated continuation values obtained in Step (ii).
- Step IV: determine exposure values at time t_m by formula (1) on each path: if the option at a path is exercised at time t_m, then the corresponding exposure values from time t_m to time t_M at this path are assigned value zero; otherwise the exposure values are the computed continuation values on the path.

Finalizing: At time $t_0 = 0$, approximate directly the coefficients $\{\hat{\beta}_m(k)\}_{k=1}^H$ and calculate the continuation value at time t_0, which is also the option value at time t_0.

The backward algorithm of calculating the exposure profile of a European option or a barrier option is the same as the algorithm for a Bermudan option, except that the pricing formula (23) in Step III needs to be replaced by formula (25) for pricing European options or formula (26) for pricing barrier options, respectively.

3.3 Sensitivities of EE

The sensitivities *Delta* (Δ_{EE}) and *Gamma* (Γ_{EE}) of EE w.r.t. the change of the underlying asset price S_0 can be computed in the same backward algorithm for the computation of the exposure profile. At time $t_M = T$, we simply assign value zero to these derivatives of the EE function. At time $t_m < T$, the sensitivities can be computed by:

$$\Delta_{\text{EE}}(t_m) := \frac{\partial \text{EE}}{\partial S_0}(t_m) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \text{E}}{\partial x_m}(t_m, \hat{\mathbf{y}}_m(i)) \cdot \frac{1}{S_0}, \tag{38}$$

$$\Gamma_{\text{EE}}(t_m) := \frac{\partial^2 \text{EE}}{\partial S_0^2}(t_m) \approx \frac{1}{N}\sum_{i=1}^{N} \left(\frac{\partial^2 \text{E}}{\partial x_m^2}(t_m, \hat{\mathbf{y}}_m(i)) - \frac{\partial \text{E}}{\partial x_m}(t_m, \hat{\mathbf{y}}_m(i)) \right) \cdot \frac{1}{S_0^2}, \tag{39}$$

where $x_m = \log(S_m)$ represents the log-asset value at time t_m. The derivation of formulas (38) and (39) is presented here. At time t_m, the first derivative of the EE function can be computed by

$$\frac{\partial \text{EE}}{\partial S_0}(t_m) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \text{E}}{\partial S_0}(t_m, \hat{\mathbf{y}}_m(i)), \tag{40}$$

by the chain rule,

$$\frac{\partial \text{E}}{\partial S_0}(t_m, \hat{\mathbf{y}}_m(i)) = \frac{\partial \text{E}}{\partial x_m} \cdot \frac{\partial x_m}{\partial S_m} \cdot \frac{\partial S_m}{\partial S_0}(t_m, \hat{\mathbf{y}}_m(i)), \tag{41}$$

where $x_m := \log S_m$, and

$$\frac{\partial x_m}{\partial S_m} = \frac{1}{S_m}, \quad \frac{\partial S_m}{\partial S_0} = \frac{S_m}{S_0}. \tag{42}$$

The second equation in (42) can be derived as follows. The asset value S_t follows a Geometric Brownian motion process, i.e.

$$d \log S_t = \mu_t dt + \sigma_t dW_t. \tag{43}$$

By integrating both sides, we obtain

$$S_t = S_0 \cdot \exp\left(\int_0^t (\mu_s ds + \sigma_s dW_s)\right), \tag{44}$$

hence the derivative of S_t w.r.t. S_0 can be expressed by

$$\frac{\partial S_t}{\partial S_0} = \exp\left(\int_0^t (\mu_s ds + \sigma_s dW_s)\right) = \frac{S_t}{S_0}. \tag{45}$$

So, the first derivative of the EE function can be expressed by

$$\frac{\partial \mathrm{EE}}{\partial S_0}(t_m) \approx \frac{1}{N}\sum_{i=1}^{N} \frac{\partial \mathrm{E}}{\partial x_m}(t_m, \hat{\mathbf{y}}_m(i)) \cdot \frac{1}{S_0}. \tag{46}$$

From (46), the second derivative can be derived by

$$\begin{aligned}\frac{\partial^2 \mathrm{EE}}{\partial S_0^2}(t_m) &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial \mathrm{E}}{\partial x_m^2}(t_m, \hat{\mathbf{y}}_m(i)) \cdot \frac{x_m}{S_m} \cdot \frac{S_m}{S_0} \cdot \frac{1}{S_0} + \frac{\partial \mathrm{E}}{\partial x_m}(t_m, \hat{\mathbf{y}}_m(i)) \cdot \left(-\frac{1}{S_0^2}\right)\right)\\ &= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial^2 \mathrm{E}}{\partial x_m^2}(t_m, \hat{\mathbf{y}}_m(i)) - \frac{\partial \mathrm{E}}{\partial x_m}(t_m, \hat{\mathbf{y}}_m(i))\right) \cdot \frac{1}{S_0^2}.\end{aligned} \tag{47}$$

For those paths on which the option is alive at time t_m, the first and the second derivatives of the exposure function are given by

$$\begin{aligned}\frac{\partial \mathrm{E}}{\partial x_m}(t_m, \mathbf{Y}_m) &:= \frac{\partial c}{\partial x_m}(t_m, \mathbf{Y}_m),\\ \frac{\partial^2 \mathrm{E}}{\partial x_m^2}(t_m, \mathbf{Y}_m) &:= \frac{\partial^2 c}{\partial x_m^2}(t_m, \mathbf{Y}_m),\end{aligned} \tag{48}$$

where the derivatives of the continuation function w.r.t. x_m within the j-th bundle are approximated by

$$\begin{aligned}\frac{\partial c}{\partial x_m}(t_m, \mathbf{Y}_m) &\approx \sum_{k=1}^{H} \hat{\beta}_m^j(k) \frac{\partial \psi_k}{\partial x_m}(\mathbf{Y}_m, \Delta t),\\ \frac{\partial^2 c}{\partial x_m^2}(t_m, \mathbf{Y}_m) &\approx \sum_{k=1}^{H} \hat{\beta}_m^j(k) \frac{\partial^2 \psi_k}{\partial x_m^2}(\mathbf{Y}_m, \Delta t),\end{aligned} \tag{49}$$

with the same coefficient set $\{\hat{\beta}_m^j(k)\}_{k=1}^H$ as in (35).

For those paths on which the option has been exercised or knocked out at time t_m, the derivatives of EE are given value zero, as the exposure values on these paths are zero.

3.4 Convergence Results

The so-called *direct estimator* is obtained in the backward algorithm by regression [20]. With convexity of the 'max' function, it can be proven by induction that the direct estimator is often higher than the true value with some bias, and that the direct estimator converges to the option value as the number of paths and the number of monomial basis functions goes to infinity. See Theorem 2 and Theorem 4 in [20].

In addition, an estimator can be made based on the average cash flow of a second set of paths, referred to as the *path estimator*. Using the coefficients obtained by regression based on one set of paths, an approximation of the optimal early exercise strategy of another set of paths can be made by comparing values of continuation and values of immediate exercise. The path estimator is often a lower bound of the option value, converging a.s. as the number of paths goes to infinity [20], since the option value computed by the optimal early exercise strategy is the supremum of the option value at time $t = 0$ by definition. Details of the proof can be found in [20, 27].

For European and barrier options, one can take the discounted average of the MC paths as our reference. For Bermudan options, the direct and path estimators provide a conservative confidence interval for the true option value [20]:

$$\left[V^{\text{path}}(0) - 1.96\frac{\hat{s}_{\text{path}}}{\sqrt{N_s}}, V^{\text{direct}}(0) + 1.96\frac{\hat{s}_{\text{direct}}}{\sqrt{N_s}} \right], \tag{50}$$

where $\hat{s}_{\text{path}}$ and $\hat{s}_{\text{direct}}$ are the sample standard deviations for the path and direct estimator respectively, and $V^{\text{path}}(0)$ and $V^{\text{direct}}(0)$ are the sample means of the path and direct estimators respectively; these sample means and sample standard deviations are based on N_s independent trials.

The approximation of the option function converges as the number of paths, the number of basis functions and the number of bundles go to infinity. Details of this can be found in Appendix 4. From the discussion of convergence in Appendix 4, we can also conclude that by using bundles, the option function can be approximated well piece-wise functions, even with a low order $p = 1$. This advantage of the SGBM approach will reduce the computational effort for increasing problem dimensions. In addition, the error of approximation of the continuation function can be uniformly bounded by the error in approximating the option function, as stated in Proposition 1. It ensures the accuracy of the computed continuation values by SGBM on each path, which is important for computing exposure profiles.

Proposition 1 *At time t_m, the error of approximating the continuation function by SGBM is uniformly bounded by the error of approximation of the option function within each bundle, given by*

$$\left| c(t_m, \mathbf{Y}_m) - c_2(t_m, \mathbf{Y}_m) \right| \leq \| V(t_{m+1}, \cdot) - Z_2(t_{m+1}, \cdot) \|_{L_2}$$

$$= \left(\int_{\mathbf{Y}_{m+1} \in \mathbb{R}^n} \left(V(t_{m+1}, \mathbf{Y}_{m+1}) - Z_2(t_{m+1}, \mathbf{Y}_{m+1}) \right)^2 d\mu_{(\mathbf{Y}_{m+1} | \mathbf{Y}_m)} \right)^{\frac{1}{2}}, \quad (51)$$

where $\mu_{(\mathbf{Y}_{m+1}|\mathbf{Y}_m)}$ is the probability measure conditioned on $\mathbf{Y}_m \in \mathbf{I}^j_m$ under the risk-neutral measure $\mathbb{Q}$.

Proof By Jensen's inequality it is proved in Appendix 5.

4 Choice of Basis Functions and Bundling

4.1 The Monomial Basis and the Discounted Moments

Essentially, the approximation of the option function expressed in (32) is its projection onto a space consisting of basis functions on the bounded domain $\mathbf{U}^j_{m+1}$. For the polynomial space, it is natural to take *monomials* as the basis, as all monomials with order lower or equal to any degree $p \in \mathbb{N}$ can form a closure. With a state vector $\mathbf{Y}_t = [Y^1_t, Y^2_t, \ldots, Y^l_t, \ldots, Y^n_t] \in \mathbb{R}^n$, a monomial basis of order $p > 0$ can be expressed by $\prod_{l=1}^{n} \left(Y^l_t\right)^{q_l}$, where $\left(\sum_{l=1}^{n} q_l\right) = p,$ with $q_l \geq 0$ for any l. The number of basis functions of a monomial basis of order less than or equal to p is $H = \frac{(n+p)!}{p!n!}$. We denote the polynomial space of order p on the bounded domain $\mathbf{U}^j_{m+1}$ by:

$$\mathscr{P}(\mathbf{U}^j_{m+1}, p) := \left\{ f \middle| f(\mathbf{y}) = \sum_{k=1}^{H} \beta(k) \phi_k(\mathbf{y}),\ \mathbf{y} \in \mathbf{U}^j_{m+1},\ \boldsymbol{\beta} \in \mathbb{R}^H. \right\}, \quad (52)$$

where $\boldsymbol{\beta} := [\beta(1), \beta(2), \cdots, \beta(H)] \in \mathbb{R}^H$, and $\{\phi_k\}_{k=1}^H$ is the monomial basis. Table 1 presents the monomial basis set for the hybrid models in this paper with degree $p = \{1, 2, 3\}$.

The monomial basis grows rapidly with the dimension of the state variable n and the polynomial order p. In the algorithm of SGBM, bundling will enhance the accuracy and thus a lower degree p can be employed to achieve a certain accuracy level, as we will see in the numerical Sect. 5.

Table 1 The monomial basis for the hybrid models

order p	Heston	BSHW	HHW →H1HW
1	$\{1, x_t, v_t\}$	$\{1, x_t, r_t\}$	$\{1, x_t, v_t, r_t\}$
2	$\{1, x_t, v_t, x_t^2, x_t v_t, v_t^2\}$	$\{1, x_t, r_t, x_t^2, x_t r_t, r_t^2\}$	$\{1, x_t, v_t, r_t, x_t^2, x_t v_t,$ $v_t^2, x_t r_t, r_t^2, v_t r_t\}$
3	$\{1, x_t, v_t, x_t^2, x_t v_t, v_t^2,$ $x_t^3, x_t^2 v_t, x_t v_t^2, v_t^3\}$	$\{1, x_t, r_t, x_t^2, x_t r_t, r_t^2,$ $x_t^3, x_t^2 r_t, x_t r_t^2, r_t^3\}$	

The expected value of a discounted monomial basis is the *discounted moment*, for which an analytic formula, the ψ-function, is needed in the calculation of the continuation function. Over a time period $[s, t]$, the k-th discounted moment of an n-dimensional vector $\mathbf{Y}_t$, corresponding to the monomial basis $\prod_{l=1}^{n} \left(Y_t^l\right)^{q_l}$ with degree $0 \leq \left(\sum_{l=1}^{n} q_l\right) \leq p$, is defined by:

$$\psi_k(\mathbf{Y}_s, t-s) := \mathbb{E}^{\mathbb{Q}}\left[\prod_{l=1}^{n} \left(Y_t^l\right)^{q_l} \cdot D(s,t) \middle| \mathbf{Y}_s\right], \tag{53}$$

which can be derived by the associated dChF of the dynamics,

$$\psi_k(\mathbf{Y}_s, t-s) = \frac{1}{(i)^p} \prod_{l=1}^{n} \frac{\partial^{q_l} \Phi}{\partial u_l^{q_l}}(\mathbf{u}; \mathbf{Y}_s, t-s)\bigg|_{\mathbf{u}=\mathbf{0}}, \tag{54}$$

where i represents the imaginary unit, vector $\mathbf{u} = [u_1, u_2, \ldots, u_l, \ldots, u_n] \in \mathbb{R}^n$ and the function $\Phi(\mathbf{u}; \mathbf{Y}_s, t-s)$ is the dChF of the underlying dynamics given in equation (9).

So, the discounted moments of AD processes of any order can be expressed in closed form, i.e. we have all discounted moments corresponding to the monomial basis presented in Table 1. For the HHW process, of course, we base them on the H1HW approximate model.

4.2 A Bundling Method

We introduce a technique for making bundles in SGBM such that there is an equal number of paths within each bundle. It is called the *equal-number bundling* technique. The same technique of clustering paths is found in [10, 26]. The advantages of this bundling technique are that the number of paths within each bundle will grow in portion to the number of paths, and that there will be a sufficient number of paths for regression when the total number of paths is large.

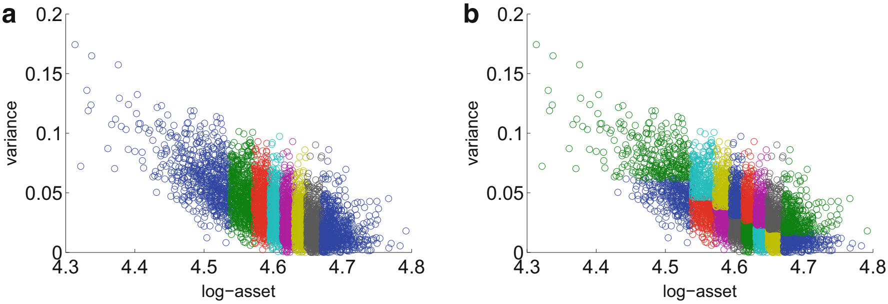


Fig. 1 Equal-number bundling. Each colored block represents a disjoint subdomain $\mathbf{I}_{m,j}$. (**a**) First iteration, J_1. (**b**) Second iteration, J_2

We use the Heston model to present the bundling technique, where the 2D state vector is denoted by $\mathbf{Y}_t = [x_t, v_t]^T$. First, all paths are sorted w.r.t. their log-asset values, and clustered into J_1 bundles with respect to their ranking, ensuring that within each bundle, the number of paths is equal to $\frac{N}{J_1}$; subsequently, within each bundle we perform a second sorting w.r.t. the variance values and cluster the paths into J_2 bundles. After these two iterations, the total number of bundles will be $J = J_1 \cdot J_2$.

The two steps are visualized in Fig. 1, where scatter plots demonstrate the 2D domain for the Heston model, at some time instant t_m. In plot (a), the paths are first clustered into 8 bundles w.r.t. the values of the log-asset, while in plot (b), the paths within each bundle are again clustered into 2 bundles w.r.t. the value of the variance. The total number of bundles is thus 16.

In a similar way, paths simulated under the HHW model can be clustered by the realized values of the log-asset (x_t), variance (v_t) and interest rate (r_t) values, in this order. We denote the number of bundles in these three dimension by J_1, J_2 and J_3, and the total number of bundles $J = J_1 \cdot J_2 \cdot J_3$.

There are other bundling approaches such as the *recursive-bifurcation-method* and the *k-means clustering method*, used in [20]. For our specific multi-dimensional problems, however, using the recursive-bifurcation-method will give rise to too few paths within some bundles when the correlation parameter ρ is close to 1 or -1, no matter how large the total number of paths is. This problem will not occur if we use the equal-number bundling technique. In addition, it is easy to implement and fast for computation compared to the k-means clustering method.

5 Numerical Tests

In this section, we will analyze the convergence and accuracy of SGBM for the Heston and the HHW models, respectively w.r.t. the following quantities:

- the value of the option at time $t = 0$;
- the EE and PFE quantities over time $[0, T]$;
- the sensitivities w.r.t. S_0 of the EE function over time $[0, T]$.

The convergence of SGBM for the computation of Bermudan options can be checked by comparing the direct and path estimators. The reference values for European and barrier options can be computed by averaging discounted cash flows for a very large number of paths.

In addition, the COS method can be connected to the MC method [31] for reference values. Under the Heston model, the COS method in [14] can be used to calculate option values and corresponding Greeks at time $t = 0$ for Bermudan and barrier options. By the MC COS method exposure profiles, quantities and sensitivities of the EE function can be computed at monitoring date t_m. We use quantities computed by the COS method as the reference values for EE, PFE and sensitivity functions under the Heston model.[2]

The Quadratic Exponential (QE) scheme is employed for accurate simulation of the Heston volatility model [1]. CVA is computed here via formula (4) with $\text{LGD} = 1$. The survival probability function defined in (2) is assumed to the independent of exposure with a constant intensity $h_t = 0.03$ in the period $[0, T]$.

5.1 The Heston Model

The parameters for the Heston model in (14) are chosen as

Test A: $S_0 = 100, r = 0.04, K = 100, T = 1; \kappa = 1.15, \gamma = 0.39, \bar{v} = 0.0348$, $v_0 = 0.0348, \rho_{x,v} = -0.64$, where the Feller condition is not satisfied.

We choose a large number of MC paths, $N = 2 \cdot 10^6$ and a relatively small time step size $\Delta t = 0.05$. The paths will be clustered into $J_1 = 2^j, J_2 = 2^j, j = 1, 2, 3, 4$ bundles. The monomial basis in SGBM is of order $p = \{1, 2, 3\}$. The number of paths is chosen large as we wish to compare the convergence and accuracy using the same set of simulated scenarios for different choices of the number of bundles J and degree p. The number of paths can be greatly reduced in real-life CVA computations because SGBM typically exhibits low variances compared to LSM.

We consider a Bermudan put option under the Heston model with parameter Test A, with 10 equally-spaced exercise dates till $T = 1$.

Figure 2a shows that the direct and path estimators converge to the option value when increasing the number of bundles (J) and the order of the monomial basis (p), as expected. Monomial basis $p = 3$ enhances the convergence speed compared to

[2] In the MC COS method, we use 400 Fourier terms, and 400 grid points in volatility direction; the COS parameter for the integration domain size is set to $L = 12$ for calculating the reference values.

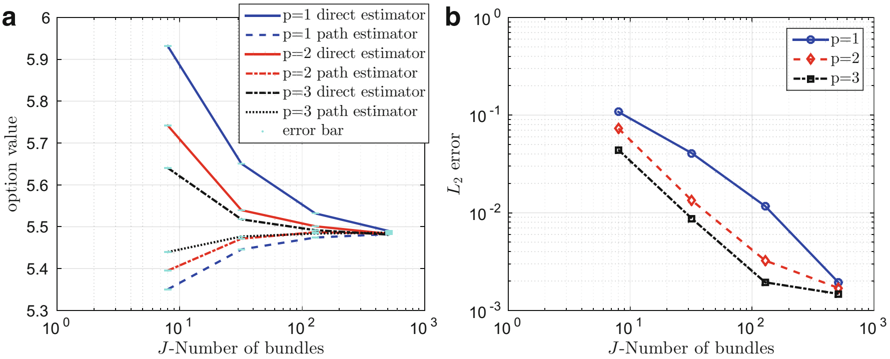


Fig. 2 Convergence of the Bermudan option value and the EE w.r.t. J—the number of bundles and p—the order of the basis functions, by comparing the direct and path estimators. Strike $K = 100$, expiry date $T = 1$ and exercise times 10. The total number of paths $N = 2 \cdot 10^6$. (**a**) Bermudan option. (**b**) Error in EE

$p = 2$ or $p = 1$. Figure 2b confirms this by showing the difference in the computed EE of the direct and path estimators, where the difference is measured in the relative L_2 norm.[3]

In Fig. 3, we present the accuracy of SGBM for the exposure quantities, EE, PFE and sensitivities of EE, by comparing to reference values by the MC COS method based on the same set of MC paths. Increasing the number of bundles J and/or the order of the monomial basis p enhances the accuracy of the results, as expected. In particular, a basis of order $p = 2$ achieves the same level of accuracy as order $p = 3$ with twice more bundles. By increasing the number of bundles, we can thus employ a monomial basis of lower order, which is an important insight.

Table 2 presents option values as well as CVA and sensitivities computed by SGBM plus the corresponding reference values. We see that the direct estimators have smaller variances compared to the path estimators.

In addition, Fig. 4 demonstrates the convergence of SGBM based on basis functions of lower order, $p = 1$, where we increase the number of bundles to 4^6. The conclusion in Appendix 4, i.e. when the size of a bundle approaches zero, the bias caused by approximating a continuous function by a simple linear function goes to zero, is confirmed. This is one advantage of SGBM compared to LSM. We need fewer basis functions by using bundles.

[3] The relative L_2 norm is defined by:

$$\frac{\|EE_{direct} - EE_{path}\|_2}{\|EE_{direct}\|_2} = \frac{\sqrt{\sum_{m=0}^{M} \left(\mathrm{EE}_{direct}(t_m) - \mathrm{EE}_{path}(t_m)\right)^2}}{\sqrt{\sum_{m=0}^{M} \left(\mathrm{EE}_{direct}(t_m)\right)^2}}. \tag{55}$$

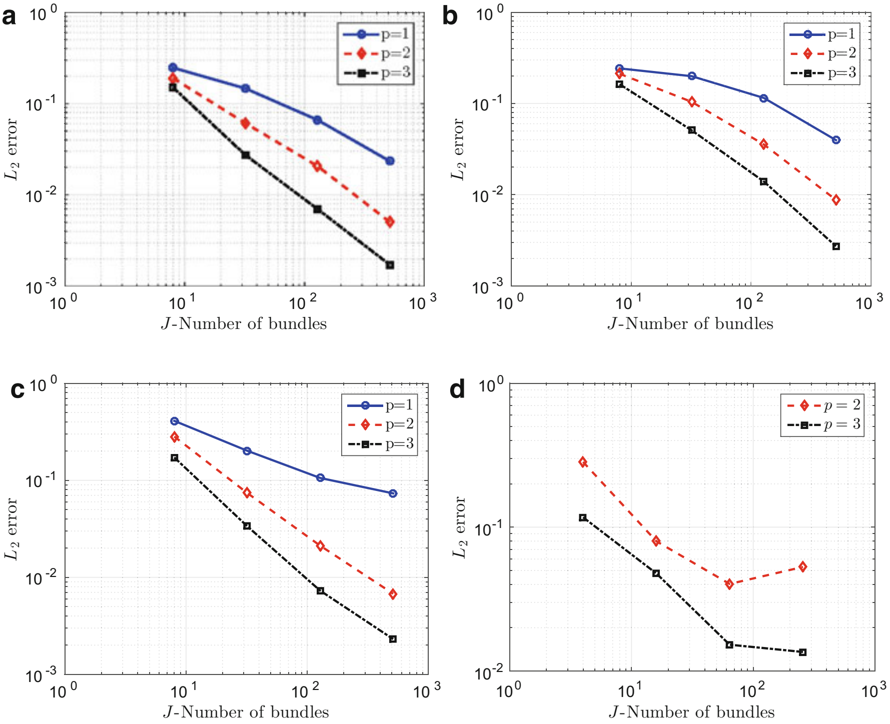


Fig. 3 Convergence of the EE, PFE and sensitivities, w.r.t. J—the number of bundles and p—the order of basis functions for a Bermudan put option; the reference is generated by the MC COS method. Strike $K = 100$, expiry date $T = 1$ and exercise times 10. The total number of paths $N = 2 \cdot 10^6$. (**a**) Error in EE. (**b**) Error in PFE. (**c**) Error in Δ_{EE}. (**d**) Error in Γ_{EE}

Table 2 Results of a Bermudan put option under the Heston model. Strike $K = 100$, expiry date $T = 1$ and exercise times 10. The total number of paths $N = 2 \cdot 10^6$, and the order $p = 2$ and the bundle number $J = 2^8$

Bermudan option under the Heston model			
Quantities	Direct estimator (std.)	Path estimator (std.)	COS
$V(0)$	5.486(0.000)	5.488 (0.005)	5.486
$\Delta_{\text{EE}}(0)$	−0.329(0.000)	–	−0.328
$\Gamma_{\text{EE}}(0)$	0.022(0.000)	–	0.025
CVA	0.093(0.000)	0.093 (0.000)	0.093 (0.000)

We also consider a put-down-out barrier option with strike $K = 100$. The option is knocked out when the asset value reaches barrier level $H = 0.9K$ before the maturity $T = 1$. After being knocked out, an investor receives a rebate value, $r_b = 10$; otherwise the investor receives the payoff at time $T = 1$. We present these quantities computed by SGBM and the corresponding reference values in Table 3.

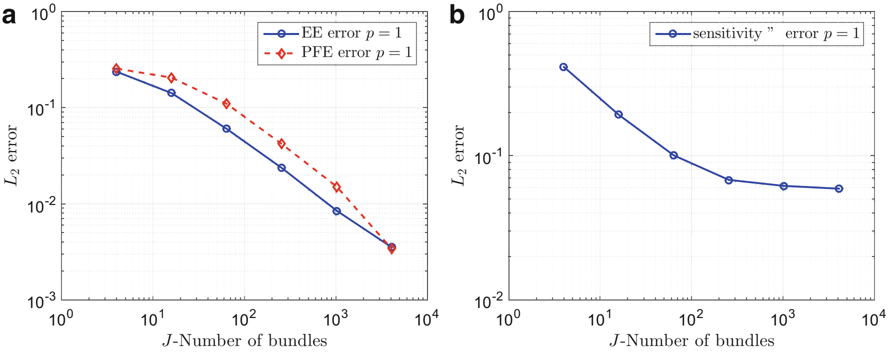


Fig. 4 Convergence of the EE, PFE and sensitivity Δ w.r.t. J—the number of bundles for a Bermudan put option when the number of paths within each bundle is 200, the order of the basis functions $p = 1$, and the total number of paths is $200J$; the reference is generated by the MC COS method. (**a**) EE and PFE when $p = 1$. (**b**) Sensitivity Δ when $p = 1$

Table 3 Results of a down-and-out barrier put option under the Heston model. Strike $K = 100$, expiry date $T = 1$, barrier level $H = 0.9K$, $r_b = 10$. The total number of paths $N = 2 \cdot 10^6$, and the order $p = 2$ and the bundle number $J = 2^8$

Barrier option under the Heston model			
Values $t = 0$	SGBM (std.)	Monte Carlo (std.)	COS
$V(0)$	4.013 (0.000)	4.016 (0.003)	4.015
$\Delta_{\text{EE}}(0)$	−0.2631 (0.000)	–	−0.263
$\Gamma_{\text{EE}}(0)$	0.0232 (0.000)	–	0.0224
CVA	0.0493 (0.000)	0.0493 (0.000)	0.0493 (0.000)

5.2 *The HHW Model*

SGBM for the Heston Hull-White model is based on forward simulation under the true HHW dynamics while the backward computation employs the discounted moments of the H1HW dynamics.There are basically two issues regarding the SGBM computation of exposure under the HHW model. We will focus on the impact of a long expiry date (say $T = 10$), and we will examine the accuracy of the approximation of the HHW model by the affine H1HW model.

We use the following parameters for the HHW and H1HW models (14):

Test B: $S_0 = 100$, $v_0 = 0.05$, $r_0 = 0.02$; $\kappa = 0.3$, $\gamma = 0.6$, $\bar{v} = 0.05$, $\lambda = 0.01$, $\eta = 0.01$, $\theta = 0.02$, $\rho_{x,v} = -0.3$ and $\rho_{x,r} = 0.6$. $T = 10$.

Simulation is done with $N = 10^6$ MC paths and $\Delta t = 0.1$. The details of the SGBM algorithm are as follows: the number of bundles varies as $J_1 = 2^{2+j}$, $J_2 = 2^j$, $J_3 = 2^j$, $j = 1, 2, 3$ and the orders of the monomial basis are $p = \{1, 2\}$.

Table 4 Implied volatility (%) obtained for a European put option with expiry date $T = 10$ under the HHW model, based on 5 simulations

Implied volatility (%)			
K/S_0	SGBM (std.)	Monte Carlo (std.)	Abs. error (%)
40%	26.481 (0.003)	26.479 (0.03)	0.0014
80%	20.699 (0.003)	20.719 (0.02)	0.0202
100%	19.200 (0.003)	19.242 (0.01)	0.0413
120%	18.369 (0.003)	18.427 (0.01)	0.0585
180%	18.220 (0.003)	18.291 (0.02)	0.0706

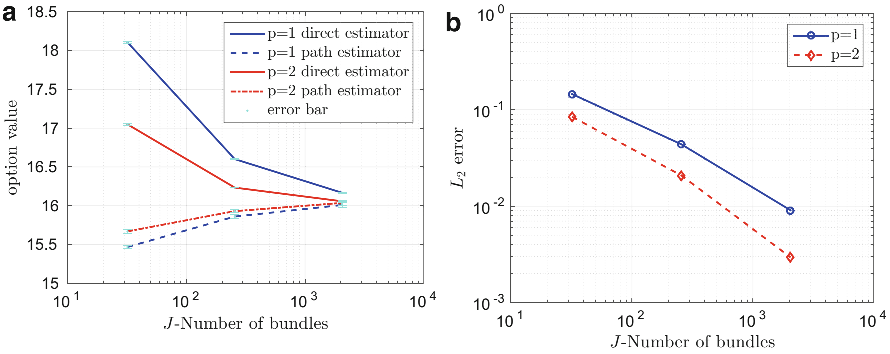


Fig. 5 Convergence w.r.t. J—the number of bundles and p—the order of the monomial basis by comparing the path and the direct estimator under the HHW model. Strike $K = 100$, $T = 10$ and 50 exercise times. The total number of paths $N = 10^6$. (**a**) Bermudan option values. (**b**) EE difference

The accuracy of SGBM is first studied by computing a European put option with $T = 10$. The implied volatility (in %) is used to demonstrate the accuracy of the computed option values, as the implied volatility is typically sensitive to the accuracy of option values [17]. The implied volatility is computed by means of the BS formula for strike values $K = \{40, 80, 100, 120, 180\}$. The reference values are computed by the average cash flows on the generated MC paths. The results are presented in Table 4. The SGBM results have smaller variances compared to results of a plain Monte Carlo simulation, and maintain a high accuracy when comparing the absolute errors.

We then consider a Bermudan put option with 50 exercise dates equally distributed in the period $[0, T]$. Figure 5 shows the SGBM convergence rate by comparing the direct and path estimators. Results of this Bermudan put are presented in Table 5. Table 6 presents results of SGBM for computing a down-and-out barrier put option. It shows that SGBM works well also for a non-continuous payoff function.

Table 5 Results for a Bermudan put option under the HHW model. Strike $K = 100$, $T = 10$ and 50 exercise times. The total number of paths is $N = 10^6$, and the order $p = 2$ and bundle number $J = 2048$

Bermudan option under the HHW model			
$T = 10$	Values $t = 0$	Direct estimator (std.)	Path estimator (std.)
	$V(0)$	16.056(0.002)	16.009 (0.018)
	$\varDelta_{\mathrm{EE}}(0)$	−0.268(0.000)	–
	$\varGamma_{\mathrm{EE}}(0)$	0.815(0.001)	–
	CVA	2.968(0.003)	–

Table 6 Results for a down-and-out barrier put option under the HHW model. Strike $K = 100$, $T = 10$, barrier level $H = 0.9K$, $r_b = 0$. The total number of paths is $N = 10^6$, and the order $p = 2$ and $J = 2048$ bundles

Barrier option under the HHW model		
Values $t = 0$	Direct estimator (std.)	Monte Carlo (std.)
$V(0)$	0.0478(0.000)	0.0477 (0.001)
$\varDelta_{\mathrm{EE}}(0)$	0.0017(0.000)	–
$\varGamma_{\mathrm{EE}}(0)$	−0.0001(0.000)	–
CVA	0.0123(0.000)	–

Table 7 Calculation time in seconds for computing exposure profiles of a Bermudan option and for that of a whole portfolio with expiry date $T = 10$ under the HHW model; SGBM with polynomial order $p = 2$, number of paths $N = 10^6$ and time step size $\varDelta t = 0.1$

Calculation time	Direct estimator	Path estimator for Bermudan
A single (Bermudan) option	151.5 (sec.)	130.2 (sec.)
Portfolio	306.3 (sec.)	131.5 (sec.)

5.3 *Speed*

One benefit of the SGBM algorithm is that one can calculate different financial derivatives on the same underlying in one backward iteration using the same set of simulated paths, as the monomial basis and the discounted moments are the same. Table 7 compares the calculation time of a single Bermudan option and of a portfolio, that consists of a Bermudan option, a European option and two barrier options with the same underlying stock. The algorithm is implemented in MATLAB, and runs on an Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz.

By using parallelization of the SGBM algorithm, the speed can be further enhanced drastically, see a study in [26].

Table 8 CVA(%) of European options with $T = 5$ and strike values $K = \{80, 100, 120\}$

European option, CVA (%)					
	K/S_0	BS	Heston	BSHW	HHW
$T = 1$	80%	2.951 (0.010)	2.959 (0.003)	2.953 (0.005)	2.949 (0.005)
	100%	2.956 (0.011)	2.958 (0.003)	2.952 (0.002)	2.952 (0.002)
	120%	2.955 (0.002)	2.959 (0.001)	2.953 (0.001)	2.952 (0.001)
$T = 5$	80%	13.925 (0.036)	13.941 (0.021)	13.882 (0.016)	13.929 (0.027)
	100%	13.951 (0.039)	13.960 (0.010)	13.901 (0.003)	13.940 (0.018)
	120%	13.919 (0.010)	13.953 (0.007)	13.901 (0.005)	13.936 (0.010)

5.4 Impact of Stochastic Volatility and Stochastic Interest Rates

We here check the impact of stochastic volatility and stochastic interest rates on exposure profiles and CVA. Next to the already discussed Heston and HHW models, we also consider the Black-Scholes (BS) and the Black-Scholes Hull-White (BSHW) models in this section. The parameter set chosen is the same as in Test B. For comparison, we use the parameters of the other models such that we can ensure that the values of a European put option with a fixed expiry date T has the same price under all models.[4]

We define a so-called CVA percentage as $\left(100 \cdot \frac{\text{CVA}}{V(0)}\right)\%$. Table 8 presents the percentage CVA for European put options with two maturity times, $T = \{1, 5\}$, for the strike values $K = \{80, 100, 120\}$. It can be seen that the CVA percentage does not change with strike; furthermore, European options with maturity $T = 5$ exhibit a higher CVA percentage than those with maturity $T = 1$. Based on the chosen parameters, we see only a small impact of stochastic volatility and stochastic interest rate on the CVA percentage.

Table 9 presents the percentage CVA for Bermudan put options with maturity times $T = \{1, 5\}$ for strike values $K = \{80, 100, 120\}$. We see that the 'in-the-money' options have the smallest CVA percentage. This is understandable as the optimal exercise strategy, in this paper, does not take into account the risk of a counterparty default. A put option is likely to be exercised before maturity when the strike value is higher than the current stock value, and thus one can expect relatively little exposure.

Figure 6 presents the EE and PFE function values w.r.t. time for a Bermudan put option which is at-the-money.

[4] For example, under the Black-Scholes model, we use the implied interest rate, i.e. $r_T = -\frac{\log(p(0,T))}{T}$, and compute the implied volatility by the analytic BS formula. Under the Heston model, the parameters of the Heston process are the same as those in Test B, and the corresponding interest rate is computed by the bisection algorithm. Under the BSHW model, the parameters of the Hull White process are the same as those in Test B, and the corresponding volatility is determined.

Table 9 CVA(%) of Bermudan options with $T = 5$ and strike values $K = \{80, 100, 120\}$

Bermudan option, CVA (%)					
	K/S_0	BS	Heston	BSHW	HHW
$T = 1$	80%	2.534 (0.007)	2.460 (0.002)	2.643 (0.003)	2.504 (0.003)
	100%	2.005 (0.003)	1.939 (0.002)	2.165 (0.001)	2.016 (0.001)
	120%	0.906 (0.002)	1.031 (0.001)	0.986 (0.001)	1.068 (0.001)
$T = 5$	80%	10.110 (0.032)	9.876 (0.030)	12.612 (0.014)	10.890 (0.029)
	100%	7.784 (0.011)	8.120 (0.012)	10.965 (0.008)	9.649 (0.019)
	120%	4.453 (0.008)	4.416 (0.020)	6.923 (0.005)	6.259 (0.013)

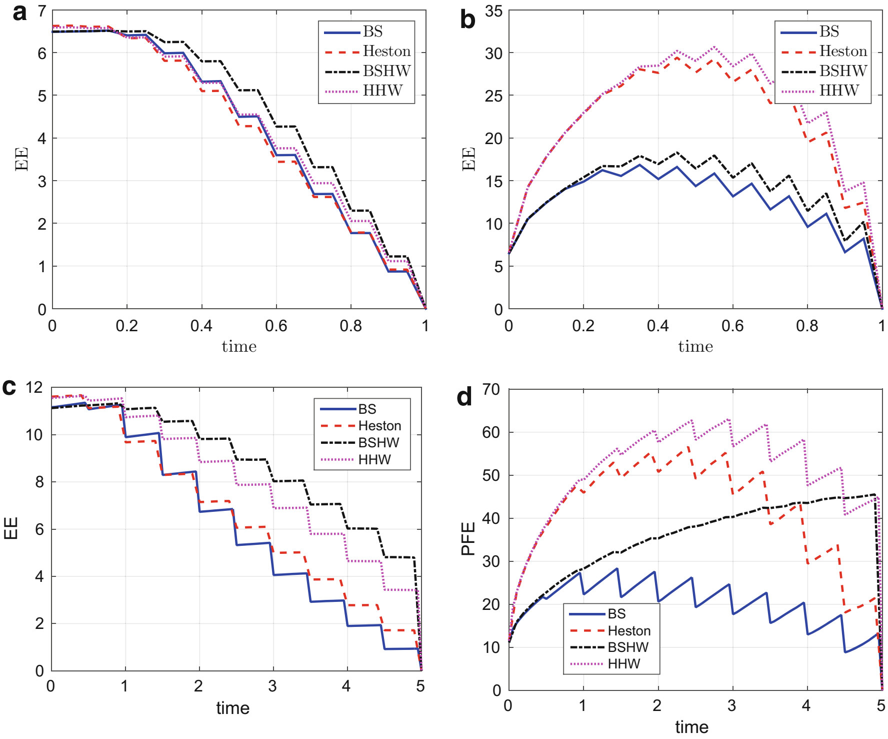


Fig. 6 Impact of stochastic volatility and interest rate on EE and PFE with different tenors and different asset dynamics, at the money $K = 100$. (**a**) T=1, EE. (**b**) T=1, PFE . (**c**) T=5, EE. (**d**) T=5, PFE

- In Fig. 6b, it can be seen that the PFE values for the HHW model are relatively close to those of the Heston model, and the PFE values for the BSHW model are very similar to those of the BS model. With a short time to maturity ($T = 1$), under our model assumptions and parameters, the stochastic volatility has a more significant contribution to the PFE values compared to the stochastic interest rate.

Compared to Fig. 6a, we can see that the EE values for the Heston and the BS models are very close. Adding stochastic volatility has more impact on the right-side tails of the exposure profiles than on the EE values.

- In Fig. 6d, in the period $t = [0, 1]$, we see similarities of PFE values between the HHW and the Heston models, and between the BS and the BSHW models; in the period $t = [1, 5]$, the PFE values for the BSHW model tend to be higher than those of the BS model, and the PFE values for the HHW model are also higher than those of the Heston model. Clearly, interest rates have more impact on the exposure profiles in the longer term (say $T = 5$).
- Figures 6a and c show that the stochastic interest rate increases the future EE values of Bermudan options, while the stochastic volatility has the opposite effect.
- The PFE curve for the BSHW model in Fig. 6d looks differently from the other curves because of the positive correlation parameter ($\rho_{x,r} = 0.6$) and the long expiry ($T = 5$). The PFE curve represents events with large option values and for a put option, this means that the associated stock values are low. In the case of a positive correlation parameter $\rho_{x,r}$, the interest rate is low as well. The investor likely holds on to the option. If we set the correlation value to zero in the BSHW model and perform the same computation, the PFE curves under the BSHW model becomes 'spiky' as well.

The stochastic interest rate plays a significant role in the case of a longer maturity derivatives, and results in increasing PFE profiles; stochastic asset volatility appears to have an effect on PFE values at the early stage of a contract. Under the parameters chosen here, at an early stage of the contract (say $t < 1$), the PFE profiles under the HHW model are very similar to those under the Heston model, but at later contract times the PFE profiles under the HHW model increase. It seems that the stochastic volatility has more effect on the right-side tail compared to the expectation of the exposure profile, while adding the stochastic interest rate increases the whole exposure profile, especially in the case of a longer maturity.

6 Conclusion

In this paper we generalize the Stochastic Grid Bundling Method (SGBM) towards the computation of exposure profiles and sensitivities for asset dynamics with stochastic asset volatility and stochastic interest rate for European, Bermudan as well as barrier options. The algorithmic structure as well as the essential method components are very similar for CVA as for the computation of early-exercise options, which makes SGBM a flexible CVA valuation framework.

We presented arguments for the choice of the basis functions for the local regression, presented a bundling technique, and showed SGBM convergence of the direct and path estimators with respect to an increasing number of bundles. Numerical experiments demonstrate SGBM's convergence and accuracy.

Using higher-order polynomials as the basis functions is especially important when accurate sensitivities values are needed; otherwise, a polynomial order $p = 1$ is sufficient for option prices and exposure quantities with a sufficiently large number of bundles and paths. The computational efficiency is connected to the number of bundles used in SGBM. A parallel algorithm will be important for a drastic reduction of the computation times, see the studies in [26].

Acknowledgements We thank Lech A. Grzelak, Shashi Jain, Kees de Graaf and Drona Kandhai for helpful discussions, as well as the CVA team at ING bank. Financial support by the Dutch Technology Foundation STW (project 12214) is gratefully acknowledged.

Appendix 1: The Joint Discounted ChF of the Heston Model

The discounted ChF of an affine model can be derived by Ricatti ODEs, as presented by Duffie et al. [13]. The expression for the joint dChF of the Heston model is given by:

$$\Phi_{\text{Heston}}(u_1, u_2, T|\mathbf{Y}_t) = \exp\left(\bar{A}_H(u_1, u_2, \tau) + \bar{B}_H(u_1, \tau)x_t + \bar{C}_H(u_1, u_2, \tau)v_t\right), \quad (56)$$

where the coefficients of the ChF are obtained via the following ODEs:

$$\frac{d\bar{B}_H}{d\tau}(u_1, \tau) = 0, \quad (57)$$

$$\frac{d\bar{C}_H}{d\tau}(u_1, u_2, \tau) = \bar{B}_H(\tau)(\bar{B}_H(\tau) - 1)/2 \quad (58)$$

$$+ \left(\gamma\rho_{x,v}\bar{B}_H(\tau) - \kappa\right)\bar{C}_H(\tau) + \gamma^2\bar{C}_H^2(\tau)/2, \quad (59)$$

$$\frac{d\bar{A}_H}{d\tau}(u_1, u_2, \tau) = \kappa\bar{v}\bar{C}_H(\tau) + r(\bar{B}_H(\tau) - 1), \quad (60)$$

where $\tau = T - t$ and initial condition $\bar{B}_H(u_1, \tau = 0) = iu_1$, $\bar{C}_H(u_1, u_2, \tau = 0) = iu_2$ and $\bar{A}_H(u_1, u_2, \tau = 0) = 0$. The solution is given by:

$$\bar{B}_H(u_1, \tau) = iu_1, \quad (61)$$

$$\bar{C}_H(u_1, u_2, \tau) = r_+ - \frac{2D_1}{\gamma^2\left(1 - ge^{-D_1\tau}\right)}, \quad (62)$$

$$\bar{A}_H(u_1, u_2, \tau) = I_1^H + I_2^H, \quad (63)$$

with

$$g = \frac{iu_2 - r_-}{iu_2 - r_+}, D_1 = \sqrt{(\kappa - \gamma\rho_{x,v}iu_1)^2 + \gamma^2 u_1(u_1 + i)}, \quad (64)$$

$$r_{\pm} = \frac{1}{\gamma^2}\left(\kappa - \gamma\rho_{x,v} i u_1 \pm D_1\right), \tag{65}$$

and

$$I_1^H = \kappa\bar{v}\left(r_-\tau - \frac{2}{\gamma^2}\log\left(\frac{1 - ge^{-D_1\tau}}{1-g}\right)\right), \tag{66}$$

$$I_2^H = r(iu_1 - 1)\tau. \tag{67}$$

The form of the characteristic function in Heston's original paper [18] is problematic due to branch cuts. A more recent reference is [15]. We use the correct form of the characteristic function in our numerical examples.

Appendix 2: The Joint Discounted ChF of the Black-Scholes Hull-White Model

The expression for the joint dChF for the BSHW model is given by:

$$\Phi_{\text{BSHW}}(u_1, u_3, T|\mathbf{Y}_t) = \exp\left(\bar{A}_S(u_1, u_3, \tau) + \bar{B}_S(u_1, \tau)x_t + \bar{D}_S(u_1, u_3, \tau)r_t\right), \tag{68}$$

where the coefficients of the ChF are obtained via the following ODEs:

$$\frac{d\bar{B}_S}{d\tau}(u_1, \tau) = 0, \tag{69}$$

$$\frac{d\bar{D}_S}{d\tau}(u_1, u_3, \tau) = -1 + \bar{B}_S(u_1, \tau) - \lambda\bar{D}_S(u_1, u_3, \tau), \tag{70}$$

$$\begin{aligned}\frac{d\bar{A}_S}{d\tau}(u_1, u_3, \tau) = {} & \frac{1}{\sigma^2}\bar{B}_S(u_1, \tau)\left(\bar{B}_S(u_1, \tau) - 1\right) + \lambda \cdot \theta(T - \tau) \cdot \bar{D}_S(u_1, u_3, \tau) \\ & + \frac{1}{2}\eta^2\bar{D}_S(u_1, u_3, \tau) + \rho_{x,r}\sigma\eta\bar{B}_S(u_1, \tau)\bar{D}_S(u_1, u_3, \tau),\end{aligned} \tag{71}$$

where $\tau = T - t$ and initial condition $\bar{B}_S(u_1, \tau = 0) = iu_1, \bar{D}_S(u_1, u_3, \tau = 0) = iu_3$, and $\bar{A}_S(u_1, u_3, \tau = 0) = 0$. The solution is now given by:

$$\bar{B}_S(u_1, \tau) = iu_1, \tag{72}$$

$$\bar{D}_S(u_1, u_3, \tau) = \frac{iu_1 - 1}{\lambda}\left(1 - e^{-\lambda\tau}\right) + iu_3e^{-\lambda\tau}, \tag{73}$$

$$\bar{A}_S(u_1, u_3, \tau) = I_1^S + I_2^S + I_3^S + I_4^S, \tag{74}$$

with

$$I_1^S = \frac{1}{2}\sigma^2 iu_1(iu_1 - 1)\tau, \tag{75}$$

$$I_2^S = \int_0^\tau \theta(T-s)\cdot \bar{D}_S(u_1, u_3, s)ds \tag{76}$$

$$\begin{aligned} I_3^S = \frac{\eta^2}{2\lambda^2}\bigg(&\frac{2}{\lambda}(u_1+i)(e^{-\lambda\tau}-1)(\lambda u_3 - u_1 - i) + \\ &\frac{1}{2\lambda}\left(e^{-2\lambda\tau}-1\right)(\lambda u_3 - u_1 - i)^2 - (u_1+i)^2\tau\bigg), \end{aligned} \tag{77}$$

$$I_4^S = \frac{\eta\theta\sigma\rho_{x,r}}{\lambda}\left(-\frac{iu_1 + u_1^2}{\lambda}(\lambda\tau + e^{-\lambda\tau} - 1) + u_1u_3(e^{-\lambda\tau}-1)\right). \tag{78}$$

When $\theta(t) = \theta$ is a constant,

$$I_2^S = \theta\left((iu_1 - 1)\tau + \frac{1}{\lambda}(e^{-\lambda\tau}-1)(iu_1-1) - iu_3\left(e^{-\lambda\tau}-1\right)\right). \tag{79}$$

Again the discounted moments are obtained by symbolic computations in MATLAB.

Appendix 3: The Joint Discounted ChF of the H1HW Model

The expression for the joint dChF of the H1HW model is given by:

$$\begin{aligned} \Phi_{\text{H1HW}}(u_1, u_2, u_3, T|\mathbf{Y}_t) = \exp\big(&\bar{A}_W(u_1, u_2, u_3, \tau) + \bar{B}_W(u_1, \tau)x_t + \bar{C}_W(u_1, u_2, \tau)v_t \\ &+ \bar{D}_W(u_1, u_3, \tau)r_t\big), \end{aligned} \tag{80}$$

where the coefficients of the ChF are here obtained via the following ODEs:

$$\frac{d\bar{B}_W}{d\tau}(u_1, \tau) = 0, \tag{81}$$

$$\begin{aligned} \frac{d\bar{C}_W}{d\tau}(u_1, u_2, \tau) = &\bar{B}_W(\tau)(\bar{B}_W(\tau) - 1)/2 + \left(\gamma\rho_{x,v}\bar{B}_W(\tau) - \kappa\right)\bar{C}_W(\tau) \\ &+\gamma^2\bar{C}_W^2(\tau)/2, \end{aligned} \tag{82}$$

$$\frac{d\bar{D}_W}{d\tau}(u_1, u_3, \tau) = -1 + \bar{B}_W(u_1, \tau) - \lambda\bar{D}_W(u_1, u_3, \tau), \tag{83}$$

$$\frac{d\bar{A}_W}{d\tau}(u_1, u_2, u_3, \tau) = \lambda\cdot\theta(T-\tau)\cdot\bar{D}_W(u_1, u_3, \tau) + \kappa\bar{v}\bar{C}_W(u_1, u_2, \tau)$$

$$+\frac{1}{2}\eta^2\bar{D}_W^2(u_1,u_3)+\eta\rho_{x,v}\mathbb{E}\left[\sqrt{v_T}\middle|v_t\right]\bar{B}_W(u_1,\tau)\bar{D}_W$$
$$\times(u_1,u_3,\tau), \tag{84}$$

where $\tau = T-t$ and initial condition $\bar{B}_W(u_1,\tau=0)=iu_1$, $\bar{C}_W(u_1,u_2,\tau=0)=iu_2$, $\bar{D}_W(u_1,u_3,\tau=0)=iu_3$ and $\bar{A}_W(u_1,u_2,u_3,\tau=0)=0$. The solution is given by:

$$\bar{B}_W(u_1,\tau) = iu_1, \tag{85}$$

$$\bar{C}_W(u_1,u_2,\tau) = r_+ - \frac{2D_1}{\gamma^2\left(1-ge^{-D_1\tau}\right)}, \tag{86}$$

$$\bar{D}_W(u_1,u_3,\tau) = \frac{iu_1-1}{\lambda}\left(1-e^{-\lambda\tau}\right)+iu_3e^{-\lambda\tau}, \tag{87}$$

$$\bar{A}_W(u_1,u_2,u_3,\tau) = I_1^W+I_2^W+I_3^W+I_4^W, \tag{88}$$

where expressions g, D_1 and $r_\pm$ are the same as in (64), and

$$I_1^W = \int_0^\tau \theta(T-s)\cdot\bar{D}_W(u_1,u_3,s)ds, \tag{89}$$

$$I_2^W = \kappa\bar{v}\left(r_-\tau-\frac{2}{\gamma^2}\log\left(\frac{1-ge^{-D_1\tau}}{1-g}\right)\right), \tag{90}$$

$$I_3^W = \frac{\eta^2}{2\lambda^2}\left(\frac{2}{\lambda}(u_1+i)(e^{-\lambda\tau}-1)(\lambda u_3-u_1-i)\right. \tag{91}$$

$$\left.+\frac{1}{2\lambda}\left(e^{-2\lambda\tau}-1\right)(\lambda u_3-u_1-i)^2-(u_1+i)^2\tau\right), \tag{92}$$

$$I_4^W = \eta\rho_{x,r}\left(-\frac{iu_1+u_1^2}{\lambda}G_1(\tau,v_t)-u_1u_3G_2(\tau,v_t)\right), \tag{93}$$

where

$$G_1(\tau,v_t) := \int_0^\tau \mathbb{E}\left[\sqrt{v_{T-x}}\middle|v_t\right]\left(1-e^{-\lambda x}\right)dx, \tag{94}$$

$$G_2(\tau,v_t) := \int_0^\tau \mathbb{E}\left[\sqrt{v_{T-x}}\middle|v_t\right]e^{-\lambda x}dx. \tag{95}$$

When $\theta(t)=\theta$ is a constant, I_1 can be integrated by

$$I_1^W = \theta\left((iu_1-1)\tau+\frac{1}{\lambda}(e^{-\lambda\tau}-1)(iu_1-1)-iu_3\left(e^{-\lambda\tau}-1\right)\right). \tag{96}$$

It is computationally expensive to calculate the integral for G_1 and G_2 over $[t, t+\tau]$. We use an approximation where, for a fixed v_t, values of the conditional expectation $\mathbb{E}\left[\sqrt{v_{t+\tau}}\middle|v_t\right]$ over a short time period can be approximated by a linear function w.r.t. time.

We will use the approximation that

$$\mathbb{E}\left[\sqrt{v_{t+\tau}}\middle|v_t\right] \approx a(v_t) + b(v_t, \Delta t)\tau, \quad \tau \leq \Delta t, \tag{97}$$

where $a(v_t) = \sqrt{v_t}$, $b(v_t, \Delta t) = \frac{v_{(t+\Delta t)} - v_t}{\Delta t}$, $\Delta t = 0.05$. Various experiments have shown that this approximation is sufficiently accurate in the present context.

The integrals expressed in (94) and (95) can be approximated by an analytic formula with the approximation in (97). To further enhance the of SGBM, we compute the integrals on a volatility grid based on the minimum and maximum values of the variance on the simulated paths. At each time step t_m, the discounted moments on all paths are computed with the help of the volatility grid plus a spline interpolation technique.

Appendix 4: Errors of Approximation of the Option Function

There are two types of errors when approximating the option function on the bounded domain $\mathbf{U}^j_{m+1}$ at time t_{m+1}. The first type of error ϵ_1 is the difference between the real option function and its projection on the polynomial space $\mathscr{P}(\mathbf{U}^j_{m+1}, p)$, and the second type of error ϵ_2 is the difference between the real projection on the polynomial space and its statistical approximation given a data set $\{\hat{v}_{m+1}(i), \hat{\mathbf{y}}_{m+1}(i)\}$. Measured in L_2 norm within the j-bundle, these two errors can be expressed by

$$\epsilon_1 = \|V(t_{m+1}, \cdot) - Z_1(t_{m+1}, \cdot)\|_{L_2}, \tag{98}$$

$$\epsilon_2 = \|Z_1(t_{m+1}, \cdot) - Z_2(t_{m+1}, \cdot)\|_{L_2}. \tag{99}$$

where the L_2 norm is defined by the conditional probability measure $\mu_{(\mathbf{Y}_{m+1}|\mathbf{Y}_m)}$, i.e. for any L_2 measurable function $f(\mathbf{Y}_{m+1})$, its L_2 norm is defined by [29]

$$\|f\|_{L_2} = \left(\int_{\mathbf{Y}_{m+1} \in \mathbb{R}^n} |f(\mathbf{Y}_{m+1})|^2 d\mu_{(\mathbf{Y}_{m+1}|\mathbf{Y}_m)}\right)^{\frac{1}{2}}. \tag{100}$$

It is trivial to see that the total error of approximation of the option function is bounded by the sum of these two types of error, i.e.

$$\mathbb{E}^{\mathbb{Q}}\left[\left(V(t_{m+1}, \mathbf{Y}_{m+1}) - Z_2(V(t_{m+1}, \mathbf{Y}_{m+1})\right)^2 \middle| \mathbf{Y}_m\right] \leq \epsilon_1 + \epsilon_2, \tag{101}$$

and we will discuss them respectively.

- For the first type of error ϵ_1: The well-known Weierstrass approximation theorem states that any continuous function defined on a closed interval can be uniformly approximated as closely as desired by a polynomial function [24]. It can ensure that ϵ_1 will go to zero as the order of the monomial basis goes to infinity.

  More specifically, the error ϵ_1 is involved with the property of the polynomial space $\mathscr{P}(\mathbf{U}^j_{m+1}, p)$, i.e. the size of the domain $\mathbf{U}^j_{m+1}$ and the order of the monomial basis p. Theorems 1.2 in [24, p.12] and Theorem 3.2 in [24, p.59] provides a priori error estimate in L_2 norm when the function needs to be approximated is twice differentiable.

  To reduce error ϵ_1, we can either reduce the size of the domain $\mathbf{U}^j_{m+1}$ or increase the order of the basis functions. By using bundles we can achieve the former goal.
- For the second type of error ϵ_2: Assuming that the function Z_2 is an unbiased statistical estimator of Z_1, i.e.

  $$Z_1(t_{m+1}, \mathbf{Y}_{m+1}) = Z_2(t_{m+1}, \mathbf{Y}_{m+1}) + \delta_{m+1}, \tag{102}$$

  where the error term $\delta_{m+1} \sim \mathscr{N}(0, \sigma^2_{m+1})$ i.i.d, where σ^2_{m+1} is the constant variance. By central limit theorem the error satisfies with probability 1 that $\epsilon_2 \to \frac{\sigma_{m+1}}{\sqrt{N_j}}$, as the number of paths $N_j \to \infty$. It implies that the error ϵ_2 approaches zero as the number of paths goes to infinity with probability 1. Error ϵ_2 can be reduced by increasing the number of paths N_j.

As a conclusion, the SGBM approach converges as the number of bundles, the number of paths within each bundle and the polynomial order p of the basis functions go to infinity.

Appendix 5: Proof of Proposition 1

Proof By Jensen's inequality:

$$\begin{aligned}
&\Big(c(t_m, \mathbf{Y}_m) - c_2(t_m, \mathbf{Y}_m)\Big)^2 \\
&\quad = \Big(\mathbb{E}^{\mathbb{Q}}\left[D(t_m, t_{m+1}) V(t_{m+1}, \mathbf{Y}_{m+1}) \middle| \mathbf{Y}_m\right] - \mathbb{E}^{\mathbb{Q}}\left[D(t_m, t_{m+1}) Z_2(t_{m+1}, \mathbf{Y}_{m+1}) \middle| \mathbf{Y}_m\right]\Big)^2 \\
&\quad \leq \mathbb{E}^{\mathbb{Q}}\left[\Big(D(t_m, t_{m+1})(V(t_{m+1}, \mathbf{Y}_{m+1}) - Z_2(t_{m+1}, \mathbf{Y}_{m+1}))\Big)^2 \middle| \mathbf{Y}_m\right] \\
&\quad \leq \mathbb{E}^{\mathbb{Q}}\left[\Big(V(t_{m+1}, \mathbf{Y}_{m+1}) - Z_2(t_{m+1}, \mathbf{Y}_{m+1})\Big)^2 \middle| \mathbf{Y}_m\right].
\end{aligned} \tag{103}$$

References

1. L. B. G. Andersen. Simple and Efficient Simulation of the Heston Stochastic Volatility Model. *Journal of Computational Finance*, 11:1–48, 2008.
2. Bank for International Settlements. Basel II: International Convergence of Capital Measurement and Capital Standards: A Revised Framework. Technical report, 2004.
3. T. R. Bielecki and M. Rutkowski. *Credit risk: Modeling, Valuation and Hedging*. Springer Science & Business Media, 2002.
4. F. Black and M. Scholes. The Pricing of Options and Corporate Liabilities. *Journal of Political Economy*, 81(3):pp. 637–654, 1973.
5. D. Brigo. Counterparty Risk FAQ: Credit VaR, PFE, CVA, DVA, Closeout, Netting, Collateral, Re-hypothecation, WWR, Basel, Funding, CCDS and Margin Lending. Papers 1111.1331, arXiv.org, Nov 2011.
6. D. Brigo and F. Mercurio. *Interest Rate Models-Theory and Practice: with Smile, Inflation and Credit*. Springer Science & Business Media, 2007.
7. M. Broadie and M. Cao. Improved Lower and Upper Bound Algorithms for Pricing American Options by Simulation. *Quantitative Finance*, 8(8):845–861, 2008.
8. M. Broadie, P. Glasserman, and Z. Ha. Pricing American Options by Simulation Using a Stochastic Mesh with Optimized Weights. In S. Uryasev, editor, *Probabilistic Constrained Optimization*, volume 49 of *Nonconvex Optimization and Its Applications*, pages 26–44. Springer US, 2000.
9. J. F. Carriere. Valuation of the Early-exercise Price for Options Using Simulations and Nonparametric Regression. *Insurance: mathematics and Economics*, 19(1):19–30, 1996.
10. F. Cong and C. W. Oosterlee. Pricing Bermudan Options under Merton Jump-Diffusion Asset Dynamics. *International Journal of Computer Mathematics, Forthcoming*, 2015.
11. J. C. Cox, J. E, Ingersoll, and S. A. Ross. A Theory of the Term Structure of Interest Rates. *Econometrica*, 53(2):385–407, 1985.
12. C. S. L. de Graaf, Q. Feng, D. Kandhai, and C. W. Oosterlee. Efficient Computation of Exposure Profiles for Couterparty Credit Risk. *International Journal of Theoretical and Applied Finance*, 17(04):1450024, 2014.
13. D. Duffie, J. Pan, and K. Singleton. Transform Analysis and Asset Pricing for Affine Jump-Diffusions. *Econometrica*, 68(6):1343–1376, 2000.
14. F. Fang and C. W. Oosterlee. A Fourier-Based Valuation Method for Bermudan and Barrier Options under Heston's Model. *SIAM Journal on Financial Mathematics*, 2(1):439–463, 2011.
15. J. Gatheral. *The Volatility Surface: A Practitioner's Guide*, volume 357. John Wiley & Sons, 2011.
16. J. Gregory. *Counterparty Credit Risk: The New Challenge for Global Financial Markets*. The Wiley Finance Series. John Wiley & Sons, 2010.
17. L. Grzelak and C. W. Oosterlee. On the Heston Model with Stochastic Interest Rates. *SIAM Journal on Financial Mathematics*, 2(1):255–286, 2011.
18. S. L. Heston. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *Review of Financial Studies*, 6(2):327–343, 1993.
19. J. Hull and A. White. Pricing Interest-Rate-Derivative Securities. *Review of financial studies*, 3(4):573–592, 1990.
20. S. Jain and C. W. Oosterlee. The Stochastic Grid Bundling Method: Efficient Pricing of Bermudan Options and Their Greeks . *Applied Mathematics and Computation*, 269:412–431, 2015.
21. R. A. Jarrow and S. M. Turnbull. Pricing Derivatives on Financial Securities Subject to Credit Risk. *Journal of Finance-New York-*, 50:53–53, 1995.
22. C. Kenyon, A. D. Green, and M. Berrahoui. Which measure for pfe?. the risk appetite measure a. *The Risk Appetite Measure A.(December 15, 2015)*, 2015.
23. D. Lando. *Credit Risk Modeling: Theory and Applications: Theory and Applications*. Princeton University Press, 2009.

24. M. G. Larson and F. Bengzon. *The Finite Element Method: Theory, Implementation, and Applications*, volume 10. Springer Science & Business Media, 2013.
25. B. Lauterbach and P. Schultz. Pricing Warrants: An Empirical Study of the Black-Scholes Model and Its Alternatives. *The Journal of Finance*, 45(4):1181–1209, 1990.
26. Á. Leitao and C. W. Oosterlee. GPU Acceleration of the Stochastic Grid Bundling Method for Early-Exercise Options. *International Journal of Computer Mathematics*, pages 1–22, 2015.
27. F. Longstaff and E. Schwartz. Valuing American Options by Simulation: A Simple Least-squares Approach. *Review of Financial Studies*, 14(1):113–147, 2001.
28. D. B. Madan and H. Unal. Pricing the Risks of Default. *Review of Derivatives Research*, 2(2–3):121–160, 1998.
29. B. Øksendal. *Stochastic Differential Equations*. Springer, 2003.
30. M. Pykhtin and S. Zhu. A Guide to Modelling Counterparty Credit Risk. *GARP Risk Review*, pages 16–22, July/August 2007.
31. Y. Shen, J. van der Weide, and J. Anderluh. A Benchmark Approach of Counterparty Credit Exposure of Bermudan Option under Lévy Process: The Monte Carlo-COS Method. *Procedia Computer Science*, 18(0):1163–1171, 2013. 2013 International Conference on Computational Science.
32. L. Stentoft. Value Function Approximation or Stopping Time Approximation: A Comparison of Two Recent Numerical Methods for American Option Pricing Using Simulation and Regression. *Journal of Computational Finance*, 18:1–56, 2010.
33. J. N. Tsitsiklis and B. Van Roy. Regression Methods for Pricing Complex American-Style Options. *Trans. Neur. Netw.*, 12(4):694–703, July 2001.
34. O. Vasicek. An Equilibrium Characterization of the Term Structure. *Journal of financial economics*, 5(2):177–188, 1977.

A Note on Independence Copula for Conditional Markov Chains

Tomasz R. Bielecki, Jacek Jakubowski, and Mariusz Niewęgłowski

Abstract Given a family $(Y^k,\ k = 1, 2, \ldots, N)$ of conditional Markov chains, we construct a conditional Markov chain $X = (X^1, \ldots, X^N)$ such that X^k, $k = 1, 2, \ldots, N$, are conditional Markov chains, which are conditionally independent given the information contained in some filtration $\mathbb{F}$, and such that for each k the conditional law of X^k coincides with the conditional law of Y^k. This is a new result that can be used to model different phenomena such as the gating behavior of multiple ion channels in a membrane patch, or credit ratings migrations.

1 Introduction

The main objective of this note is to construct the conditionally independent Markov copula, which we also call the conditionally independent multivariate Markov coupling, for a family Y^k, $k = 1, 2, \ldots, N$, of conditional Markov chains (CMCs for short). That is, to construct an N-variate conditional Markov chain $X = (X^1, \ldots, X^N)$, so that each X^k, $k = 1, 2, \ldots, N$, is a conditional Markov chain, and such that the conditional law of X^k coincides with the conditional law of Y^k, and, moreover, X^i and X^j for $i \neq j$ are conditionally independent given the information contained in some filtration $\mathbb{F}$.

T.R. Bielecki (✉)
Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616, USA
e-mail: bielecki@iit.edu

J. Jakubowski
Institute of Mathematics, University of Warsaw, Banacha 2, 02-097 Warszawa, Poland

Faculty of Mathematics and Information Science, Warsaw University of Technology,
Koszykowa 75, 00-662 Warszawa, Poland
e-mail: jakub@mimuw.edu.pl

M. Niewęgłowski
Faculty of Mathematics and Information Science, Warsaw University of Technology,
Koszykowa 75, 00-662 Warszawa, Poland
e-mail: M.Nieweglowski@mini.pw.edu.pl

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute
Communications 79, DOI 10.1007/978-1-4939-6969-2_10

Thus, the paper provides a contribution to the theory of structured dependence between conditional Markov chains. The conditioning is done with respect to σ-fields comprising the so called reference filtration, which is denoted by $\mathbb{F}$. Typically, inclusion of a reference filtration in a dynamical stochastic model is aimed at accounting for random factors that are believed to affect evolution of the processes of primary interest.

The conditionally independent copula models dependence between processes X^k, $k = 1, 2, \ldots, N$, via their dependence on the common information, but, at the same time, features conditional independence property between X^k, $k = 1, 2, \ldots, N$, when conditioning is done with respect to this information. Often this information is generated by stochastic processes.

In a very special case when the filtration $\mathbb{F}$ is trivial, the conditionally independent copula X reduces to the family of independent Markov chains X^k, $k = 1, 2, \ldots, N$. Such families have been quite extensively studied and applied in modeling of gating behavior of multiple ion channels in a membrane patch (see Dabrowski and McDonald [7], Kijima and Kijima [11], Ball and Yeo [2]).

Using the theory of CMCs and the respective conditionally independent copula, we can model conditional independence between multiple ion channels that are otherwise linked via common stochastic factors, embedded in filtration $\mathbb{F}$, which models a random environment. Ball, Milne and Yao [1] considered a special case of CMCs. They have assumed that single ion channels are independent, conditionally on some environmental process, which is a Markov chain. Thus, this note provides a theoretical foundation for generalization of the model studied in [1]. It is worth to note that Biagini, Groll and Widenmann in a recent paper [3] considered an application of CMCs for the evaluation of rational premia for unemployment insurance products. They were assuming that the processes representing employment status of individuals in a pool of employable individuals are conditionally independent CMCs.

Consequently, this note does not only provide the theoretical contribution in the area of structured dependence of conditional Markov chains, but also a potential contribution to the area of modeling of gating behavior of multiple ion channels in a membrane patch.

We close this brief introduction by noting that, in general, if the conditional Markov chains Y^k, $k = 1, 2, \ldots, N$, admit intensity processes, say Ψ^k, $k = 1, 2, \ldots, N$, then the structured dependence between Y^k, $k = 1, 2, \ldots, N$, can be modeled by appropriate perturbation of the conditionally independent Markov copula. Specifically, structured (conditionally Markovian) dependence between CMCs Y^k, $k = 1, 2, \ldots, N$, can be modeled in terms of a matrix valued stochastic intensity process, say $\Gamma_t,\ t \geq 0$, which is represented as a sum

$$\Gamma_t = \underbrace{\bigoplus_{k=1}^{N} \Psi_t^k}_{\text{conditionally independent Markov copula}} + \text{ dependence terms},$$

where the formula for $\bigoplus_{k=1}^{N} \Psi_t^k$ is given in (17). This is another good reason for the importance of the contribution of this note.

2 CMCs and Their Structured Dependence

Let us recall the basic set-up and the basic definitions from Bielecki, Jakubowski and Nieweęgłowski [5, 6].

Let $T > 0$ be a fixed finite time horizon. Let $(\Omega, \mathscr{A}, \mathbb{P})$ be the underlying complete probability space, which is endowed with two filtrations, the reference filtration $\mathbb{F} = (\mathscr{F}_t)_{t\in[0,T]}$ and another filtration $\mathbb{G} = (\mathscr{G}_t)_{t\in[0,T]}$, that are assumed to satisfy the usual conditions, i.e. they are right-continuous and complete. So, processes considered in this paper are defined on $(\Omega, \mathscr{A}, \mathbb{P})$ with the time interval $[0, T]$. Moreover, for any process U we denote by $\mathbb{F}^U$ the completed right-continuous filtration generated by this process.

In addition, we fix a finite set S, and we denote by d the cardinality of S. Without loss of generality we take $S = \{1, 2, 3, \ldots, d\}$.

Definition 2.1 An S-valued, $\mathbb{G}$-adapted càdlàg process X is called an $(\mathbb{F}, \mathbb{G})$-conditional Markov chain if for every $x_1, \ldots, x_k \in S$ and for every $0 \leq t \leq t_1 \leq \ldots \leq t_k \leq T$ it satisfies

$$\mathbb{P}(X_{t_k} = x_k, \ldots, X_{t_1} = x_1 | \mathscr{F}_t \vee \mathscr{G}_t) = \mathbb{P}(X_{t_k} = x_k, \ldots, X_{t_1} = x_1 | \mathscr{F}_t \vee \sigma(X_t)). \quad (1)$$

As in [5] we write $(\mathbb{F}, \mathbb{G})$-CMC, for short, in place of $(\mathbb{F}, \mathbb{G})$-conditional Markov chain.

Given an $(\mathbb{F}, \mathbb{G})$-CMC process X, we define its indicator process,

$$H_t^x := \mathbb{1}_{\{X_t = x\}}, \quad x \in S, \quad t \in [0, T]. \quad (2)$$

Accordingly, we define a column vector $H_t = (H_t^x,\ x \in S)^\top$, where $\top$ denotes transposition. For $x, y \in S,\ x \neq y$, we define the process H^{xy} that counts the number of transitions of X from x to y,

$$H_t^{xy} := \#\{u \leq t : X_{u-} = x \text{ and } X_u = y\} = \int_{]0,t]} H_{u-}^x dH_u^y, \quad t \in [0, T]. \quad (3)$$

Definition 2.2 We say that an $\mathbb{F}$-adapted matrix valued process $\Lambda_t = [\lambda_t^{xy}]_{x,y\in S}$ satisfying

$$\lambda_t^{xy} \geq 0, \quad \forall x, y \in S, x \neq y, \quad \text{and} \quad \sum_{y\in S} \lambda_t^{xy} = 0, \quad \forall x \in S, \quad (4)$$

is an $\mathbb{F}$-intensity matrix process for X, if the process $M = (M_t^x,\ x \in S)^\top$ defined as

$$M_t = H_t - \int_0^t \Lambda_u^\top H_u du, \quad t \in [0, T], \quad (5)$$

is an $\mathbb{F} \vee \mathbb{G}$ – local martingale with values in $\mathbb{R}^d$.

Now recall the concept of $(\mathbb{F},\mathbb{G})$-doubly stochastic Markov chain, $(\mathbb{F},\mathbb{G})$–DSMC for short, that was introduced in Jakubowski and Niewęgłowski [9].

Definition 2.3 A $\mathbb{G}$-adapted càdlàg process $X = (X_t)_{t\in[0,T]}$ is called an $(\mathbb{F},\mathbb{G})$–DSMC with state space S if for any $0 \le s \le t \le T$ and every $y \in S$ we have

$$\mathbb{P}(X_t = y \mid \mathscr{F}_T \vee \mathscr{G}_s) = \mathbb{P}(X_t = y \mid \mathscr{F}_t \vee \sigma(X_s)). \tag{6}$$

Most of the analysis done in [5] regards $(\mathbb{F},\mathbb{G})$–CMCs that are also $(\mathbb{F},\mathbb{G})$ doubly stochastic Markov chains. This is because doubly stochastic Markov chains enjoy very useful analytical properties. We recall that with any X, which is an $(\mathbb{F},\mathbb{G})$-DSMC, we associate a matrix valued random field $P = (P(s,t),\ 0 \le s \le t \le T)$, called the conditional transition probability matrix field (c–transition field), where $P(s,t) = (p_{xy}(s,t))_{x,y\in S}$ is defined by

$$p_{x,y}(s,t) = \frac{\mathbb{P}(X_t = y, X_s = x \mid \mathscr{F}_t)}{\mathbb{P}(X_s = x|\mathscr{F}_t)}\mathbb{1}_{\{\mathbb{P}(X_s=x|\mathscr{F}_t)>0\}} + \mathbb{1}_{\{x=y\}}\mathbb{1}_{\{\mathbb{P}(X_s=x|\mathscr{F}_t)=0\}}. \tag{7}$$

By [5, Proposition 4.2] we know that for any $0 \le s \le t \le T$ and for every $y \in S$ we have

$$\mathbb{P}(X_t = y \mid \mathscr{F}_T \vee \mathscr{G}_s) = \sum_{x\in S} \mathbb{1}_{\{X_s=x\}} p_{xy}(s,t). \tag{8}$$

Moreover, the $\mathbb{F}$–adapted matrix-valued process $\Gamma = (\Gamma_s)_{s\ge 0} = ([\gamma_s^{xy}]_{x,y\in S})_{s\ge 0}$ is the intensity of an $(\mathbb{F},\mathbb{G})$-DSMC X if:

1)

$$\int_{]0,T]} \sum_{x\in S} \left\| \gamma_s^{xx} \right\| ds < \infty. \tag{9}$$

2)

$$\gamma_s^{xy} \ge 0 \quad \forall x,y \in S, x \ne y, \quad \gamma_s^{xx} = -\sum_{y\in S: y\ne x} \gamma_s^{xy} \quad \forall x \in S. \tag{10}$$

3) The Kolmogorov backward equation holds: for all $v \le t$,

$$P(v,t) - \mathrm{I} = \int_v^t \Gamma_u P(u,t) du. \tag{11}$$

4) The Kolmogorov forward equation holds: for all $v \le t$,

$$P(v,t) - \mathrm{I} = \int_v^t P(v,u) \Gamma_u du. \tag{12}$$

We refer to [5] for discussion of the notion of intensity process of an $(\mathbb{F},\mathbb{G})$–DSMC, as well as for a discussion of the relationship between the concept of the $(\mathbb{F},\mathbb{G})$–CMC and the concept of $(\mathbb{F},\mathbb{G})$–DSMC. In particular, sufficient conditions under which an $(\mathbb{F},\mathbb{G})$–DSMC is an $(\mathbb{F},\mathbb{G})$–CMC are given in [5]. Moreover, it is shown in [5] that one can construct an $(\mathbb{F},\mathbb{G})$–CMC, which is also an $(\mathbb{F},\mathbb{G})$–DSMC. It is done for Λ satisfying canonical conditions.

Condition 2.1 *We say that a matrix valued process* $\Lambda = [\lambda^{xy}]_{x,y\in S}$ *satisfies canonical conditions relative to the pair* $(S,\mathbb{F})$ *if:*

(C1) Λ *is an* $\mathbb{F}$*-progressively measurable and it satisfies* (4).

(C2) *The processes* λ^{xy}, $x,y\in S$, $x\neq y$, *have countably many jumps* $\mathbb{P}$*-a.s., and their trajectories admit left limits.*

Any $\mathbb{F}$-adapted càdlàg process $\Lambda_t = [\lambda^{xy}_t]_{x,y\in S}$, for which (4) holds, satisfies canonical conditions.

In what follows, we will use the acronym $(\mathbb{F},\mathbb{G})$–CDMC for any process that is both an $(\mathbb{F},\mathbb{G})$–CMC and an $(\mathbb{F},\mathbb{G})$–DSMC.

Let X be an $(\mathbb{F},\mathbb{F}^X)$–CDMC. Let us note that in view of [5, Theorem 4.15], the intensity of X considered as an $(\mathbb{F},\mathbb{F}^X)$–DSMC coincides, in the sense of [5, Definition 2.5], with the $\mathbb{F}$-intensity Λ of X considered as an $(\mathbb{F},\mathbb{F}^X)$–CMC. Consequently, we will say that X is an $(\mathbb{F},\mathbb{F}^X)$–CDMC with intensity Λ.

In this paper we consider processes X satisfying the following assumptions

Assumption 2.1 *(i)* X *is an* $(\mathbb{F},\mathbb{F}^X)$*–CDMC admitting an intensity.*
(ii) $\mathbb{P}(X_0 = x_0|\mathscr{F}_T) = \mathbb{P}(X_0 = x_0|\mathscr{F}_0)$ *for every* $x_0\in S$.

2.1 Strong Markovian Consistency of Conditional Markov Chains

We consider multivariate processes, so that the state space $S := \times_{k=1}^N S_k$, where S_k is a finite set, $k = 1,\ldots,N$ and X being a multivariate $(\mathbb{F},\mathbb{F}^X)$-CDMC can be written as $X = (X^1,\ldots,X^N)$. Now we introduce the notion of strong Markovian consistency.

Definition 2.4 Let us fix $k\in\{1,\ldots,N\}$. We say that process X satisfies the strong Markovian consistency property with respect to $(X^k,\mathbb{F})$ if for every $x^k_1,\ldots,x^k_m\in S_k$ and for all $0\le t\le t_1\le\ldots\le t_m\le T$, it holds that

$$\mathbb{P}\left(X^k_{t_m}=x^k_m,\ldots,X^k_{t_1}=x^k_1|\mathscr{F}_t\vee\mathscr{F}^X_t\right)=\mathbb{P}\left(X^k_{t_m}=x^k_m,\ldots,X^k_{t_1}=x^k_1|\mathscr{F}_t\vee\sigma(X^k_t)\right), \tag{13}$$

or, equivalently, if X^k is an $(\mathbb{F},\mathbb{F}^X)$-CMC.[1]

[1] In more generality, one might define strong Markovian consistency with respect to a collection $X^I := \{X^k, k\in I\subset\{1,2,\ldots\}\}$ of components of X. This will not be done in this paper though.

The next definition extends the previous one by requiring that the laws of the marginal processes $X^k,\ k=1,\dots,N$, are predetermined. This definition will be a gateway to the concept of strong CMC copula that we introduce in Sect. 3.1.

Definition 2.5 Let $\mathscr{Y}=\left\{Y^1,\dots,Y^N\right\}$ be a family of processes such that each Y^k is an $(\mathbb{F},\mathbb{F}^{Y^k})$-CMC with values in S_k.

(i) Let us fix $k\in\{1,2,\dots,N\}$ and let process X satisfy the strong Markovian consistency property with respect to $(X^k,\mathbb{F})$. If the conditional law of X^k given $\mathscr{F}_T$ coincides with the conditional law of Y^k given $\mathscr{F}_T$, then we say that process X satisfies the strong Markovian consistency property with respect to $(X^k,\mathbb{F},Y^k)$.

(ii) If X satisfies the strong Markovian consistency property with respect to $(X^k,\mathbb{F},Y^k)$ for every $k\in\{1,2,\dots,N\}$, then we say that X satisfies the strong Markovian consistency property with respect to $(\mathbb{F},\mathscr{Y})$.

Now we provide sufficient and necessary conditions for strong Markovian consistency property of X with respect to $(\mathbb{F},\mathscr{Y})$.

Theorem 2.1 ([6, Theorem 3.6]) *Let $\mathscr{Y}=\left\{Y^1,\dots,Y^N\right\}$ be a family of processes such that each Y^k is an $(\mathbb{F},\mathbb{F}^{Y^k})$-CDMC, with values in S_k, and with $\mathbb{F}$-intensity $\Psi_t^k=[\psi_t^{k;x^ky^k}]_{x^k,y^k\in S_k}$. Let process X satisfy Assumption (A). Then, X satisfies the strong Markovian consistency property with respect to $(\mathbb{F},\mathscr{Y})$ if and only if for all $k=1,2,\dots,N$, the following hold:*

(i) For every $x^k,y^k\in S_k, x^k\neq y^k$

$$\begin{aligned}\mathbb{1}_{\{X_t^k=x^k\}}\sum_{\substack{y^n\in S_n,\\ n=1,2,\dots,N,n\neq k}}&\lambda_t^{(X_t^1,\dots,X_t^{k-1},x^k,X_t^{k+1},\dots,X_t^N)(y^1,\dots,y^k,\dots,y^N)}\\ &=\mathbb{1}_{\{X_t^k=x^k\}}\psi^{k;x^ky^k},\quad dt\otimes d\mathbb{P}\text{-a.e.}\end{aligned}\tag{14}$$

(ii) The law of X_0^k given $\mathscr{F}_T$ coincides with the law of Y_0^k given $\mathscr{F}_T$.

The necessary and sufficient condition for strong Markov consistency of X with respect to $(\mathbb{F},\mathscr{Y})$ formulated in Theorem 2.1 may not be easily verified. Here, we provide an algebraic sufficient condition for that, which typically is easily verified. We illustrate this in Sect. 3.2, where Theorem 2.2 will play the key role.

Theorem 2.2 ([6, Proposition 3.9]) *Let $\mathscr{Y}=\left\{Y^1,\dots,Y^N\right\}$ be a family of processes such that each Y^k is an $(\mathbb{F},\mathbb{F}^{Y^k})$-CDMC with values in S_k, and with $\mathbb{F}$-intensity $\Psi_t^k=[\psi_t^{k;x^ky^k}]_{x^k,y^k\in S_k}$. Let process X satisfy Assumption (A). Assume that* *(i) There exists a version of* $\mathbb{F}$*–intensity* Λ *which satisfies the following condition: for each* $k = 1, 2, \ldots, N$*,* $x^k, y^k \in S_k,\ x^k \neq y^k$*,*

$$\psi_t^{k;x^k y^k} = \sum_{\substack{y^n \in S_n, \\ n=1,2,\ldots,N, n\neq k}} \lambda_t^{(x^1,\ldots,x^k,\ldots,x^N)(y^1,\ldots,y^k,\ldots,y^N)}. \tag{15}$$

(ii) The law of X_0^k *given* $\mathscr{F}_T$ *coincides with the law of* Y_0^k *given* $\mathscr{F}_T$ *for all* $k = 1, 2, \ldots, N$*.*

Then, X satisfies the strong Markovian consistency property with respect to $(\mathbb{F}, \mathscr{Y})$*.*

In general, condition (15) is not necessary for the strong Markovian consistency property. However, it needs to be stressed, that this condition is so powerful that it implies strong Markovian consistency property regardless of the initial distribution of process X. On the other hand, whether or not condition (14) holds depends also on the initial distribution of X.

In the next section we will give a construction of a conditionally independent strong CMC copula, which, as stated in the Introduction, finds applications in physics and chemistry, as well as in other disciplines.

3 Conditionally Independent Strong CMC Copula

We first introduce the concept of strong CMC copula, and then we proceed with construction of a conditionally independent strong CMC copula.

3.1 *Strong CMC Copulae*

We begin with

Definition 3.1 Let $\mathscr{Y} = \{Y^1, \ldots, Y^N\}$ be a family of processes, defined on some underlying probability space $(\Omega, \mathscr{A}, \mathbb{Q})$, such that each Y^k is an $(\mathbb{F}, \mathbb{F}^{Y^k})$-CMC with values in S_k. A *strong CMC copula* between processes $Y^1, \ldots, Y^N$ is any multivariate process $X = (X^1, \ldots, X^N)$, given on $(\Omega, \mathscr{A})$ endowed with some probability measure $\mathbb{P}$, such that X is an $(\mathbb{F}, \mathbb{F}^X)$–CMC, and such that it satisfies the strong Markovian consistency property with respect to $(\mathbb{F}, \mathscr{Y})$.

The methodology developed in [5] allows us to construct strong CMC copulae between processes $Y^1, \ldots, Y^N$, that are defined on some underlying probability space $(\Omega, \mathscr{A}, \mathbb{Q})$ endowed with a reference filtration $\mathbb{F}$, and are such that each Y^k is $(\mathbb{F}, \mathbb{F}^{Y^k})$-CDMC with $\mathbb{F}$–intensity, say, $\Psi^k = [\psi^{k;x^k y^k}]_{x^k, y^k \in S_k}$. The additional feature of our construction is that, typically, the constructed CMC copulae X are also $(\mathbb{F}, \mathbb{F}^X)$-DSMC.

According to [6] a natural starting point for constructing a strong copula between $Y^1, \dots, Y^N$ is to determine a system of stochastic processes $[\lambda^{xy}]_{x,y \in S}$ and an S-valued random variable $\xi = (\xi^1, \dots, \xi^N)$ on $(\Omega, \mathscr{A})$, such that they satisfy the following conditions:

(CMC-1)

$$\psi_t^{k;x^k y^k} = \sum_{\substack{y^n \in S_n, \\ n=1,2,\dots,N, n \neq k}} \lambda_t^{(x^1,\dots,x^k,\dots,x^N)(y^1,\dots,y^k,\dots,y^N)}, \quad \begin{array}{l} x^n \in S_n, n = 1, \dots, N, \\ y^k \in S_k, y^k \neq x^k, \\ k = 1, \dots, N, t \in [0,T]. \end{array}$$

(CMC-2) The matrix process $\Lambda_t = [\lambda_t^{xy}]_{x,y \in S}$ satisfies canonical conditions relative to the pair $(S, \mathbb{F})$ (cf. Condition 2.1).

(CMC-3)

$$\mathbb{Q}(\xi = y | \mathscr{F}_T) = \mathbb{Q}(\xi = y | \mathscr{F}_0), \quad \forall y \in S.$$

(CMC-4)

$$\mathbb{Q}(\xi^k = y^k | \mathscr{F}_T) = \mathbb{Q}(Y_0^k = y^k | \mathscr{F}_T), \quad \forall y^k \in S_k, k = 1, \dots, N.$$

We will call any pair (Λ, ξ) satisfying conditions (CMC-1)–(CMC-4) *strong CMC pre-copula* between processes $Y^1, \dots, Y^N$. Given a strong CMC pre-copula between processes $Y^1, \dots, Y^N$, we can construct on $(\Omega, \mathscr{A})$ probability measure $\mathbb{P}$ and process X, starting from measure $\mathbb{Q}$ as above, such that, in view of Theorem 2.2, it satisfies the strong Markovian consistency property with respect to $(\mathbb{F}, \mathscr{Y})$. Thus, it is a strong CMC copula between processes $Y^1, \dots, Y^N$.

Moreover for $\mathbb{P}$ constructed in [5] we have

$$\mathbb{P}(\xi = y | \mathscr{F}_T) = \mathbb{P}(\xi = y | \mathscr{F}_0), \quad \forall y \in S.$$

$$\mathbb{P}(\xi^k = y^k | \mathscr{F}_T) = \mathbb{Q}(Y_0^k = y^k | \mathscr{F}_T), \quad \forall y^k \in S_k, k = 1, \dots, N.$$

Remark 3.1 (i) Note that in the definition of strong CMC copula it is required that $\mathscr{F}_T$-conditional distribution of X_0^k coincides with $\mathscr{F}_T$-conditional distribution of Y_0^k, for $k \in 1, \dots, N$, but, the $\mathscr{F}_T$-conditional distribution of the multivariate random variable $X_0 = (X_0^1, \dots, X_0^N)$ can be arbitrary. Thus, in principle, a strong CMC copula X between processes $Y^1, \dots, Y^N$ can be constructed with help of a strong CMC pre-copula between processes $Y^1, \dots, Y^N$, as well as a copula between the $\mathscr{F}_T$-conditional distributions of X_0^ks, for $k \in 1, \dots, N$. The constructed CMC copulae X are also $(\mathbb{F}, \mathbb{F}^X)$-CDMC.

(ii) In general, there exist numerous systems of stochastic processes that satisfy conditions (CMC-1) and (CMC-2), so that there exist numerous strong pre-copulae between conditional Markov chains Y^1, …, Y^N, and, consequently, there exists numerous strong CMC copulae between conditional Markov chains Y^1, …, Y^N. This is an important feature in applications (see e.g. [4] and [6]).

3.2 Construction of Conditionally Independent Strong CMC Copula

Let $Y^1, \dots, Y^N$ be processes such that each Y^k is an $(\mathbb{F}, \mathbb{F}^{Y^k})$-CDMC with values in S_k, and with $\mathbb{F}$–intensity $\Psi^k_t = [\psi^{k;x^k y^k}_t]_{x^k, y^k \in S_k}$. Assume that for each k the process Ψ^k satisfies canonical conditions relative to the pair $(S_k, \mathbb{F})$. Additionally assume that

$$\mathbb{Q}(Y^k_0 = x^k | \mathscr{F}_T) = \mathbb{Q}(Y^k_0 = x^k | \mathscr{F}_0), \quad \forall x^k \in S_k,\ k = 1, \dots, N. \tag{16}$$

Consider a matrix valued random process Λ given as the following Kronecker sum

$$\Lambda_t = \bigoplus_{k=1}^{N} \Psi^k_t := \sum_{k=1}^{N} I_1 \otimes \dots \otimes I_{k-1} \otimes \Psi^k_t \otimes I_{k+1} \otimes \dots \otimes I_N, \quad t \in [0, T], \tag{17}$$

where I_k denotes the identity matrix of dimensions $|S_k| \times |S_k|$ and $\otimes$ is the Kronecker product of two matrices.[2] Moreover, let us take an S-valued random variable $\xi = (\xi^1, \dots, \xi^N)$, which has $\mathscr{F}_T$-conditionally independent coordinates, that is

$$\mathbb{Q}(\xi^1 = x^1, \dots, \xi^N = x^N | \mathscr{F}_T) = \prod_{i=1}^{N} \mathbb{Q}(\xi^i = x^i | \mathscr{F}_T), \quad \forall x = (x^1, \dots, x^N) \in S. \tag{18}$$

Additionally assume that $\mathscr{F}_T$-conditional distributions of coordinates of ξ and Y_0 coincide, meaning that

$$\mathbb{Q}(\xi^k = x^k | \mathscr{F}_T) = \mathbb{Q}(Y^k_0 = x^k | \mathscr{F}_T), \quad \forall x^k \in S_k,\ k = 1, \dots, N. \tag{19}$$

Now our goal is to prove that

1. (Λ, ξ) is a strong CMC pre-copula between CDMC $Y^1, \dots, Y^N$.
2. The multivariate process X, that is a strong CMC copula constructed from (Λ, ξ) in a way described above, has components which are conditionally independent given $\mathscr{F}_T$.

The process X in 2 above is called *conditionally independent strong CMC copula* or *independence strong copula for CMCs*.

In what follows, we denote by I the identity matrix of dimension $|S|$.

[2] Let us recall that for two given matrices, say $A = [a_{x_k x_l}]_{x_k, x_l \in E_1}$ and $B = [b_{y_m y_n}]_{y_m, y_n \in E_2}$ indexed by elements of some finite sets E_1,E_2, its Kronecker product is the matrix $A \otimes B = [(a \otimes b)_{(x_k, y_m)(x_l, y_n) \in E_1 \times E_2}]$ with entries defined by $(a \otimes b)_{(x_k, y_m)(x_l, y_n)} = a_{x_k x_l} b_{y_m y_n}$. See, e.g., Horn and Johnson [8].

Theorem 3.1 *Suppose that we are given an N-tuple of process $Y^1, \dots, Y^N$ such that each Y^k is an $(\mathbb{F}, \mathbb{F}^{Y^k})$-CDMC with values in S_k, and with $\mathbb{F}$–intensity $\Psi_t^k = [\psi_t^{k;x^ky^k}]_{x^k,y^k\in S_k}$ which satisfy canonical conditions relative to the pair $(S_k, \mathbb{F})$. Moreover, suppose that ξ satisfy (18) and (19) and let Λ be given by (17). Then (Λ, ξ) is a strong CMC pre-copula between $Y^1, \dots, Y^N$.*

Proof In what follows, we will use a convention that for $A \subset \widetilde{S}$, where $\widetilde{S}$ is a finite set, the characteristic function

$$\mathbb{1}_A(j) = \begin{cases} 1 & \text{if } j \in A, \\ 0 & \text{if } j \notin A, \end{cases}$$

is interpreted as a vector in $\mathbf{R}^{|\widetilde{S}|}$, written as $\mathbb{1}_A^{\widetilde{S}}$; for simplicity, we will also denote $\mathbb{1}_{\widetilde{S}} = \mathbb{1}_{\widetilde{S}}^{\widetilde{S}}$. By 0_{S_p} we denote zero vector indexed by elements of S_p.

First we prove that Λ satisfies (CMC-1). Let us fix $k \in \{1, \dots, N\}$, $x^k, y^k \in S_k$. Fix $\bar{x} = (\bar{x}^1, \dots, \bar{x}^N) \in S$ such that $\bar{x}^k = x^k$. Now we observe that

$$\sum_{\substack{y^n \in S_n, \\ n=1,2,\dots,N, n\neq k}} \lambda_t^{(\bar{x}^1,\dots,\bar{x}^k,\dots,\bar{x}^N)(y^1,\dots,y^k,\dots,y^N)} = \left(\mathbb{1}_{\{\bar{x}\}}^S\right)^\top \Lambda_t \mathbf{v}^{y^k},$$

where

$$\mathbf{v}^{y^k} := \mathbb{1}_{S_1} \otimes \dots \otimes \mathbb{1}_{S_{k-1}} \otimes \mathbb{1}_{\{y^k\}}^{S_k} \otimes \dots \otimes \mathbb{1}_{S_N}. \tag{20}$$

Next, we see that

$$\Lambda_t \mathbf{v}^{y^k} = \sum_{m=1}^N \Phi_t^m,$$

where Φ^m are defined by

$$\Phi_t^m = \left((\otimes_{p=1}^{m-1} I_p) \otimes \Psi_t^m \otimes (\otimes_{q=m+1}^N I_n)\right) \mathbf{v}^{y^k}.$$

We have for $m = k$, by using (20) and the mixed-product rule (cf. [8, Lemma 4.2.10]),

$$\begin{aligned}
\Phi_t^k &= \left((\otimes_{p=1}^{k-1} I_p) \otimes \Psi_t^k \otimes (\otimes_{q=k+1}^N I_n)\right) \left((\otimes_{p=1}^{k-1} \mathbb{1}_{S_p}) \otimes \mathbb{1}_{\{y^k\}}^{S_k} \otimes (\otimes_{q=k+1}^N \mathbb{1}_{S_q})\right) \\
&= \left((\otimes_{p=1}^{k-1} I_p \mathbb{1}_{S_p}) \otimes \Psi_t^k \mathbb{1}_{\{y^k\}}^{S_k} \otimes (\otimes_{q=k+1}^N I_q \mathbb{1}_{S_q})\right) \\
&= \left((\otimes_{p=1}^{k-1} \mathbb{1}_{S_p}) \otimes \Psi_t^k \mathbb{1}_{\{y^k\}}^{S_k} \otimes (\otimes_{q=k+1}^N \mathbb{1}_{S_q})\right).
\end{aligned}$$

Analogously, we have, for $m > k$,

$$\begin{aligned}\Phi_t^m &= \big((\otimes_{p=1}^{m-1} I_p) \otimes \Psi_t^m \otimes (\otimes_{q=m+1}^{N} I_n)\big)\Big((\otimes_{p=1}^{k-1} \mathbb{1}_{S_p}) \otimes \mathbb{1}_{\{y^k\}}^{S_k} \otimes (\otimes_{q=k+1}^{N} \mathbb{1}_{S_q})\Big)\\ &= \Big((\otimes_{p=1}^{k-1} I_p \mathbb{1}_{S_p}) \otimes (I_p \mathbb{1}_{\{y^k\}}^{S_k}) \otimes (\otimes_{r=k}^{m-1} I_p \mathbb{1}_{S_p}) \otimes (\Psi_t^m \mathbb{1}_{S_m}) \otimes (\otimes_{q=m+1}^{N} I_q \mathbb{1}_{S_q})\Big)\\ &= \Big((\otimes_{p=1}^{k-1} \mathbb{1}_{S_p}) \otimes (\mathbb{1}_{\{y^k\}}^{S_k}) \otimes (\otimes_{r=k}^{m-1} \mathbb{1}_{S_p}) \otimes (0_{S_m}) \otimes (\otimes_{q=m+1}^{N} \mathbb{1}_{S_q})\Big)\\ &= \otimes_{p=1}^{N} 0_{S_p} = 0_{\times_{p=1}^{N} S_p}\end{aligned}$$

and, for $m < k$,

$$\begin{aligned}\Phi_t^m &= \big((\otimes_{p=1}^{m-1} I_p) \otimes \Psi_t^m \otimes (\otimes_{q=m+1}^{N} I_n)\big)\Big((\otimes_{p=1}^{k-1} \mathbb{1}_{S_p}) \otimes \mathbb{1}_{\{y^k\}}^{S_k} \otimes (\otimes_{q=k+1}^{N} \mathbb{1}_{S_q})\Big)\\ &= \Big((\otimes_{p=1}^{m-1} I_p \mathbb{1}_{S_p}) \otimes (\Psi_t^m \mathbb{1}_{S_m}) \otimes (\otimes_{r=m+1}^{k-1} I_r \mathbb{1}_{S_r}) \otimes (I_k \mathbb{1}_{\{y^k\}}^{S_k})(\otimes_{q=k+1}^{N} I_q \mathbb{1}_{S_q})\Big)\\ &= \Big((\otimes_{p=1}^{m-1} \mathbb{1}_{S_p}) \otimes (0_{S_m}) \otimes (\otimes_{r=m+1}^{k-1} \mathbb{1}_{S_r}) \otimes (\mathbb{1}_{\{y^k\}}^{S_k})(\otimes_{q=k+1}^{N} \mathbb{1}_{S_q})\Big)\\ &= \otimes_{p=1}^{N} 0_{S_p} = 0_{\times_{p=1}^{N} S_p}.\end{aligned}$$

Consequently, for any $\bar{x} = (\bar{x}^1, \dots, \bar{x}^N) \in S$ such that $\bar{x}^k = x^k$ and $y^k \in S_k$, we have that

$$\begin{aligned}\big(\mathbb{1}_{\{\bar{x}\}}^{S}\big)^\top \Lambda_t \mathbf{v}^{y^k} &= \sum_{m=1}^{N} \big(\mathbb{1}_{\{\bar{x}\}}^{S}\big)^\top \Phi_t^m = \big(\mathbb{1}_{\{\bar{x}\}}^{S}\big)^\top \Phi_t^k\\ &= \Big(\prod_{p=1}^{k-1} \mathbb{1}_{S_p}(\bar{x}^p)\Big)\Psi_t^k \mathbb{1}_{\{y^k\}}^{S_k}(\bar{x}^k)\Big(\prod_{q=k+1}^{N} \mathbb{1}_{S_q}(\bar{x}^q)\Big) = \psi_t^{k;x^k y^k}.\end{aligned}$$

This finishes the prove that Λ satisfies (CMC-1).

The fact that Λ satisfies (CMC-2) follows from the assumption that $\Psi_t^k = [\psi_t^{k;xy}]_{x,y\in S}$, satisfies canonical conditions relative to the pair $(S_k, \mathbb{F})$ for every $k = 1, \dots, N$, and from the following representation of the entries of Λ_t:

$$\lambda_t^{(x^1,\dots,x^N)(y^1,\dots,y^N)} = \sum_{m=1}^{N} \Big(\prod_{\substack{n=1\\ n\neq m}}^{N} \mathbb{1}_{\{y^n = x^n\}}\Big)\psi_t^{m;x^m y^m}.$$

It is clear from Assumption 2.1, that any ξ satisfying (18) and (19) satisfies (CMC-3) and (CMC-4). Therefore (Λ, ξ) is a pre-copula between processes $Y^1, \dots, Y^N$ □

Our next aim is to demonstrate that components of the process X constructed from (Λ, ξ) are conditionally independent given $\mathscr{F}_T$. We start with

Lemma 3.1 *Suppose that we are given an N-tuple of matrix valued processes* $\Psi_t^k = [\psi_t^{k;xy}]_{x,y\in S_k}$, $k = 1, \ldots, N$, *which satisfy canonical conditions relative to the pair* $(S_k, \mathbb{F})$ *and*

$$\sum_{x^k \in S_k} \int_0^T \left\| \psi_s^{k;x^k,x^k} \right\| ds < \infty, \quad \forall k = 1, \ldots, N. \tag{21}$$

Let us fix $s \in [0, T]$, *and let* $P(s, \cdot)$ *be the solution of*

$$dP(s,t) = P(s,t)\Lambda_t dt, \quad P(s,s) = I, \quad t \in [s,T], \tag{22}$$

where Λ *is defined by* (17). *Then,*

$$P(s,t) = \bigotimes_{k=1}^{N} P_k(s,t), \tag{23}$$

where

$$dP_k(s,t) = P_k(s,t)\Psi_t^k dt, \quad P_k(s,s) = I_k, \quad t \in [s,T]$$

for $k = 1, \ldots, N$.

Proof We will verify that P defined by (23) satisfies (22), which, by uniqueness of solutions of (22), will imply the desired result. We will proceed by induction on N. First, we take $N = 2$ and we prove that $P^{(2)}(s, \cdot)$ given as

$$P^{(2)}(s,t) := P_1(s,t) \otimes P_2(s,t),$$

satisfies (22), which takes the form

$$dP^{(2)}(s,t) = P^{(2)}(s,t)(\Psi_t^1 \otimes I_2 + I_1 \otimes \Psi_t^2)dt, \quad P^{(2)}(s,s) = I. \tag{24}$$

By the mixed-product rule (cf. [8, Lemma 4.2.10]) we can write $P^{(2)}(s,t)$ as

$$P^{(2)}(s,t) = (P_1(s,t)I_1) \otimes (I_2 P_2(s,t)) = Q_1(s,t)Q_2(s,t), \tag{25}$$

where

$$Q_1(s,t) = P_1(s,t) \otimes I_2, \quad Q_2(s,t) = I_1 \otimes P_2(s,t).$$

Thus, to show (24) we need to prove that

$$d(Q_1(s,t)Q_2(s,t)) = (Q_1(s,t)Q_2(s,t))(\Psi_t^1 \otimes I_2 + I_1 \otimes \Psi_t^2)dt.$$

We have

$$\begin{aligned} dQ_1(s,t) &= d(P_1(s,t)\otimes I_2) = (dP_1(s,t))\otimes I_2 = (P_1(s,t)\Psi_t^1 dt)\otimes I_2 \\ &= (P_1(s,t)\otimes I_2)(\Psi_t^1\otimes I_2)dt = Q_1(s,t)(\Psi_t^1\otimes I_2)dt, \end{aligned}$$

and, similarly,

$$dQ_2(s,t) = Q_2(s,t)(I_1\otimes \Psi_t^2)dt.$$

The matrices $Q_2(s,t)$ and $(\Psi_t^1\otimes I_2)$ commute, because definition of Q_2 and the mixed-product property imply

$$Q_2(s,t)(\Psi_t^1\otimes I_2) = (I_1\otimes P_2(s,t))(\Psi_t^1\otimes I_2) = (I_1\Psi_t^1)\otimes(P_2(s,t)I_2) = \Psi_t^1\otimes P_2(s,t),$$

and analogously

$$(\Psi_t^1\otimes I_2)Q_2(s,t) = (\Psi_t^1\otimes I_2)(I_1\otimes P_2(s,t)) = (\Psi_t^1 I_1)\otimes(I_2P_2(s,t)) = \Psi_t^1\otimes P_2(s,t).$$

Using the above results and integration by parts we get

$$\begin{aligned} d(Q_1(s,t)Q_2(s,t)) &= (dQ_1(s,t))Q_2(s,t) + Q_1(s,t)dQ_2(s,t) \\ &= Q_1(s,t)(\Psi_t^1\otimes I_2)Q_2(s,t)dt + Q_1(s,t)Q_2(s,t)(I_1\otimes\Psi_t^2)dt \\ &= Q_1(s,t)Q_2(s,t)(\Psi_t^1\otimes I_2)dt + Q_1(s,t)Q_2(s,t)(I_1\otimes\Psi_t^2)dt \\ &= Q_1(s,t)Q_2(s,t)(\Psi_t^1\otimes I_2 + I_1\otimes\Psi_t^2)dt, \end{aligned}$$

where the third equality follows since the matrices $Q_2(s,t)$ and $(\Psi_t^1\otimes I_2)$ commute. This demonstrates that $P^{(2)}(s,\cdot)$ satisfies (24). Consequently, in view of the uniqueness of the solution of (24), the result of the lemma is proved in case $N=2$.

Now, let us assume that the assertion of the lemma holds for some $N\geq 2$. We want to show that

$$P^{(N+1)}(s,t) := \bigotimes_{k=1}^{N+1} P_k(s,t)$$

satisfies

$$dP^{(N+1)}(s,t) = P^{(N+1)}(s,t)\Lambda_t^{(N+1)}dt,$$

where

$$\Lambda_t^{(N+1)} := \sum_{k=1}^{N+1} I_1\otimes\ldots\otimes I_{k-1}\otimes\Psi_t^k\otimes I_{k+1}\otimes\ldots\otimes I_{N+1}.$$

Note that using

$$I^{(N)} := \bigotimes_{k=1}^{N} I_k$$

we have

$$\begin{aligned} P^{(N+1)}(s,t) &= P^{(N)}(s,t) \otimes P_{N+1}(s,t) = (P^{(N)}(s,t)I^{(N)}) \otimes (I_{N+1}P_{N+1}(s,t)) \\ &= (P^{(N)}(s,t) \otimes I_{N+1})(I^{(N)} \otimes P_{N+1}(s,t)), \end{aligned} \tag{26}$$

where the third equality follows from the mixed product rule. Now, we will calculate the differentials of components of (26). We have

$$\begin{aligned} d(P^{(N)}(s,t) \otimes I_{N+1}) &= (P^{(N)}(s,t) \otimes I_{N+1})(\Lambda_t^{(N)} \otimes I_{N+1})dt, \\ P^{(N)}(s,s) \otimes I_{N+1} &= I^{(N+1)}, \end{aligned}$$

and

$$\begin{aligned} d(I^{(N)} \otimes P_{N+1}(s,t)) &= (I^{(N)} \otimes P_{N+1}(s,t))(I^{(N)} \otimes \Psi^{N+1})dt, \\ I^{(N)} \otimes P_{N+1}(s,s) &= I^{(N+1)}. \end{aligned}$$

In a similar way as before we prove that matrices $(\Lambda_t^{(N)} \otimes I_{N+1})$ and $(I^{(N)} \otimes P_{N+1}(s,t))$ commute. Integration by parts in (26) yields

$$\begin{aligned} dP^{(N+1)}(s,t) &= P^{(N+1)}(s,t)(\Lambda_t^{(N)} \otimes I_{N+1} + (I_1 \otimes \ldots \otimes I_N) \otimes \Psi_t^{N+1})dt, \\ P^{(N+1)}(s,s) &= I^{(N+1)}. \end{aligned}$$

Since we have

$$\Lambda_t^{(N+1)} = \Lambda_t^{(N)} \otimes I_{N+1} + (I_1 \otimes \ldots \otimes I_N) \otimes \Psi_t^{N+1},$$

this completes the proof. □

Theorem 3.2 *Suppose that $X = (X^1, \ldots, X^N)$ is an S-valued $(\mathbb{F}, \mathbb{G})$-DSMC with c-transition field of the form*

$$P(s,t) = \bigotimes_{k=1}^{N} P_k(s,t), \tag{27}$$

where $P_k = [p_{k;xy}]_{x,y \in S_k}$ is a stochastic matrix valued random field, for $k = 1, \ldots N$. Moreover assume that for all $x = (x^1, \ldots, x^N) \in S$ it holds

$$\mathbb{P}\left(\bigcap_{k=1}^{N}\{X_0^k = x^k\}\Big|\mathscr{F}_T\right) = \prod_{m=1}^{N}\mathbb{P}\left(X_0^k = x^k\Big|\mathscr{F}_T\right). \tag{28}$$

Then, the components $X^1, \dots, X^N$ of X are conditionally independent given $\mathscr{F}_T$.

Proof It suffices to prove that for any $t_1, \dots, t_n \in [0, T]$, and for any sets $A_k^m \subset S_m$, $m = 1, \dots N$, $k = 1, \dots, n$ it holds

$$\mathbb{P}\left(\bigcap_{m=1}^{N}\bigcap_{k=1}^{n}\{X_{t_k}^m \in A_k^m\}\Big|\mathscr{F}_T\right) = \prod_{m=1}^{N}\mathbb{P}\left(\bigcap_{k=1}^{n}\{X_{t_k}^m \in A_k^m\}\Big|\mathscr{F}_T\right). \tag{29}$$

For simplicity, we will give the proof of (29) for $N = 2$. The proof in the general case proceeds along the same lines and will be omitted. We prove (29) in three steps.

Step 1: Let us first note that (27) and the definition of the Kronecker product imply that for any (x^1, x^2), $(y^1, y^2) \in S_1 \times S_2$ we have

$$p_{(x^1,x^2)(y^1,y^2)}(s,t) = p_{1;x^1y^1}(s,t)p_{2;x^2y^2}(s,t). \tag{30}$$

In addition, as we will show now, if $P_1(s,t)$ and $P_2(s,t)$ satisfy $\mathscr{F}_T$-conditional Chapmann-Kolmogorov equations (cf. [9, Theorem 3.6]), then $(P(s,t))_{0\le s\le t\le T}$ defined by (27) satisfies $\mathscr{F}_T$-conditional Chapmann-Kolmogorov equations as well. Indeed, applying the mixed-product rule to the right hand side of (27) we obtain

$$\begin{aligned}
P(s,t)P(t,u) &= (P_1(s,t) \otimes P_2(s,t))(P_1(t,u) \otimes P_2(t,u)) \\
&= (P_1(s,t)P_1(t,u)) \otimes (P_2(s,t)P_2(t,u)) \\
&= P_1(s,u) \otimes P_2(s,u) = P(s,u).
\end{aligned}$$

Step 2: We will show that X^1 and X^2 are $(\mathbb{F}, \mathbb{G})$-DSMC with c-transition fields P_1 and P_2. We first observe that

$$\begin{aligned}
&\mathbb{P}(X_t^1 = y^1|\mathscr{F}_T \vee \mathscr{G}_s)\mathbb{1}_{\{X_s^1=x^1,X_s^2=x^2\}} \\
&\quad = \mathbb{1}_{\{X_s^1=x^1,X_s^2=x^2\}}\sum_{y^2\in S_2}\mathbb{P}(X_t^1 = y^1, X_t^2 = y^2|\mathscr{F}_T \vee \mathscr{G}_s) \\
&\quad = \mathbb{1}_{\{X_s^1=x^1,X_s^2=x^2\}}\sum_{y^2\in S_2}p_{1;x^1y^1}(s,t)p_{2;x^2y^2}(s,t) \\
&\quad = \mathbb{1}_{\{X_s^1=x^1,X_s^2=x^2\}}p_{1;x^1y^1}(s,t)\left(\sum_{y^2\in S_2}p_{2;x^2y^2}(s,t)\right) \\
&\quad = \mathbb{1}_{\{X_s^1=x^1,X_s^2=x^2\}}p_{1;x^1y^1}(s,t),
\end{aligned}$$

where the second equality follows from (30). Now, summing this equality over $x^2 \in S_2$ yields

$$\mathbb{P}(X^1_t = y^1|\mathscr{F}_T \vee \mathscr{G}_s)\mathbb{1}_{\{X^1_s=x^1\}} = \mathbb{1}_{\{X^1_s=x^1\}} p_{1;x^1y^1}(s,t),$$

which means that X^1 is an $(\mathbb{F},\mathbb{G})$-DSMC with c-transition field P_1. Analogously we can prove that X^2 is an $(\mathbb{F},\mathbb{G})$-DSMC with c-transition field P_2.

Step 3: Now, we will prove that (29) holds.

Let us restate (29) in the following equivalent form: for every $y^1_1,\dots,y^1_n \in S_1$ and $y^2_1,\dots,y^2_n \in S_2$ it holds

$$\begin{aligned}
&\mathbb{P}\left(\bigcap_{k=1}^{n}\{(X^1_{t_k},X^2_{t_k}) = (y^1_k,y^2_k)\}|\mathscr{F}_T\right)\\
&\quad = \mathbb{P}\left(\bigcap_{k=1}^{n}\{X^1_{t_k} = y^1_k\}|\mathscr{F}_T\right)\mathbb{P}\left(\bigcap_{k=1}^{n}\{X^2_{t_k} = y^2_k\}|\mathscr{F}_T\right). \qquad (31)
\end{aligned}$$

Next, using the tower property of conditional expectations, the definition of $(\mathbb{F},\mathbb{G})$-DSMC, [5, Proposition 4.6], and (28) we can rewrite the left hand side of (31) as follows

$$\begin{aligned}
&\mathbb{P}\Big((X^1_{t_1},X^2_{t_1}) = (y^1_1,y^2_1),\dots,(X^1_{t_n},X^2_{t_n}) = (y^1_n,y^2_n)|\mathscr{F}_T\Big)\\
&= \mathbb{E}\Big(\mathbb{P}\Big((X^1_{t_1},X^2_{t_1}) = (y^1_1,y^2_1),\dots(X^1_{t_n},X^2_{t_n}) = (y^1_n,y^2_n)|\mathscr{F}_T \vee \mathscr{G}_0\Big)|\mathscr{F}_T\Big)\\
&= \mathbb{E}\Bigg(\sum_{(y^1_0,y^2_0)\in S_1\times S_2} \mathbb{1}_{\{X^1_0=y^1_0,X^2_0=y^2_0\}}\prod_{k=1}^{n} p_{(y^1_{k-1},y^2_{k-1})(y^1_k,y^2_k)}(t_{k-1},t_k)|\mathscr{F}_T\Bigg)\\
&= \sum_{(y^1_0,y^2_0)\in S_1\times S_2} \mathbb{P}\left(X^1_0 = y^1_0, X^2_0 = y^2_0|\mathscr{F}_T\right)\prod_{k=1}^{n} p_{(y^1_{k-1},y^2_{k-1})(y^1_k,y^2_k)}(t_{k-1},t_k).
\end{aligned}$$

Now, employing (28), (30) and some elementary manipulations we obtain

$$\begin{aligned}
&\mathbb{P}\Big((X^1_{t_1},X^2_{t_1}) = (y^1_1,y^2_1),\dots,(X^1_{t_n},X^2_{t_n}) = (y^1_n,y^2_n)|\mathscr{F}_T\Big)\\
&= \sum_{y^1_0\in S_1}\sum_{y^2_0\in S_2} \mathbb{P}\left(X^1_0 = y^1_0|\mathscr{F}_T\right)\mathbb{P}\left(X^2_0 = y^2_0|\mathscr{F}_T\right)\prod_{k=1}^{n} p_{1;y^1_{k-1}y^1_k}(t_{k-1},t_k)p_{2;y^2_{k-1},y^2_k}(t_{k-1},t_k)\\
&= \Bigg(\sum_{y^1_0\in S_1} \mathbb{P}\left(X^1_0 = y^1_0|\mathscr{F}_T\right)\prod_{k=1}^{n} p_{1;y^1_{k-1}y^1_k}(t_{k-1},t_k)\Bigg)\\
&\quad\Bigg(\sum_{y^2_0\in S_2} \mathbb{P}\left(X^2_0 = y^2_0|\mathscr{F}_T\right)\prod_{k=1}^{n} p_{2;y^2_{k-1},y^2_k}(t_{k-1},t_k)\Bigg).
\end{aligned}$$

Summing the above equality over all $y_1^2 \ldots, y_n^2 \in S_2$ yields

$$\mathbb{P}(X_{t_1}^1 = y_1^1, \ldots, X_{t_n}^1 = y_n^1 | \mathscr{F}_T) = \sum_{y_0^1 \in S_1} \mathbb{P}\left(X_0^1 = y_0^1 | \mathscr{F}_T\right) \prod_{k=1}^{n} p_{1;y_{k-1}^1 y_k^1}(t_{k-1}, t_k).$$

Applying analogous reasoning to X^2 we obtain

$$\mathbb{P}(X_{t_1}^2 = y_1^2, \ldots, X_{t_n}^2 = y_n^2 | \mathscr{F}_T) = \sum_{y_0^2 \in S_2} \mathbb{P}\left(X_0^2 = y_0^2 | \mathscr{F}_T\right) \prod_{k=1}^{n} p_{2;y_{k-1}^2, y_k^2}(t_{k-1}, t_k).$$

These facts conclude the proof of (31). □

Finally, using the above facts, we derive the main theorem:

Theorem 3.3 *Suppose that we are given processes Y^1, ..., Y^N such that each Y^k is an $(\mathbb{F}, \mathbb{F}^{Y^k})$-CDMC with values in S_k, and with $\mathbb{F}$–intensity $\Psi_t^k = [\psi_t^{k;x^k y^k}]_{x^k, y^k \in S_k}$ which satisfy canonical conditions relative to the pair $(S_k, \mathbb{F})$. Let $X = (X^1, \ldots, X^N)$ be a CMC copula constructed from a strong CMC pre-copula (Λ, ξ) between Y^1, ..., Y^N, where ξ satisfy (18) and (19) and Λ is given by (17). Then the components of X are $(\mathbb{F}, \mathbb{F}^X)$-CMCs conditionally independent given $\mathscr{F}_T$.*

Proof In view of our assumptions, using Theorem 3.1, we see that (Λ, ξ) is pre-copula. Therefore, X is a copula between Y^1, ..., Y^N by construction (see Remark 3.1.(i)). Moreover, X is $(\mathbb{F}, \mathbb{F}^X)$-CDMC. Thus, the conditional independence of components of X follows from Lemma 3.1 and Theorem 3.2. □

Remark 3.2 Ball, Milne and Yao [1] considered a model that corresponds to a special case of CMCs. They were assuming that single ion channels $X^1, \ldots X^N$ are independent conditionally on an environmental factor process, say Z^E, which is a Markov chain. Their model corresponds to setting the $\mathbb{F}^{Z^E}$ intensity of X^k as

$$\Psi_t^k = \Psi(Z_t^E)$$

and the $\mathbb{F}^{Z^E}$ intensity of the joint process X as

$$\Lambda(Z_t^E) = \sum_{k=1}^{N} I_1 \otimes \ldots \otimes I_{k-1} \otimes \Psi(Z_t^E) \otimes I_{k+1} \otimes \ldots \otimes I_N, \quad t \in [0, T].$$

In our setting, of course, a random environmental factor process driving the intensities of channels between open and closed states can be far more general.

Remark 3.3 Biagini, Groll and Widenmann [3] studied a model for the rational evaluation of premia for unemployment insurance products. They were assuming that the employment status process X^k of a single individual in the pool of employable individuals is a CMC process with state space $S_k = \{1, 2\}$, where 1 stands for employed and 2 for unemployed. The matrix intensity process of X^k is of the form

$$\Psi_t^k := \begin{pmatrix} -\psi^{k;1,2}(Z_t) & \psi^{k;1,2}(Z_t) \\ \psi^{k;2,1}(Z_t) & -\psi^{k;2,1}(Z_t) \end{pmatrix},$$

where Z is a multidimensional process of covariates influencing the modeled evolution of the employment statuses. These covariates represent macro- and microeconomic risk factors, as well as individual-related risk factors. Assuming the conditional independence, and assuming that

$$\psi^{k;i,j}(Z_t) = \alpha^{i,j}(t)e^{(\beta^{i,j},Z_t)}, \quad i \neq j,$$

the authors were able to estimate the stochastic intensities of individuals using Cox proportional hazards model. In [6] we suggest a possible generalization, using CMC copulae, of the model studied in [3]. This generalization, we believe, may provide a more adequate way to deal with computation of the premia.

4 Conclusion

To a great extent, the progress in the emerging theory and practice of structured dependence between stochastic processes will be measured by our ability to construct all sorts of Markov copulae. The present note provides a considerable contribution in this direction.

In particular, in this note we constructed the conditionally independent Markov copula for a family of conditional Markov chains. As mentioned in the Introduction, this is important since conditionally independent Markov copula serves as starting point for modeling structured dependence between CMCs.

In addition, construction given here may be applied, for example, in modeling of gating behavior of multiple ion channels in a membrane patch or in the problem of evaluation of premia for unemployment insurance products. Also, calculation of decision functions, discussed in Jakubowski and Pytel [10], may be undertaken using conditionally independent Markov copula.

There is much more to be done, though. For example, the great challenge is posed by effective construction of weak Markov copulae and weak CMC copulae, that were studied in [4] and in [6], respectively. This will be objective of our future work.

Acknowledgements We thank the referees and the editors for valuable comments and suggestions, which we used revising the original version of this note.

Research of T.R. Bielecki was partially supported by NSF grant DMS-1211256.

References

1. Ball, F., Milne, R.K., Yeo, G.F.: Continuous-time Markov chains in a random environment, with applications to ion channel modelling. Adv. in Appl. Probab. **26**(4), 919–946 (1994). DOI 10.2307/1427898. URL http://dx.doi.org/10.2307/1427898
2. Ball, F., Yeo, G.F.: Lumpability and marginalisability for continuous-time Markov chains. J. Appl. Probab. **30**(3), 518–528 (1993)
3. Biagini, F., Groll, A., Widenmann, J.: Intensity-based premium evaluation for unemployment insurance products. Insurance Math. Econom. **53**(1), 302–316 (2013). DOI 10.1016/j.insmatheco.2013.06.001. URL http://dx.doi.org/10.1016/j.insmatheco.2013.06.001
4. Bielecki, T.R., Jakubowski, J., Niewęgłowski, M.: Intricacies of dependence between components of multivariate Markov chains: weak Markov consistency and weak Markov copulae. Electron. J. Probab. **18**, no. 45, 21 (2013). DOI 10.1214/EJP.v18-2238. URL http://dx.doi.org/10.1214/EJP.v18-2238
5. Bielecki, T.R., Jakubowski, J., Niewęgłowski, M.: Conditional Markov chains, part I: construction and properties (2015). URL http://arxiv.org/abs/1501.05531
6. Bielecki, T.R., Jakubowski, J., Niewęgłowski, M.: Conditional Markov chains, part II: consistency and copulae (2015). URL http://arxiv.org/abs/1501.05535
7. Dabrowski, A.R., McDonald, D.: Statistical analysis of multiple ion channel data. Ann. Statist. **20**(3), 1180–1202 (1992). DOI 10.1214/aos/1176348765. URL http://dx.doi.org/10.1214/aos/1176348765
8. Horn, R.A., Johnson, C.R.: Topics in matrix analysis. Cambridge University Press, Cambridge (1994). Corrected reprint of the 1991 original
9. Jakubowski, J., Niewęgłowski, M.: A class of $\mathbb{F}$-doubly stochastic Markov chains. Electron. J. Probab. **15**, no. 56, 1743–1771 (2010). DOI 10.1214/EJP.v15-815. URL http://dx.doi.org/10.1214/EJP.v15-815
10. Jakubowski, J., Pytel, A.: The Markov consistency of Archimedean survival processes. J. Appl. Probab. **53**(2), 293–409 (2015)
11. Kijima, S., Kijima, H.: Statistical analysis of channel current from a membrane patch. I. Some stochastic properties of ion channels or molecular systems in equilibrium. J. Theoret. Biol. **128**(4), 423–434 (1987). DOI 10.1016/S0022-5193(87)80188-1. URL http://dx.doi.org/10.1016/S0022-5193(87)80188-1

The Construction and Properties of Assortative Configuration Graphs

T.R. Hurd

Abstract In the new field of financial systemic risk, the network of interbank counterparty relationships can be described as a directed random graph. In *cascade models* of systemic risk, this *skeleton* acts as the medium through which financial contagion is propagated. It has been observed in real networks that such counterparty relationships exhibit *negative assortativity*, meaning that a bank's counterparties are more likely to have unlike characteristics. This paper introduces and studies a general class of random graphs called the assortative configuration model, parameterized by an arbitrary node-type distribution P and edge-type distribution Q. The first main result is a law of large numbers that says the empirical edge-type distributions converge in probability to Q as the number of nodes N goes to infinity. The second main result is a formula for the large N asymptotic probability distribution of general graphical objects called *configurations*. This formula exhibits a key property called *locally tree-like* that in simpler models is known to imply strong results of percolation theory on the size of large connected clusters. Thus this paper provides the essential foundations needed to prove rigorous percolation bounds and cascade mappings in assortative networks.

Keywords Skeleton • Systemic risk • Banking network • Configuration graph • Assortativity • Random graph simulation • Large graph asymptotics • Laplace method • Locally tree-like • Percolation theory

The *skeleton* of a financial network at a moment in time is the directed graph whose directed edges indicate which pairs of banks are deemed to have a significant counterparty relationship at this time. The arrow on each edge points from debtor to creditor. It has been often observed in financial networks (and as it happens, also the world wide web) that they are highly *disassortative*, or as we prefer to say, *negatively assortative* (see for example [14] and [1]). This refers to the property

T.R. Hurd (✉)
Department of Mathematics and Statistics, McMaster University, Hamilton, ON, Canada L8S 4K1
e-mail: hurdt@mcmaster.ca

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_11

that any bank's counterparties (i.e. their graph neighbours) have a marked tendency to be banks of an opposite character. For example, it is observed that small banks tend to lend preferentially to large banks rather than other small banks. On the other hand, social networks are commonly observed to have positive assortativity: the friends of highly popular people are more likely to be highly popular. Structural characteristics such as degree distribution and assortativity are felt by some (see [10, 12]) to be highly relevant to *systemic risk*, meaning the stability properties of financial networks, notably their susceptibility to the propagation of contagion effects, that are the subject of the book [9].

The present paper introduces and studies a general class of assortative directed random graphs that is both rich enough to describe real financial, engineered and social networks, and amenable to analytic treatment. In this class, one can determine the relationships between local network topology and global connectivity properties (a theory that is called *percolation*) and ultimately to understand what essential graph characteristics control the stability of systems such as financial networks that rest on such a skeleton. The main aim here is to put a firm theoretical foundation under the class of configuration graphs on N nodes with arbitrary node type distribution P and edge type distribution Q. The class of configuration graphs with general Q has not been well studied previously, and we will generalize some of the classic large N asymptotic results known to be true for the nonassortative configuration graph construction introduced by [2] and others, and described in Sect. 1.2. To this end, an analytical technique based on the Laplace asymptotic method is developed. These techniques turn out to be powerful enough to prove a property we call *locally tree-like* that is known to be key to understanding the percolation properties of graph models similar to the ACG model. Finally, at the end of the paper, an approximate Monte Carlo simulation algorithm for assortative configuration graphs is proposed.

1 Definitions and Basic Results

This section provides some standard graph theoretic definitions and develops an efficient notation for what will follow. Since this paper deals only with directed graphs rather than undirected graphs, the term *graph* will have that meaning. Undirected graphs fit in easily as a subcategory of the directed case.

Definition 1.1 1. For any $N \geq 1$, the collection of *directed graphs* on N nodes is denoted $\mathscr{G}(N)$. The set of *nodes* $\mathscr{N}$ is numbered by integers, i.e. $\mathscr{N} = \{1, \ldots, N\} := [N]$. Then $g \in \mathscr{G}(N)$, a graph on N nodes, is a pair $(\mathscr{N}, \mathscr{E})$ where the set of edges is a subset $\mathscr{E} \subset \mathscr{N} \times \mathscr{N}$ and each element $\ell \in \mathscr{E}$ is an ordered pair $\ell = (v, w)$ called an *edge* or *link*. Links are labelled by integers $\ell \in \{1, \ldots, E\} := [E]$ where $E = |\mathscr{E}|$. Normally, *self-edges* with $v = w$ are excluded from $\mathscr{E}$, that is, $\mathscr{E} \subset \mathscr{N} \times \mathscr{N} \setminus \text{diag}$.

2. A given graph $g \in \mathscr{G}(N)$ can be represented by its $N \times N$ *adjacency matrix* $M(g)$ with components

$$M_{vw}(g) = \begin{cases} 1 \text{ if } (v, w) \in g \\ 0 \text{ if } (v, w) \in \mathscr{N} \times \mathscr{N} \setminus g \end{cases} .$$

3. The *in-degree* $\deg^-(v)$ and *out-degree* $\deg^+(v)$ of a node v are

$$\deg^-(v) = \sum_w M_{wv}(g), \quad \deg^+(v) = \sum_w M_{vw}(g).$$

4. A node $v \in \mathscr{N}$ has *node type* (j, k) if its in-degree is $\deg^-(v) = j$ and its out-degree is $\deg^+(v) = k$. The node set $\mathscr{N} = \cup_{jk}\mathscr{N}_{jk}$ partitions into sets $\mathscr{N}_{jk}$ with the given node type. One writes $k_v = k, j_v = j$ for any $v \in \mathscr{N}_{jk}$ and allow degrees to be any non-negative integer.
5. An edge $\ell = (v, w) \in \mathscr{E}$ is said to have *edge type* (k, j) with in-degree j and out-degree k if it is an out-edge of a node v with out-degree $k_v = k$ and an in-edge of a node w with in-degree $j_w = j$. The edge set $\mathscr{E} = \cup_{kj}\mathscr{E}_{kj}$ partitions into sets $\mathscr{E}_{kj}$ with the given edge type. One writes $\deg^+(\ell) = k_\ell = k$ and $\deg^-(\ell) = j_\ell = j$ whenever $\ell \in \mathscr{E}_{kj}$.
6. For completeness, an undirected graph can be defined as a directed graph g for which $M(g)$ is symmetric.

The standard visualization of a graph g on N nodes is to plot nodes as dots with labels $v \in \mathscr{N}$, and any edge (v, w) as an arrow pointing "downstream" from node v to node w. In the financial system application, such an arrow signifies that bank v is a debtor of bank w and the in-degree $\deg^-(w)$ is the number of banks in debt to w, in other words the existence of the edge (v, w) means "v owes w". Figure 1 illustrates the labelling of types of nodes and edges.

There are constraints on the collections of node type $(j_v, k_v)_{v\in\mathscr{N}}$ and edge type $(k_\ell, j_\ell)_{\ell\in\mathscr{E}}$ if they derive from a graph. By computing the total number of edges $E = |\mathscr{E}|$, the number of edges with $k_\ell = k$ and the number of edges with $j_\ell = j$, one finds three conditions:

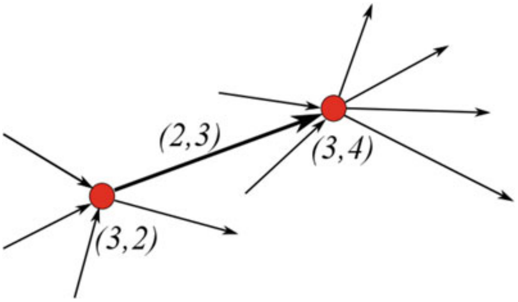


Fig. 1 A type $(3, 2)$ debtor bank that owes to a type $(3, 4)$ creditor bank through a type $(2, 3)$ link

$$\begin{aligned} E &:= |\mathscr{E}| = \sum_v k_v = \sum_v j_v \\ e_k^+ &:= | \cup_j \mathscr{E}_{kj}| = \sum_\ell \mathbb{I}(k_\ell = k) = \sum_v k\mathbb{I}(k_v = k) \\ e_j^- &:= | \cup_k \mathscr{E}_{kj}| = \sum_\ell \mathbb{I}(j_\ell = j) = \sum_v j\mathbb{I}(j_v = j). \end{aligned} \tag{1}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

It is useful to define some further graph theoretic objects and notation in terms of the adjacency matrix $M(g)$:

1. The *in-neighbourhood* of a node v is the set $\mathscr{N}_v^- := \{w \in \mathscr{N} | M_{wv}(g) = 1\}$ and the *out-neighbourhood* of v is the set $\mathscr{N}_v^+ := \{w \in \mathscr{N} | M_{vw}(g) = 1\}$.
2. One writes $\mathscr{E}_v^+$ (or $\mathscr{E}_v^-$) for the set of out-edges (respectively, in-edges) of a given node v and v_ℓ^+ (or v_ℓ^-) for the node for which ℓ is an out-edge (respectively, in-edge).
3. Similarly, second-order neighbourhoods $\mathscr{N}_v^{--}, \mathscr{N}_v^{-+}, \mathscr{N}_v^{+-}, \mathscr{N}_v^{++}$ have the obvious definitions. Second and higher order neighbours can be determined directly from the powers of M and its transpose $M^\top$. For example, $w \in \mathscr{N}_v^{-+}$ whenever $(M^\top M)_{wv} \geq 1$.
4. One often writes j, j', j'', j_1, etc. to refer to in-degrees and k, k', k'', k_1, etc. refer to out-degrees.

Financial network models typically have a sparse adjacency matrix $M(g)$ when N is large, meaning that the number of edges taken to be a small $O(N)$ fraction of the $N(N-1)$ potential edges. This reflects the fact that bank counterparty relationships are expensive to build and maintain, and thus $\mathscr{N}_v^+$ and $\mathscr{N}_v^-$ typically contain relatively few nodes even in a very large network.

1.1 Random Graphs

Random graphs are probability distributions on the sets $\mathscr{G}(N)$:

Definition 1.2 1. A random graph of size N is a probability distribution $\mathbb{P}$ on the finite set $\mathscr{G}(N)$. When the size N is itself random, the probability distribution $\mathbb{P}$ is on the countable infinite set $\mathscr{G} := \cup_N \mathscr{G}(N)$. Normally, it is assumed that $\mathbb{P}$ is invariant under permutations of the N node labels.

2. Given $\mathbb{P}$, the *node-type distribution* is defined to have probabilities $P_{jk} := \mathbb{P}[v \in \mathscr{N}_{jk}]$ for a randomly drawn node v and the *edge-type distribution* is defined to have probabilities $Q_{kj} := \mathbb{P}[\ell \in \mathscr{E}_{kj}]$ for a randomly drawn edge ℓ.

P and Q can be viewed as bivariate distributions on the natural numbers, with marginals $P_k^+ := \sum_j P_{jk}, P_j^- := \sum_k P_{jk}$ and $Q_k^+ := \sum_j Q_{kj}, Q_j^- := \sum_k Q_{kj}$. Edge and node type distributions cannot be chosen independently however, but must be consistent with the fact that they derive from actual graphs which is true if one imposes that equations (2) hold in expectation, that is, P and Q are *consistent* :

$$\begin{aligned} z := \sum_k kP_k^+ &= \sum_j jP_j^- \\ Q_k^+ = kP_k^+/z, \quad Q_j^- &= jP_j^-/z \quad \forall k,j. \end{aligned} \tag{2}$$

Thus z is both the mean in-degree and mean out-degree.

A number of random graph construction algorithms have been proposed in the literature, motivated by the desire to create families of graphs that match the types and measures of network topology that have been observed in nature and society. The present paper focusses on so-called configuration graphs. The textbook "Random Graphs and Complex Networks" by van der Hofstad [15] provides a complete and up-to-date review of the entire subject.

In the analysis to follow, asymptotic results are expressed in terms of convergence of random variables in probability, defined as:

Definition 1.3 A sequence $\{X_n\}_{n\geq 1}$ of random variables is said to *converge in probability* to a random variable X, written $\lim_{n\to\infty} X_n \overset{\mathrm{P}}{=} X$ or $X_n \xrightarrow{\mathrm{P}} X$, if for any $\epsilon > 0$

$$\mathbb{P}[|X_n - X| > \epsilon] \to 0.$$

Recall further standard notation for asymptotics of sequences of real numbers $\{x_n\}_{n\geq 1}, \{y_n\}_{n\geq 1}$ and random variables $\{X_n\}_{n\geq 1}$:

1. Landau's "little oh": $x_n = o(1)$ means $x_n \to 0$; $x_n = o(y_n)$ means $x_n/y_n = o(1)$;
2. Landau's "big oh": $x_n = O(y_n)$ means there is $N > 0$ such that x_n/y_n is bounded for $n \geq N$;
3. $x_n \sim y_n$ means $x_n/y_n \to 1$;
4. $X_n \overset{\mathrm{P}}{=} o(y_n)$ means $X_n/y_n \xrightarrow{\mathrm{P}} 0$.

1.2 Configuration Random Graphs

In their classic paper [7], Erdös and Renyi introduced the undirected model $G(N, M)$ that consists of N nodes and a random subset of exactly M edges chosen uniformly from the collection of $\binom{N}{M}$ possible such edge subsets. This model can be regarded as the Mth step of a random graph process that starts with N nodes and no edges, and adds edges one at a time selected uniformly randomly from the set of available undirected edges. Gilbert's random graph model $G(N,p)$, which takes N nodes and selects each possible edge independently with probability $p = z/(N-1)$, has mean degree z and similar large N asymptotics provided $M = zN/2$. In fact, it was proved by [3] and [13] that the undirected Erdös-Renyi graph $G(N, zN/2)$ and $G(N, p_N)$ with probability $p_N = z/(N-1)$ both converge in probability to the same model as $N \to \infty$ for all $z \in \mathbb{R}_+$. Because of their popularity, the two models $G(N,p) \sim G(N, zN/2)$ have come to be known as "the" random graph. Since the degree distribution of $G(N,p)$ is $\text{Bin}(N-1,p) \sim_{N\to\infty} \text{Pois}(z)$, this is also called the *Poisson graph* model. Both these constructions have obvious directed graph analogues.

The well known directed configuration multigraph model introduced by Bollobás [2] with general degree distribution $P = \{P_{jk}\}_{j,k=0,1,\ldots}$ and size N is constructed by the following random algorithm:

1. Draw a sequence of N node-type pairs $(j_1,k_1),\ldots,(j_N,k_N)$ independently from P, and accept the draw if and only if it is feasible, i.e. $\sum_{n\in[N]}(j_n - k_n) = 0$. Label the nth node with k_n *out-stubs* (a half-edge with an out-arrow) and j_n *in-stubs*.
2. While there remain available unpaired stubs, select (according to any rule, whether random or deterministic) any unpaired out-stub and pair it with an in-stub selected uniformly amongst unpaired in-stubs. Each resulting pair of stubs is a directed edge of the multigraph.

The algorithm leads to objects with self-loops and multiple edges, which are usually called *multigraphs* rather than graphs. Only multigraphs that are free of self-loops and multiple edges, a condition called *simple* , are considered to be graphs. For the most part, one does not care over much about the distinction, because the density of self-loops and multiple edges goes to zero as $N \to \infty$. In fact, Janson [11] has proved in the undirected case that the probability for a multigraph to be simple is bounded away from zero for well-behaved sequences $(g_N)_{N>0}$ of size N graphs with given P.

Exact simulation of the adjacency matrix in the configuration model with general P is problematic because the feasibility condition met in the first step occurs only with asymptotic frequency $\sim \frac{\sigma}{\sqrt{2\pi N}}$, which is vanishingly small for large graphs. For this reason, practical Monte Carlo implementations use some kind of rewiring or *clipping* to adjust each infeasible draw of node-type pairs.

Because of the uniformity of the matching in step 2 of the above construction, the edge-type distribution of the resultant random graph is

$$Q_{kj} = \frac{jkP^+_k P^-_j}{z^2} = Q^+_k Q^-_j \tag{3}$$

which is called the *independent edge condition*. For many reasons, financial and otherwise, one is interested in the more general situation when *assortativity*, defined to be the Pearson correlation of Q:

$$\rho_Q := \frac{\sum_{kj} kj(Q_{kj} - Q_k^+ Q_j^-)}{\sqrt{\mathrm{Var}^+ \mathrm{Var}^-}}$$

$$\mathrm{Var}^+ := \sum_k k^2 Q_k^+ - (\sum_k k Q_k^+)^2, \quad \mathrm{Var}^- := \sum_j j^2 Q_j^- - (\sum_j j Q_j^j)^2$$

is not zero. We will now show how such an extended class of assortative configuration graphs can be defined. The resultant class encompasses all reasonable type distributions (P, Q) and has special properties that make it suitable for exact analytical results, including the possibility of a detailed percolation analysis.

2 The ACG Construction

The assortative configuration (multi-)graph (ACG) of size N parametrized by the node-edge type distribution pair (P, Q) that satisfy the consistency conditions (2) is defined by the *ACG algorithm*:

1. Draw a sequence of N node-type pairs $X = ((j_1, k_1), \ldots, (j_N, k_N))$ independently from P, and accept the draw if and only if it is feasible, i.e. $\sum_{n\in[N]} j_n = \sum_{n\in[N]} k_n$, and this defines the number of edges E that will result. Label the nth node with k_n *out-stubs* (picture each out-stub as a half-edge with an out-arrow, labelled by its degree k_n) and j_n *in-stubs*, labelled by their degree j_n. Define the partial sums $u_j^- := \sum_n \mathbb{I}(j_n = j), u_k^+ := \sum_n \mathbb{I}(k_n = k), u_{jk} := \sum_n \mathbb{I}(j_n = j, k_n = k)$, the number $e_k^+ := k u_k^+$ of *k-stubs* (out-stubs of degree k) and the number of *j-stubs* (in-stubs of degree j), $e_j^- := j u_j^-$.
2. Conditioned on X from Step 1, Step 2 matches k-stubs to j-stubs to form edges of type (k,j), with matching probabilities determined by Q. Given an arbitrary ordering ℓ^- and ℓ^+ of the E in-stubs and E out-stubs, the *matching sequence* or *wiring* W of edges is selected by choosing a pair of permutations $\sigma, \tilde{\sigma} \in S(E)$ of the set $[E]$. This leads to the edge sequence $\ell = (\ell^- = \sigma(\ell), \ell^+ = \tilde{\sigma}(\ell))$ labelled by $\ell \in [E]$, to which is assigned a probability weighting factor
$$\prod_{\ell\in[E]} Q_{k_{\sigma(\ell)} j_{\tilde{\sigma}(\ell)}}. \tag{4}$$

Given the wiring W determined in Step 2, the number of type (k,j) edges is

$$e_{kj} = e_{kj}(W) := \sum_{\ell\in[E]} \mathbb{I}(k_{\tilde{\sigma}(\ell)} = k,\ j_{\sigma(\ell)} = j). \tag{5}$$

The collection $e = (e_{kj})$ of edge-type numbers are constrained by the e_k^+, e_j^- that are determined by Step 1:

$$e_k^+ = \sum_j e_{kj}, \quad e_j^- = \sum_k e_{kj}, \quad E = \sum_{kj} e_{kj}. \tag{6}$$

This construction serves to characterize the precise class of random graphs that shall be called ACG. Its large group of permutation symmetries make it amenable for proving the basic properties of the ACG class, as shall be done in this and the following two sections. However, this defining algorithm is not intended to be an efficient method for simulating random graphs. Efficient approximate simulation methods will be discussed in detail later on in Sect. 5.

Intuitively, since Step 1 leads to a product probability measure subject to a single linear constraint that is true in expectation, one expects that it will lead to the independence of node degrees for large N, with the probability P. Similar logic suggests that since the matching weights in Step 2 define a product probability measure conditional on a set of linear constraints that are true in expectation, it should lead to edge type independence in the large N limit, with the limiting probabilities given by Q. However, the verification of these facts is not so easy, and their justification is the main object of this paper. First, certain combinatorial properties of the wiring algorithm of Step 2, conditioned on the node-type sequence X resulting from Step 1 for a finite N will be derived. One result says that the probability of any wiring sequence $W = (\ell \in [E])$ in step 2 depends only on the set of quantities (e_{kj}) where for each k,j, $e_{kj} := |\{\ell \in [E] \mid \ell \in \mathscr{E}_{kj}\}|$. Another is that the conditional expectation of e_{kj}/E is the exact edge-type probability for all edges in W.

Proposition 1 *Consider Step 2 of the assortative configuration graph construction for finite N with probabilities P, Q conditioned on the $X = (j_i, k_i), i \in [N]$.*

1. *The conditional probability of any wiring sequence $W = (\ell \in [E])$ is:*

$$\mathbb{P}[W \mid X] = C^{-1} \prod_{kj} (Q_{kj})^{e_{kj}(W)}, \tag{7}$$

$$C = C(e^-, e^+) = E! \sum_e \prod_{kj} \frac{(Q_{kj})^{e_{kj}}}{e_{kj}!} \prod_j \left(e_j^-!\right) \prod_k \left(e_k^+!\right), \tag{8}$$

where the sum in (8) is over collections $e = (e_{kj})$ satisfying the constraints (6).

2. *The conditional probability p of any edge of the wiring sequence $W = (\ell \in [E])$ having type k, j is*

$$p = \mathbb{E}[e_{kj} \mid X]/E. \tag{9}$$

Proof of Proposition 1 The denominator of (7) is $C = \sum_{\sigma,\tilde{\sigma} \in S(E)} \prod_{l \in [E]} Q_{k_{\sigma(\ell)} j_{\tilde{\sigma}(\ell)}}$, from which (8) can be verified by induction on E. Assuming (8) is true for $E - 1$, one can verify the inductive step for E:

$$\begin{aligned}
C &= \sum_{\tilde{k},\tilde{j}} \sum_{\sigma,\tilde{\sigma}\in S(E)} \mathbb{I}(k_{\sigma(E)} = \tilde{k}, j_{\tilde{\sigma}(E)} = \tilde{j}) \prod_{l\in[E]} Q_{k_{\sigma(\ell)} j_{\tilde{\sigma}(\ell)}} \\
&= \sum_{\tilde{k},\tilde{j}} e^+_{\tilde{k}} e^-_{\tilde{j}}\, Q_{\tilde{k}\tilde{j}} \sum_{\sigma',\tilde{\sigma}'\in S(E-1)} \prod_{l\in[E-1]} Q_{k_{\sigma'(\ell)} j_{\tilde{\sigma}'(\ell)}} \\
&= \sum_{\tilde{k},\tilde{j}} e^+_{\tilde{k}} e^-_{\tilde{j}}\, Q_{\tilde{k}\tilde{j}}\, (E-1)! \sum_{e'} \prod_{kj} \frac{(Q_{kj})^{e'_{kj}}}{e'_{kj}!} \prod_{j} \left(e'^-_j!\right) \prod_{k} \left(e'^+_k!\right).
\end{aligned}$$

Here, $e'_{kj} = e_{kj} - \mathbb{I}(k = \tilde{k}, j = \tilde{j})$, $e'^-_j = e^-_j - \mathbb{I}(j = \tilde{j})$, $e'^+_k = e^+_k - \mathbb{I}(k = \tilde{k})$. After noting cancellations that occur in the last formula, and re-indexing the collection e' one finds

$$\begin{aligned}
C &= \sum_{\tilde{k},\tilde{j}} \sum_{e'} e_{\tilde{k}\tilde{j}}\, (E-1)! \prod_{kj} \frac{(Q_{kj})^{e_{kj}}}{e_{kj}!} \prod_{j} (e^-_j!) \prod_{k} (e^+_k!) \\
&= \sum_{e} \left(\sum_{\tilde{k},\tilde{j}} e_{\tilde{k}\tilde{j}} \right) (E-1)! \prod_{kj} \frac{(Q_{kj})^{e_{kj}}}{e_{kj}!} \prod_{j} (e^-_j!) \prod_{k} (e^+_k!) \\
&= E! \sum_{e} \prod_{kj} \frac{(Q_{kj})^{e_{kj}}}{e_{kj}!} \prod_{j} (e^-_j!) \prod_{k} (e^+_k!)
\end{aligned}$$

which is the desired result.

Because of the edge-permutation symmetry, it is enough to prove (9) for the last edge. For this, one can follow the same logic and steps as in Part 1 to find:

$$\begin{aligned}
p &= \frac{1}{C(e^-,e^+)} \sum_{\sigma,\tilde{\sigma}\in S(E)} \mathbb{I}(k_{\sigma(E)} = k, j_{\tilde{\sigma}(E)} = j) \prod_{l\in[E]} Q_{k_{\sigma(\ell)} j_{\tilde{\sigma}(\ell)}} \\
&= \frac{E!}{C(e^-,e^+)} \sum_{e} \frac{e_{kj}}{E} \prod_{k'j'} \frac{(Q_{k'j'})^{e_{k'j'}}}{e_{k'j'}!} \prod_{j'} \left(e^-_{j'}!\right) \prod_{k'} (e^+_{k'}!) = \mathbb{E}[e_{kj} \mid X]/E.
\end{aligned}$$

□

An easy consequence of the above proof is that the number of wirings W consistent with a collection $e = (e_{kj})$ is given by

$$|\{W : e(W) = e\}| = \frac{E! \left(\prod_j e^-_j!\right) \left(\prod_k e^+_k!\right)}{\prod_{kj} e_{kj}!}. \tag{10}$$

Because of the permutation symmetries of the construction, a host of more complex combinatorial identities hold for this model. The most important is that

Part 2 of the Proposition can be extended inductively to determine the joint edge distribution for the first M edges conditioned on X. To see how this goes, define two sequences $e_j^-(m), e_k^+(m)$ for $0 \le m \le M$ to be the number of available j-stubs and k-stubs available after m wiring steps.

Proposition 2 *Consider Step 2 of the assortative configuration graph construction for finite N with probabilities P, Q conditioned on X from Step 1. The conditional probability p of the first M edges of the wiring sequence $W = (\ell \in [E])$ having types $(k_i, j_i)_{i\in[M]}$ is*

$$\mathbb{P}[(k_i,j_i)_{i\in[M]} \mid X] = \frac{(E-M)!}{E!} \prod_{i\in[M]} \mathbb{E}[e_{k_i j_i} \mid e^-(i-1), e^+(i-1)]. \tag{11}$$

Proof of Proposition 2 Note that Part 2 of Proposition 1 gives the correct result when $M = 1$. For any m, an extension of the argument that proves Part 2 of Proposition 1 also shows that

$$\begin{aligned}
&\mathbb{P}[(k_i,j_i)_{i\in[m]} \mid X] \\
&= \frac{1}{C(e^-(0), e^+(0))} \sum_{\sigma,\tilde\sigma\in S(E)} \prod_{\ell=1}^{m} \mathbb{I}(k_{\sigma(\ell)} = k_\ell, j_{\tilde\sigma(\ell)} = j_\ell) \prod_{l\in[E]} Q_{k_{\sigma(\ell)} j_{\tilde\sigma(\ell)}} \\
&= \frac{1}{C(e^-(0), e^+(0))} \prod_{\ell=1}^{m} \left[e^-_{j_\ell}(\ell-1) e^+_{k_\ell}(\ell-1)\, Q_{k_\ell j_\ell}\right] \sum_{\sigma',\tilde\sigma'\in S(E-m)} \prod_{\ell=m+1}^{E} Q_{k_{\sigma'(\ell)} j_{\tilde\sigma'(\ell)}}.
\end{aligned} \tag{12}$$

Now assume inductively that the result (11) is true for $M-1$ and compute (11) for M:

$$\begin{aligned}
&\mathbb{P}[(k_i,j_i)_{i\in[M]} \mid X] = \\
&\qquad \frac{\mathbb{P}[(k_i,j_i)_{i\in[M]} \mid X]}{\mathbb{P}[(k_i,j_i)_{i\in[M-1]} \mid X]} \times \frac{(E-M+1)!}{E!} \prod_{i\in[M-1]} \mathbb{E}[e_{k_i j_i} \mid e^-(i-1), e^+(i-1)].
\end{aligned}$$

The ratio in the first factor can be treated using (13), and the resulting cancellations lead to the formula

$$\begin{aligned}
\mathbb{P}[(k_i,j_i)_{i\in[M]} \mid X] &= \frac{(E-M+1)!}{E!} \prod_{i\in[M-1]} \mathbb{E}[e_{k_i j_i} \mid e^-(i-1), e^+(i-1)] \\
&\times \frac{\left[e^-_{j_M}(M-1) e^+_{k_M}(M-1)\, Q_{k_M j_M}\right] \sum_{\sigma',\tilde\sigma'\in S(E-M)} \prod_{\ell=M+1}^{E} Q_{k_{\sigma'(\ell)} j_{\tilde\sigma'(\ell)}}}{\sum_{\sigma',\tilde\sigma'\in S(E-M+1)} \prod_{\ell=M}^{E} Q_{k_{\sigma'(\ell)} j_{\tilde\sigma'(\ell)}}}.
\end{aligned}$$

The desired result follows because Part 2 of Proposition 1 can be applied to show

$$\frac{\left[e^-_{j_M}(M-1)e^+_{k_M}(M-1)\;Q_{k_M j_M}\right]\;\sum_{\sigma',\tilde{\sigma}'\in S(E-M)}\prod_{\ell=M+1}^{E}\;Q_{k_{\sigma'(\ell)}j_{\tilde{\sigma}'(\ell)}}}{\sum_{\sigma',\tilde{\sigma}'\in S(E-M+1)}\prod_{\ell=M}^{E}\;Q_{k_{\sigma'(\ell)}j_{\tilde{\sigma}'(\ell)}}}$$

$$=\frac{1}{E-M+1}\mathbb{E}[e_{k_M j_M}\mid e^-(M-1),e^+(M-1)].$$

□

3 Asymptotic Analysis

It is quite easy to prove that the empirical node-type distributions (u_{jk}, u^-_j, u^+_k) resulting from Step 1 of the ACG algorithm satisfy a law of large numbers:

$$N^{-1}u_{jk}\overset{\mathrm{P}}{=}P_{jk},\quad N^{-1}u^-_j\overset{\mathrm{P}}{=}P^-_j,\quad N^{-1}u^+_k\overset{\mathrm{P}}{=}P^+_k,\tag{13}$$

as $N\to\infty$. In this section, we focus on the new and more difficult problem to determine the asymptotic law of the empirical edge-type distribution, conditioned on the node-type sequence X. To keep the discussion as clear as possible, we confine the analysis to the case the distributions P and Q have support on the finite set $(j,k)\in(\{0\}\cup[K])^2$ for a fixed K. This technical restriction should be removed in future work, since it precludes graph models with fat-tailed degree distributions that are of interest in network applications.

One can see from Proposition 2 that the probability distribution of the first M edge types will be given asymptotically by $\prod_{i\in[M]}Q_{k_i j_i}$ provided our intuition is correct that $\mathbb{E}[E^{-1}e_{kj}]\overset{\mathrm{P}}{=}Q_{kj}(1+o(1))$ asymptotically for large N. To validate this intuition, it turns out one can apply the Laplace asymptotic method to the joint cumulant generating function for the empirical edge-type random variables e_{kj}, conditioned on any feasible collection of (e^+_k, e^-_j) with total number $E=\sum_k e^+_k=\sum_j e^-_j$:

$$F(\mathbf{w};e^-,e^+):=\log\mathbb{E}[e^{\sum_{kj}w_{kj}e_{kj}}\mid e^-,e^+],\quad\forall\,\mathbf{w}=(w_{kj})_{j,k\in[K]}\tag{14}$$

$$=\log\frac{\sum_e\prod_{kj}\frac{(Q_{kj}e^{w_{kj}})^{e_{kj}}}{e_{kj}!}\prod_j\left(e^-_j!\right)\prod_k\left(e^+_k!\right)}{\sum_e\prod_{kj}\frac{(Q_{kj})^{e_{kj}}}{e_{kj}!}\prod_j\left(e^-_j!\right)\prod_k\left(e^+_k!\right)},\tag{15}$$

The constraints on $e=(e_{kj})$ on the sums in both the numerator and denominator of (15) can be introduced by auxiliary integrations over $2K$ new variables v^-_j, v^+_k of the form

$$\mathbb{I}(\sum_j e_{kj}=e^+_k)=\frac{1}{2\pi}\int_0^{2\pi}dv^+_k\;e^{iv^+_k(\sum_j e_{kj}-e^+_k)}.$$

This substitution leads to closed formulas for the sums over e_{kj} and the expression:

$$e^{F(\mathbf{w};\,\mathbf{e})} = \frac{\int_I d^{2K}\mathbf{v}\,\exp[H(\mathbf{w},-i\mathbf{v};\ \mathbf{e})]}{\int_I d^{2K}\mathbf{v}\ \exp[H(0,-i\mathbf{v};\ \mathbf{e})]} \tag{16}$$

where

$$H(\mathbf{w},\alpha;\ \mathbf{e}) = \sum_{kj} e^{(\alpha_j^- + \alpha_k^+)} e^{w_{kj}} Q_{kj} - \Big(\sum_j \alpha_j^- e_j^- + \sum_k \alpha_k^+ e_k^+\Big) = \sum_{kj} e^{\alpha\cdot\mathbf{d}_{jk}} e^{w_{kj}} Q_{kj} - \alpha\cdot\mathbf{e}. \tag{17}$$

The integration in (16) is over the set $I := [0, 2\pi]^{2K}$.

Here a "double vector" notation has been introduced for $\mathbf{v} = (v^-; v^+)$, $\mathbf{e} = (e^-; e^+), \alpha = (\alpha^-; \alpha^+)$ where $v^-, v^+ \in \mathbb{C}^K$ etc. and where K is the number of possible in and out degrees (which one may want to take to be infinite). Define double vectors $\mathbf{1}^- = (1, 1, \dots\ 1; 0, \dots,\ 0), \mathbf{1}^+ = (0, \dots,\ 0; 1, \dots,\ 1), \mathbf{1} = \mathbf{1}^- + \mathbf{1}^+, \tilde{\mathbf{1}} = \mathbf{1}^- - \mathbf{1}^+$. For any pair $(j, k) \in [K]^2$, let $\mathbf{d}_j^-$ be the double vector with a 1 in the jth place and zeros elsewhere, let $\mathbf{d}_k^+$ be the double vector with a 1 in the $K+k$th place and zeros elsewhere and $\mathbf{d}_{jk} = \mathbf{d}_j^- + \mathbf{d}_k^+$. Using the natural inner product for double vectors $\alpha\cdot\mathbf{e} := \sum_j \alpha_j^- e_j^- + \sum_k \alpha_k^+ e_k^+$, etc., the number of stubs is $\mathbf{e}\cdot\mathbf{1} = 2E$ and the feasibility condition on stubs can be written $\mathbf{e}\cdot\tilde{\mathbf{1}} = 0$.

The main aim of the paper is to prove a conditional law of large numbers for $E^{-1}e_{jk}$ as $E \to \infty$, conditioned on $\mathbf{e} = (e^-; e^+)$ satisfying $\mathbf{e}\cdot\tilde{\mathbf{1}} = 0$. By explicit differentiation of the cumulant generating function, and some further manipulation, one finds that

$$\mathbb{E}[e_{kj} \mid \mathbf{e}] = \frac{\partial F}{\partial w_{kj}}\Big|_{\mathbf{w}=0} = Q_{kj}\frac{\int_I d^{2K}\mathbf{v}\ \exp[H(0,-i\mathbf{v}; \mathbf{e}-\mathbf{d}_{jk})]}{\int_I d^{2K}\mathbf{v}\ \exp[H(0,-i\mathbf{v};\ \mathbf{e})]} \tag{18}$$

$$\mathrm{Var}[e_{kj} \mid \mathbf{e}] = \frac{\partial^2 F}{\partial w_{kj}^2}\Big|_{\mathbf{w}=0} = Q_{kj}\frac{\int_I d^{2K}\mathbf{v}\ \exp[H(0,-i\mathbf{v};\ \mathbf{e}-\mathbf{d}_{jk})]}{\int_I d^{2K}\mathbf{v}\ \exp[H(0,-i\mathbf{v};\ \mathbf{e})]} + \tag{19}$$

$$(Q_{kj})^2\left[\frac{\int_I d^{2K}\mathbf{v}\ \exp[H(0,-i\mathbf{v};\ \mathbf{e}-2\mathbf{d}_{jk})]}{\int_I d^{2K}\mathbf{v}\ \exp[H(0,-i\mathbf{v};\ \mathbf{e})]} - \left(\frac{\int_I d^{2K}\mathbf{v}\ \exp[H(0,-i\mathbf{v};\ \mathbf{e}-\mathbf{d}_{jk})]}{\int_I d^{2K}\mathbf{v}\ \exp[H(0,-i\mathbf{v};\ \mathbf{e})]}\right)^2\right]$$

Since our present aim is to understand (18) and (19), we henceforth set $\mathbf{w} = 0$ in the H-function. The H function defined by (17) with $\mathbf{w} = 0$ has special combinatorial features:

Lemma 3.1 *For all $\mathbf{e} \in \mathbb{Z}_+^{2K}$ satisfying $\mathbf{e}\cdot\tilde{\mathbf{1}} = 0$, the function $H = H(\alpha; \mathbf{e})$ satisfies the following properties:*

1. *H is convex for $\alpha \in \mathbb{R}^{2K}$ and entire analytic for $\alpha \in \mathbb{C}^{2K}$;*
2. *H is periodic: $H(\alpha + 2\pi i\eta;\ \mathbf{e}) = H(\alpha;\ \mathbf{e})$ for all $\eta \in \mathbb{Z}^{2K}$.*
3. *For any $\lambda \in \mathbb{C}$, $H(\alpha + \lambda\tilde{\mathbf{1}};\ \mathbf{e}) = H(\alpha;\ \mathbf{e})$;*

4. *For any* $\lambda > 0$, $H(\alpha; \lambda\mathbf{e}) = \lambda H(\alpha - \frac{\log\lambda}{2}\mathbf{1};\ \mathbf{e}) - \frac{\lambda}{2}\ \log\lambda\mathbf{1}\cdot\mathbf{e}$.
5. *The mth partial derivative of H with respect to* α *is given by*

$$\nabla^m H(\alpha;\ \mathbf{e}) = \begin{cases} \sum_{jk} \mathbf{d}_{jk} e^{\alpha\cdot\mathbf{d}_{jk}}\ Q_{kj} - \mathbf{e}, & m = 1; \\ \sum_{jk} (\mathbf{d}_{jk})^{\bigotimes m} e^{\alpha\cdot\mathbf{d}_{jk}}\ Q_{kj},\ m = 2, 3, \ldots \end{cases} \tag{20}$$

Here $(\mathbf{d}_{jk})^{\bigotimes m}$ *denotes the mth tensor power of the double vector* $\mathbf{d}_{jk}$.

The Laplace asymptotic method (or saddlepoint method), reviewed for example in [8], involves shifting the $\mathbf{v}$ integration into the complex by an imaginary vector. The Cauchy Theorem, combined with the periodicity of the integrand in $\mathbf{v}$, will ensure the value of the integral is unchanged under the shift. The desired shift is determined by the $\mathbf{e}$-dependent critical points α^* of H which by Part 5 of Lemma 3.1 are solutions of

$$\sum_{jk} \mathbf{d}_{jk} e^{\boldsymbol{\alpha}\cdot\mathbf{d}_{jk}}\ Q_{kj} = \mathbf{e}. \tag{21}$$

In view of Parts 1 and 2 of the Lemma, for each $\mathbf{e} \in \mathbb{Z}^{2K}$ there is a unique critical point $\alpha^*(\mathbf{e})$ such that $\tilde{\mathbf{1}} \cdot \alpha^*(\mathbf{e}) = 0$. The imaginary shift of the $\mathbf{v}$-integration is implemented by writing $\mathbf{v} = i\alpha^*(\mathbf{e}) + \zeta$ where now ζ is integrated over I.

To unravel the E dependence, one uses rescaled variables $\mathbf{x} = E^{-1}\mathbf{e}$ that lie on the plane $\mathbf{1}\cdot\mathbf{x} = 2$ and by Part 4 of the Lemma with $\lambda = E^{-1}$ one has that $\alpha^*(\mathbf{e}) = \alpha^*(\mathbf{x}) + \frac{\log E}{2}\ \mathbf{1}$. Now one can use the third order Taylor expansion with remainder to write

$$\begin{aligned} H(\alpha^*(\mathbf{e}) - i\zeta;\ \mathbf{e}) &= EH(\alpha^*(\mathbf{x}) - i\zeta; \mathbf{x}) - \frac{E}{2}\log E(\mathbf{1}\cdot\mathbf{x}) \\ &= -E\log E + E\left[H(\alpha^*(\mathbf{x}); \mathbf{x}) - \frac{1}{2}\zeta^{\bigotimes 2}\cdot\nabla^2 H + i\frac{1}{6}\zeta^{\bigotimes 3}\cdot\nabla^3 H\right] + EO(|\zeta|^4) \end{aligned} \tag{22}$$

where $\nabla^2 H$, $\nabla^3 H$ are evaluated at $\alpha^*(\mathbf{x})$ and the square-bracketed quantities are all E independent. From (17) one can observe directly that $|e^H|$ has a unique maximum on the domain of integration at $\zeta = 0$:

$$\max_{\zeta\in I} |e^{H(\alpha^*(\mathbf{e})-i\zeta;\ \mathbf{e})}| = e^{H(\alpha^*(\mathbf{e});\ \mathbf{e})}. \tag{23}$$

The uniqueness of the maximum is essential to validate the following Laplace asymptotic analysis, and leads to the main result of the paper:

Theorem 3.1 *For any double vector* $\mathbf{x}^* \in (0, 1)^{2K} \cap \tilde{\mathbf{1}}^\perp$, *let* $\mathbf{e}(E) = E\mathbf{x}(E)$ *be a sequence in* $\mathbb{Z}_+^{2K} \cap \tilde{\mathbf{1}}^\perp$ *such that*

$$\lim_{E\to\infty} \mathbf{x}(E) = \mathbf{x}^*. \tag{24}$$

Then asymptotically as $E \to \infty$,

$$\mathcal{I}(E) = \int_I d^{2K}\mathbf{v}\ \exp[H(-i\mathbf{v};\ \mathbf{e}(E))] \tag{25}$$
$$= (2\pi)^{K+1/2} E^{1/2-K} e^{-E\log E + EH(\alpha^*(\mathbf{x}^*);\mathbf{x}^*)} \left[\det{}_0 \nabla^2 H\right]^{-1/2} \left[1 + O(E^{-1})\right].$$

Here $\det_0 \nabla^2 H$ *represents the determinant of the matrix projection onto* $\tilde{\mathbf{1}}^{\perp}$*, the subspace orthogonal to* $\tilde{\mathbf{1}}$*, of* $\nabla^2 H$ *evaluated at the critical point* $\alpha^*(\mathbf{x}^*)$*.*

When applied to (18) and (19) this Theorem is powerful enough to yield the desired results on the edge-type distribution in the ACG model for fixed $\mathbf{e} = (e^-, e^+) = E\mathbf{x}$ for large E.

Corollary 1 *Consider the ACG model with* (P, Q) *supported on* $(\{0\} \cup [K])^2$*.*

1. *Conditioned on X,*

$$E^{-1} e_{kj} \overset{\mathrm{P}}{=} [Q_{kj} e^{1-H(0,\alpha^*(\mathbf{x});\mathbf{x}) - \alpha^*(\mathbf{x})\cdot \mathbf{d}_{jk}} [1 + O(E^{-1/2})]$$

where $\mathbf{x} = E^{-1}\mathbf{e}$ *and* $\mathbf{e} = (e^-(X), e^+(X))$*.*

2. *Unconditionally,*

$$E^{-1} e_{kj} \overset{\mathrm{P}}{=} Q_{kj}[1 + O(N^{-1/2})].$$

Combining this Law of Large Numbers result with the easier result for the empirical node-type distribution confirms that the large N asymptotics of the empirical node- and edge-type distributions agree with the target (P, Q) distributions.

Proof of Corollary 1 By applying Part 4 of Lemma 3.1 and the Theorem to (18) one finds that

$$\mathbb{E}[e_{kj} \mid \mathbf{e}] = Q_{kj} \frac{\int_I d^{2K}\mathbf{v}\ \exp[H(-i\mathbf{v};\ \mathbf{e} - \mathbf{d}_{jk})]}{\int_I d^{2K}\mathbf{v}\ \exp[H(-i\mathbf{v};\ \mathbf{e})]}$$
$$= Q_{kj}\ \exp[-(E-1)\log(E-1) + E\log E + (E-1)H(\alpha^*(\mathbf{x}'); \mathbf{x}') - EH(\alpha^*(\mathbf{x}); \mathbf{x})]$$
$$\times \left[\frac{\det{}_0 \nabla^2 H(\alpha^*(\mathbf{x}))}{\det{}_0 \nabla^2 H(\alpha^*(\mathbf{x}'))}\right]^{1/2} \left[1 + O(E^{-1})\right]$$

where $\mathbf{x} = E^{-1}\mathbf{e}$ and $\mathbf{x}' = (E-1)^{-1}(\mathbf{e} - \mathbf{d}_{jk})$ are such that $\Delta\mathbf{x} = \mathbf{x}' - \mathbf{x} = O(E^{-1})$. Now, one can show that as long as $\mathbf{x}, \mathbf{x}'$ lie on the plane $\mathbf{1} \cdot \mathbf{x} = 2$ as they do here, and $\Delta\mathbf{x} = \mathbf{x}' - \mathbf{x}$ is $O(E^{-1})$ then

$$H(\alpha^*(\mathbf{x}'); \mathbf{x}') - H(\alpha^*(\mathbf{x}); \mathbf{x}) = \alpha^*(\mathbf{x}) \cdot \Delta\mathbf{x} + O(|\Delta\mathbf{x}|^2). \tag{26}$$

It is also true that $\Delta\alpha^* = \alpha^*(\mathbf{x}') - \alpha^*(\mathbf{x}) = O(|\Delta\mathbf{x}|)$ and satisfies

$$\Delta\alpha^* \cdot \mathbf{x} = O(|\Delta\mathbf{x}|^2). \tag{27}$$

Since $\det_0 \nabla^2 H(\alpha)$ is analytic in α with $O(1)$ derivatives, and $\Delta\alpha^* = O(|\Delta\mathbf{x}|)$

$$\left[\frac{\det_0 \nabla^2 H(\alpha^*(\mathbf{x}))}{\det_0 \nabla^2 H(\alpha^*(\mathbf{x}'))}\right]^{1/2} = \left[1 + O(E^{-1})\right].$$

Also,

$$(E-1)H(\alpha^*(\mathbf{x}');\mathbf{x}') - (E-1)H(0,\alpha^*(\mathbf{x});\mathbf{x}) = -\alpha^*(\mathbf{x}) \cdot \mathbf{d}_{jk} + O(|\Delta\mathbf{x}|)$$

and $E\log E - (E-1)\log(E-1) \sim \log E + 1 + O(E^{-1})$, from which one concludes

$$\mathbb{E}[e_{kj} \mid \mathbf{e}] = Q_{kj}\, E\, \exp[1 - H(\alpha^*(\mathbf{x});\mathbf{x}) - \alpha^*(\mathbf{x}) \cdot \mathbf{d}_{jk}]\left[1 + O(E^{-1})\right]. \tag{28}$$

The conclusion of the Part 1 of the Corollary now follows from the Chebyshev inequality if one shows that (19) is $O(E)$. Since the first term of(19) equals $\mathbb{E}[e_{kj} \mid \mathbf{e}]$, which is $O(E)$, it is only necessary to show that the $O(E^2)$ parts of the second term cancel. Each ratio in the second term can be analyzed exactly as above, leading to

$$\begin{aligned}
&\left[Q_{kj}E\exp[1 - H(\alpha^*(\mathbf{x});\mathbf{x}) - \alpha^*(\mathbf{x}) \cdot \mathbf{d}_{jk}]\right]^2 \\
&\qquad \times \left(\exp[H(\alpha^*(\mathbf{x});\mathbf{x}) - H(\alpha^*(\mathbf{x}');\mathbf{x}') - \Delta\alpha^*(\mathbf{x}') \cdot \mathbf{d}_{jk}]\right)\left[1 + O(E^{-1})\right] \\
&\quad = \left[Q_{kj}E\exp[1 - H(\alpha^*(\mathbf{x});\mathbf{x}) - \alpha^*(\mathbf{x}) \cdot \mathbf{d}_{jk}]\right]^2 \\
&\qquad \times \left(\exp[-\alpha^*(\mathbf{x}) \cdot \Delta\mathbf{x} - \Delta\alpha^*(\mathbf{x}') \cdot \mathbf{d}_{jk}] - 1\right)\left[1 + O(E^{-1})\right] \; = \; O(E)
\end{aligned}$$

where one uses (26) again in the second last equality.

To prove Part 2, it is sufficient to note that $E^{-1}(e^-(X), e^+(X)) = (Q^-, Q^+)[1 + O(N^{-1/2})]$ and that $\alpha^*(Q^-, Q^+) = 0, H(\alpha^*(Q^-, Q^+); Q^-, Q^+) = 1$. □

Proof of Theorem 3.1 For each E, since the integrand of $\mathscr{I}(E)$ is entire analytic and periodic, its integral is unchanged under a purely imaginary shift of the contour. Also, since by Part 3 of Lemma 3.1 the integrand is constant in directions parallel to $\tilde{\mathbf{1}}$, the integrand can be reduced to the set $I \cap \tilde{\mathbf{1}}^\perp$. Thus, using (22) for $\mathbf{e} = \mathbf{e}(E)$ and $\mathbf{x} = \mathbf{x}(E)$, $\mathscr{I}(E)$ can be written

$$\begin{aligned}
\mathscr{I}(E) &= 2\pi \int_{I\cap\tilde{\mathbf{1}}^\perp} d^{2K-1}\zeta\; \exp[H(\alpha^*(\mathbf{e}) - i\zeta;\; \mathbf{e})] = \; 2\pi \int_{I\cap\tilde{\mathbf{1}}^\perp} d^{2K-1}\zeta \\
&\times \exp\left[-E\log E + E\left(H(\alpha^*(\mathbf{x});\mathbf{x}) - \frac{1}{2}\zeta^{\otimes 2} \cdot \nabla^2 H + i\frac{1}{6}\zeta^{\otimes 3} \cdot \nabla^3 H + O(|\zeta|^4)\right)\right].
\end{aligned}$$

In rescaled variables $\tilde{\zeta} = E^{1/2}\zeta$ this becomes

$$\mathscr{I}(E) = 2\pi E^{1/2-K} \exp\left[-E\log E + EH(\alpha^*(\mathbf{x});\mathbf{x})\right] \times \tilde{\mathscr{I}}(E)$$

where

$$\tilde{\mathscr{I}}(E) := \int_{E^{1/2}I\cap\tilde{\mathbf{1}}^{\perp}} d^{2K-1}\tilde{\zeta}\ \exp[EH(\alpha^*(\mathbf{x}) - iE^{-1/2}\tilde{\zeta};\mathbf{x}) - EH(\alpha^*(\mathbf{x});\mathbf{x})]$$
$$= \int_{E^{1/2}I\cap\tilde{\mathbf{1}}^{\perp}} d^{2K-1}\tilde{\zeta}\ \exp[-\frac{1}{2}\tilde{\zeta}^{\otimes 2}\cdot\nabla^2 H]\left(1 + i\frac{E^{-1/2}}{6}\tilde{\zeta}^{\otimes 3}\cdot\nabla^3 H + O(E^{-1})\right).$$

In this last integral the $O(E^{-1/2})$ term is odd in $\tilde{\zeta}$ and makes no contribution. Now,

$$|\exp[H(\alpha^*(\mathbf{x}) - iE^{-1/2}\tilde{\zeta};\mathbf{x}) - H(\alpha^*(\mathbf{x});\mathbf{x})]| =$$
$$= \exp\left[\sum_{kj} e^{\alpha^*(\mathbf{x})\cdot\mathbf{d}_{jk}}\ (\cos(E^{-1/2}\tilde{\zeta}\cdot\mathbf{d}_{jk}) - 1)\ Q_{kj}\right]$$

clearly has a unique maximum at $\tilde{\zeta} = 0$. Therefore, a standard version of the Laplace method such as that found in [6] is sufficient to imply that as $E \to \infty$,

$$\tilde{\mathscr{I}}(E) = (2\pi)^{K-1/2}\left[\det_0\nabla^2 H\right]^{-1/2}\left[1 + O(E^{-1})\right] \qquad (29)$$

where $\nabla^2 H$ is evaluated at $\alpha^*(\mathbf{x}^*)$. □

4 Locally Tree-Like Property

To understand percolation theory on random graphs, or to derive a rigorous treatment of cascade mappings on random financial networks, it turns out to be important that the underlying random graph model have a property sometimes called *locally tree-like*. In this section, the local tree-like property of the ACG model will be characterized as a particular large N property of the probability distributions associated with graphical objects we call *configurations*, that are roughly speaking finite connected subgraphs g of the skeleton labelled by their degree types.

Before the definition of *configuration* is made clear in the next subsection, first consider what it means in the (P, Q) ACG model with size N to draw a random configuration g consisting of a pair of vertices v_1, v_2 joined by a link, that is, $v_2 \in \mathscr{N}^-_{v_1}$. In view of the permutation symmetry of the ACG algorithm, the random link can without loss of generality be taken to be the first link $W(1)$ of the wiring sequence W. Following the ACG algorithm, Step 1 constructs a feasible node degree sequence $X = (j_i, k_i), i \in [N]$ on nodes labelled by $v_i = i$ and conditioned on X, Step 2 constructs a random Q-wiring sequence $W = \big(\ell = (v_\ell^+, v_\ell^-)\big)_{\ell \in [E]}$ with $E = \sum_i k_i = \sum_i j_i$ edges. By an abuse of notation, we label their edge types by $k_\ell = k_{v_\ell^+}, j_\ell = j_{v_\ell^-}$ for $\ell \in [E]$. The configuration event in question, namely that the first link in the wiring sequence W attaches to nodes of the required degrees $(j_1, k_1), (j_2, k_2)$, has probability $p = \mathbb{P}[v_i \in \mathcal{N}_{j_i,k_i}, i = 1, 2 \mid v_2 \in \mathcal{N}_{v_1}^-]$. To compute this, note that the fraction $j_1 u_{j_1k_1}/e^-_{j_1}$ of available j_1-stubs come from a j_1k_1 node and the fraction $k_2 u_{j_2k_2}/e^+_{k_2}$ available k_2-stubs come from a j_2k_2 node. Combining this fact with Part 2 of Proposition 1, Eq. (9) implies the configuration probability conditioned on X is exactly

$$p = j_1 u_{j_1k_1} k_2 u_{j_2k_2} \, \frac{\mathbb{E}[e_{k_2j_1} \mid e^-, e^+]}{E e^+_{k_2} e^-_{j_1}}. \tag{30}$$

By the Corollary:

$$p \overset{\mathrm{P}}{=} \frac{j_1 k_2 P_{j_1k_1} P_{j_2k_2} Q_{k_2j_1}}{z^2 Q^+_{k_2} Q^-_{j_1}}[1 + O(N^{-1/2})]. \tag{31}$$

This argument justifies the following informal computation of the correct asymptotic expression for p by successive conditioning:

$$\begin{aligned} p &= \mathbb{P}[v_i \in \mathcal{N}_{j_ik_i}, i = 1, 2 \mid v_2 \in \mathcal{N}_{v_1}^-] && (32)\\ &= \mathbb{P}[v_1 \in \mathcal{N}_{j_1k_1} \mid v_2 \in \mathcal{N}_{v_1}^- \cap \mathcal{N}_{j_2k_2}]\mathbb{P}[v_2 \in \mathcal{N}_{j_2k_2} \mid v_2 \in \mathcal{N}_{v_1}^-] && (33)\\ &= P_{k_1|j_1} Q_{j_1|k_2} P_{j_2|k_2} Q^+_{k_2} = \frac{P_{j_1k_1} P_{j_2k_2} Q_{k_2j_1}}{P^+_{k_2} P^-_{j_1}} && (34) \end{aligned}$$

where we introduce conditional degree probabilities $P_{k|j} = P_{jk}/P^-_j$ etc.

Occasionally in the above matching algorithm, the first edge forms a self-loop, i.e. $v_1 = v_2$. The probability of this event, jointly with fixing the degree of v_1, can be computed exactly for finite N as follows:

$$\tilde{p} := \mathbb{E}[v_1 = v_2, v_1 \in \mathcal{N}_{jk} \mid v_2 \in \mathcal{N}_{v_1}^- \mid X] = \left(\frac{jk u_{jk}}{e^-_j e^+_k}\right) \frac{\mathbb{E}[e_{kj} \mid X]}{E}.$$

As $N \to \infty$ this goes to zero, however $N\tilde{p}$ approaches a finite value:

$$N\tilde{p} \xrightarrow{\mathrm{P}} \frac{jk P_{jk} Q_{kj}}{z^2 Q^+_k Q^-_j} \tag{35}$$

which says that the relative fraction of edges being self loops is the asymptotically small $\sum_{jk} \frac{jkP_{jk}Q_{kj}}{Nz^2Q^+_k Q^-_j}$. In fact, following results of [11] and others on the undirected configuration model, one expects that the total number of self loops in the multigraph converges in probability to a Poisson random variable with finite parameter

$$\lambda = \sum_{jk} \frac{jkP_{jk}Q_{kj}}{z^2Q^+_k Q^-_j}. \tag{36}$$

4.1 General Configurations

A general *configuration* is a connected subgraph g of an ACG graph $(\mathscr{N}, \mathscr{E})$ with L ordered edges and with each node labelled by its degree type. It results from a growth process that starts from a fixed node w_0 called the root and at step $\ell \le L$ adds one edge ℓ that connects a node w_ℓ to a specific existing node w'_ℓ. The following is a precise definition:

Definition 4.4 A configuration rooted to a node w_0 with degree $(j,k) := (j_0,k_0)$ is a connected subgraph g consisting of a sequence of L edges that connect nodes $(w_\ell)_{\ell\in[L]}$ of types (j_ℓ,k_ℓ), subject to the following condition: For each $\ell \ge 1$, w_ℓ is connected by the edge labelled with ℓ to a node $w'_\ell \in \{w_j\}_{j\in\{0\}\cup[\ell-1]}$ by either an in-edge (that points into w'_ℓ) (w_ℓ, w'_ℓ) or an out-edge (that points out of w'_ℓ).

A random realization of the configuration results when the construction of the size N ACG graph $(\mathscr{N}, \mathscr{E})$ is conditioned on X arising from Step 1 and the first L edges of the wiring sequence of Step 2. The problem is to compute the probability of the node degree sequence $(j_\ell,k_\ell)_{\ell\in[L]}$ conditioned on X, the graph g with its root $w_0 := v$ having degree (j,k), that is

$$p = \mathbb{P}[w_\ell \in \mathscr{N}_{j_\ell,k_\ell}, \ell \in [L] \mid v \in \mathscr{N}_{jk}, g, X]. \tag{37}$$

Note that there is no condition that the node w_ℓ at step ℓ is distinct from the earlier nodes $w_{\ell'}, \ell' \in \{0\} \cup [\ell-1]$. With high probability each w_ℓ will be new, and the resultant subgraph g will be a tree with L distinct added nodes (not including the root) and L edges. With small probability one or more of the w_ℓ will be preexisting, i.e. equal to $w_{\ell'}$ for some $\ell' \in \{0\} \cup [\ell-1]$: in this case the subgraph g will have $M < L$ added nodes, will have cycles and not be a tree.

The following sequences of numbers are determined given X and g:

- $e_{j,k}(\ell)$ is the number of available j-stubs connected to (j,k) nodes after ℓ wiring steps;
- $e_{k,j}(\ell)$ is the number of available k-stubs connected to (j,k) nodes after ℓ wiring steps.

- $e^-_j(\ell) := \sum_k e_{j,k}(\ell)$ and $e^+_k(\ell) := \sum_j e_{k,j}(\ell)$ are the number of available j-stubs and k-stubs respectively after ℓ wiring steps.

Note that $e_{j,k}(0) = ju_{jk}$ and $e_{k,j}(0) = ku_{jk}$, and both decrease by at most 1 at each step.

The analysis of configuration probabilities that follows is inductive on the step ℓ.

Theorem 4.2 *Consider the ACG sequence with* (P, Q) *supported on* $(\{0\} \cup [K])^2$. *Let* g *be any fixed finite configuration rooted to* $w_0 \in \mathcal{N}_{jk}$, *with* M *added nodes and* $L \geq M$ *edges, labelled by the node-type sequence* $(j_m, k_m)_{m\in[M]}$. *Then, as* $N \to \infty$, *the joint probability conditioned on* X,

$$p = \mathbb{P}[w_m \in \mathcal{N}_{j_m k_m}, m \in [M] \mid v \in \mathcal{N}_{jk}, g, X],$$

is given by

$$\prod_{m\in[M],(m',m)\ \textit{out-edge}} P_{k_m|j_m} Q_{j_m|k_{m'}} \prod_{m\in[M],\ \textit{in-edge}} P_{j_m|k_m} Q_{k_m|j_{m'}} \left[1 + O(N^{-1/2})\right] \quad (38)$$

if $L = M$ *and thus* g *is a tree. If* $L > M$ *and so* g *has cycles then*

$$p = O(N^{M-L}). \quad (39)$$

The factors in (38) *depend on whether the mth edge is an in- or out-edge and* $m' \in \{0\} \cup [m-1]$ *numbers the node to which* w_m *attaches.*

Remarks 1 1. Formula (38) shows clearly what is meant by saying that configuration graphs are *locally tree-like* as $N \to \infty$. It means the number of occurrences of any fixed finite size graph g with cycles embedded within a configuration graph of size N remains bounded with high probability as $N \to \infty$.

2. Even more interesting is that (38) shows that large configuration graphs exhibit a strict type of conditional independence. Selection of any root node v of the tree graph g splits it into two (possibly empty) trees g_1, g_2 with node-types $(j_m, k_m), m \in [M_1]$ and $(j_m, k_m), m \in [M_1 + M_2] \setminus [M_1]$ where $M = M_1 + M_2$. When we condition on the node-type of v, (38) shows that the remaining node-types form independent families:

$$\begin{aligned}\mathbb{P}[w_m \in \mathcal{N}_{j_m k_m}, m \in [M], g \mid X, v \in \mathcal{N}_{jk}] &= \mathbb{P}[w_m \in \mathcal{N}_{j_m k_m}, m \in [M_1], g_1 \mid X, v \in \mathcal{N}_{jk}] \\ &\times \mathbb{P}[w_m \in \mathcal{N}_{j_m k_m}, m \in [M_1 + M_2] \setminus [M_1], g_2 \mid X, v \in \mathcal{N}_{jk}]. \quad (40)\end{aligned}$$

We call this deep property of the general configuration graph the *locally tree-like independence property* (LTI property). In [9], the LTI property provides the key to unravelling cascade dynamics in large configuration graphs.

Proof of Theorem 4.2 First, suppose Step 1 generates the node-type sequence X. Conditioned on X, now suppose the first step generates an in-edge (w_1, v). Then, by refining Part 2 of Proposition 1, the conditional probability that node w_1 has degree j_1, k_1 can be written

$$\begin{aligned}
&\frac{\mathbb{P}[w_1 \in \mathscr{N}_{j_1k_1}, w_0 \in \mathscr{N}_{jk} \mid g, X]}{\mathbb{P}[w_0 \in \mathscr{N}_{jk} \mid g, X]} \\
&\quad = \frac{C^{-1}(e^-(0), e^+(0)) e_{k_1,j_1}(0) e^-_{j,k}(0) Q_{k_1j} C(e^-(1), e^+(1))}{C^{-1}(e^-(0), e^+(0)) \sum_{k'} e^+_{k'}(0) e^-_{j,k}(0) Q_{k'j} C(e^-(1), e^+(1))} \\
&\quad = \left(\frac{e_{k_1,j_1}(0) e^-_{j,k}(0)}{e^+_{k_1}(0) e^-_j(0)}\right) \left(\frac{\mathbb{E}[e_{k_1j} \mid e^-(0), e^+(0)]}{E}\right) \left(\frac{e^-_{j,k}(0)}{E}\right)^{-1} \\
&\quad = \left(\frac{k_1 u_{k_1,j_1}}{k_1 v^+_{k_1}}\right) \left(\frac{\mathbb{E}[e_{k_1j} \mid e^-(0), e^+(0)]}{e^-_j(0)}\right).
\end{aligned}$$

Be aware that $C(e^-(1), e^+(1))$ in the denominator after the first equality depends on k' and hence does not cancel a factor in the numerator. Now, for $N \to \infty$, Part 2 of the Corollary applies to the second factor, and (13) applies to the first factor, and shows that for the case of an in-edge on the first step, with high probability, X is such that:

$$\mathbb{P}[w_1 \in \mathscr{N}_{j_1k_1} \mid v \in \mathscr{N}_{jk}, g, X] \; = \; P_{j_1|k_1} \, Q_{k_1|j} \left[1 + O(N^{-1/2})\right].$$

The case of an out-edge is similar.

Now we continue conditionally on X from Step 1 and assume inductively that (38) is true for $M - 1$ and prove it for M. Suppose the final node w_M is in-connected to the node $w_{M'}$ for some $M' \leq M$. The ratio $\mathbb{P}[w_m \in \mathscr{N}_{j_mk_m}, m \in [M] \mid v \in \mathscr{N}_{jk}, g, X] / \mathbb{P}[w_m \in \mathscr{N}_{j_mk_m}, m \in [M-1] \mid v \in \mathscr{N}_{jk}, g, X]$ can be treated just as in the previous step and shown to be

$$\left(\frac{e_{k_M j_M}(M-1)}{e^+_{k_M}(M-1)}\right) \left(\frac{\mathbb{E}[e_{k_M j_{M'}} \mid e^-(M-1), e^+(M-1)]}{e^-_{j_{M'}}(M-1)}\right)$$

which with high probability equals

$$\mathbb{P}[w_1 \in \mathscr{N}_{j_1k_1} \mid v \in \mathscr{N}_{jk}, g, X] \; = \; P_{j_M|k_M} \, Q_{k_M|j_{M'}} \left[1 + O(N^{-1/2})\right].$$

The case w_M is out-connected to the node $w_{M'}$ is similar.

The first step m that a cycle is formed can be treated by imposing a condition that $w_m = w_{m''}$ for some fixed $m'' < m$. One finds that the conditional probability of this is

$$\begin{aligned}&\mathbb{P}[w_m = w_{m''}, w_\ell \in \mathcal{N}_{j_\ell k_\ell}, \ell \in [m-1] \mid v \in \mathcal{N}_{jk}, g, X]\\&\quad = \frac{k_{m''}}{e^+_{k_{m''}}(m-1)} \times \mathbb{P}[w_\ell \in \mathcal{N}_{j_\ell k_\ell}, \ell \in [m-1] \mid v \in \mathcal{N}_{jk}, g, X].\end{aligned}$$

The first factor is $O(N^{-1})$ as $N \to \infty$, which proves the desired statement (39) for cycles.

Finally, since (39) is true for cycles, with high probability all finite configurations are trees. Therefore their asymptotic probability laws are given by (38), as required. □

5 Approximate ACG Simulation

It was observed in Sect. 1.2 that Step 1 of the configuration graph construction draws a sequence $(j_i, k_i)_{i\in[N]}$ of node types that is iid with the correct distribution P, but is only feasible, $\sum_i (k_i - j_i) = 0$, with small probability. Step 2 of the exact ACG algorithm in Sect. 2 is even less feasible in practice. Practical simulation algorithms address the first problem by *clipping* the drawn node bidegree sequence when the discrepancy $D = D_N := \sum_i (k_i - j_i)$ is not too large, meaning it is adjusted by a small amount to make it feasible, without making a large change in the joint distribution. Step 1 of the following simulation algorithm generalizes slightly the method introduced by [4] who verify that the effect of clipping vanishes with high probability as $N \to \infty$. The difficulty with Step 2 of the ACG construction is overcome in this section by an approximate sequential wiring algorithm that we conjecture has the correct asymptotic properties. An alternative simulation algorithm that also has the correct asymptotics of the ACG model has been studied in [5].

The *approximate assortative configuration graph (ACG) simulation algorithm* for multigraphs of size N, parametrized by the node-edge type distribution pair (P, Q) that have support on the finite set $(j, k) \in \{0, 1, \ldots, K\}^2$, involves choosing a suitable threshold $T = T(N)$ and modifying the steps identified in Sect. 2:

1. Draw a sequence of N node-type pairs $X = ((j_1, k_1), \ldots, (j_N, k_N))$ independently from P, and accept the draw if and only if $0 < |D| \le T(N)$. When the sequence $(j_i, k_i)_{i\in[N]}$ is accepted, the sequence is adjusted by adding a few stubs, either in- or out- as needed. First draw a random subset $\sigma \subset \mathcal{N}$ of size $|D|$ with uniform probability $\binom{N}{|D|}^{-1}$, and then define the feasible sequence $\tilde{X} = (\tilde{j}_i, \tilde{k}_i)_{i\in[N]}$ by adjusting the degree types for $i \in \sigma$ as follows:

$$\tilde{j}_i = j_i + \xi_i^-; \quad \xi_i^- = \mathbb{I}(i \in \sigma, D > 0) \tag{41}$$

$$\tilde{k}_i = k_i + \xi_i^+; \quad \xi_i^+ = \mathbb{I}(i \in \sigma, D < 0). \tag{42}$$

2. Conditioned on $\tilde{X}$, the result of Step 1, randomly wire together available in and out stubs *sequentially*, with suitable weights, to produce the sequence of edges W. At each $\ell = 1, 2, \ldots, E$, match from available in-stubs and out-stubs weighted according to their degrees j, k by
$$C^{-1}(\ell)\frac{Q_{kj}}{Q^+_k Q^-_j}. \tag{43}$$
In terms of the bivariate random process $(e^-_j(\ell), e^+_k(\ell))$ with initial values $(e^-_j(1), e^+_k(1)) = (e^-_j, e^+_k)$ that at each ℓ counts the number of available degree j in-stubs and degree k out-stubs, the ℓ dependent normalization factor $C(\ell)$ is given by:
$$C(\ell) = \sum_{jk} e^-_j(\ell)e^+_k(\ell)\frac{Q_{kj}}{Q^+_k Q^-_j}. \tag{44}$$

Remark 5.1 Chen and Olvera-Cravioto, [4], addresses the clipping in Step 1 and shows that the discrepancy of the approximation is negligible as $N \to \infty$:

Theorem 5.3 *Fix $\delta \in (0, 1/2)$, and for each N let the threshold be $T(N) = N^{1/2+\delta}$. Then:*

1. *The acceptance probability $\mathbb{P}[|D_N| \le T(N)] \to 1$ as $N \to \infty$;*
2. *For any fixed finite M, Λ, and bounded function $f : (\mathbb{Z}_+ \times \mathbb{Z}_+)^M \to [-\Lambda, \Lambda]$*
$$\left|\mathbb{E}[f\left((\tilde{j}_i, \tilde{k}_i)_{i\in[M]}\right)] - \mathbb{E}[f\left((\hat{j}_i, \hat{k}_i)_{i\in[M]}\right)]\right| \to 0\,; \tag{45}$$
where $(\hat{j}_i, \hat{k}_i)_{i\in[M]}$ is an independent sequence of P distributed random variables.
3. *The following limits in probability hold:*
$$\frac{1}{N}\tilde{u}_{jk} \xrightarrow{\mathrm{P}} P_{jk}, \quad \frac{1}{N}\tilde{v}^+_k \xrightarrow{\mathrm{P}} P^+_k, \quad \frac{1}{N}\tilde{v}^-_j \xrightarrow{\mathrm{P}} P^-_j. \tag{46}$$

Similarly it is intuitively clear that the discrepancy of the approximation in Step 2 is negligible as $N \to \infty$. As long as $e^-_j(\ell), e^+_k(\ell)$ are good approximations of $(E-\ell)Q^-_j, (E-\ell)Q^+_k$, (43) shows that the probability that edge ℓ has type (k,j) will be approximately Q_{kj}. Since the detailed analysis of this problem is not yet complete, we state the desired properties as a conjecture:

Conjecture 1 In the approximate assortative configuration graph construction with probabilities P, Q, the following convergence properties hold as $N \to \infty$.

1. The fraction of type (k,j) edges in the matching sequence $(k_\ell, j_\ell)_{\ell\in[E]}$ concentrates with high probability around the nominal edge distribution Q_{kj}:
$$\frac{e_{kj}}{E} = Q_{kj} + o(1). \tag{47}$$

2. For any fixed finite number L, the first L edges $\ell \in [L]$ have degree sequence $(k_\ell, j_\ell)_{\ell \in [L]}$ that converges in distribution to $(\hat{k}_\ell, \hat{j}_\ell)_{\ell \in [L]}$, an independent sequence of identical Q distributed random variables.

Although this conjecture is not yet proven, extensive simulations have verified the consistency of the approximate ACG algorithm with the theoretical large N probabilities.

6 Conclusions

The ACG algorithm that is the main contribution of this paper, while unwieldy from the point of view of simulation, has rich combinatorial properties that have made it amenable to exact study via the Laplace asymptotic method. The consequences of this approach have not yet been explored in depth. For example, it will be of interest to get more accurate bounds on the large N asymptotics by using higher order Laplace methods, allowing us to better understand for example, P and Q distributions with fat-tails. Such bounds would also give a better understanding of the locally-treelike property of the model.

Numerous possible methods for efficient simulation of ACG graphs, including the method conjectured in Sect. 5 to be consistent, and the method studied in [5], can be imagined. However, it will take some time to decide which simulation methods are both consistent with the ACG model and computationally efficient.

Given the potential of the ACG model to describe a wide range of real world networks for which the original configuration graph model is inadequate, including the cascade models for systemic risk in financial networks that was its original motivation, future investigations along these lines are likely to prove fruitful.

References

1. M. Bech and E. Atalay. The topology of the federal funds market. *Physica A: Statistical Mechanics and its Applications*, 389(22):5223–5246, 2010.
2. B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1:311, 1980.
3. B. Bollobás. *Random Graphs*. Cambridge studies in advanced mathematics. Cambridge University Press, 2 edition, 2001.
4. Ningyuan Chen and Mariana Olvera-Cravioto. Directed random graphs with given degree distributions. *Stochastic Systems*, 3(1):147–186, 2013.
5. Philippe Deprez and Mario V. Wüthrich. Construction of directed assortative configuration graphs. arXiv:1510.00575, October 2015.
6. Arthur Erdélyi. *Asymptotic expansions*. Dover, New York, 1956.
7. P. Erdös and A. Rényi. On random graphs. *I. Publ. Math. Debrecen*, 6:290–297, 1959.
8. T. R. Hurd. Saddlepoint approximation. In Rama Cont, editor, *Encyclopedia of Quantitative Finance*. John Wiley & Sons, Ltd, 2010.

9. T. R. Hurd. *Contagion! Systemic Risk in Financial Networks*. SpringerBriefs in Quantitative Finance. Springer Verlag, Berlin Heidelberg New York, 2016. Available at http://ms.mcmaster.ca/tom/tom.html.
10. T. R. Hurd, Davide Cellai, Sergey Melnik, and Quentin Shao. Double cascade model of financial crises. *International Journal of Theoretical and Applied Finance*, (to appear), 2016. http://arxiv.org/abs/1310.6873v3.
11. Svante Janson. The probability that a random multigraph is simple. *Combinatorics, Probability and Computing*, 18:205–225, 3 2009.
12. Robert M. May and Nimalan Arinaminpathy. Systemic risk: the dynamics of model banking systems. *Journal of The Royal Society Interface*, 7(46):823–838, 2010.
13. Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2–3):161–180, 1995.
14. Kimmo Soramäki, M. Bech, J. Arnold, R. Glass, and W. Beyeler. The topology of interbank payment flows. *Physica A: Statistical Mechanics and its Applications*, 379(1):317–333, 2007.
15. R. van der Hofstad. *Random Graphs and Complex Networks*. unpublished, available at http://www.win.tue.nl/rhofstad/NotesRGCN.html, 2016. Book, to be published.

Part VII
Life and Environmental Sciences

Coexistence in the Face of Uncertainty

Sebastian J. Schreiber

Lest men believe your tale untrue, keep probability in view.

—John Gay

Abstract Over the past century, nonlinear difference and differential equations have been used to understand conditions for coexistence of interacting populations. However, these models fail to account for random fluctuations due to demographic and environmental stochasticity which are experienced by all populations. I review some recent mathematical results about persistence and coexistence for models accounting for each of these forms of stochasticity. Demographic stochasticity stems from populations and communities consisting of a finite number of interacting individuals, and often are represented by Markovian models with a countable number of states. For closed populations in a bounded world, extinction occurs in finite time but may be preceded by long-term transients. Quasi-stationary distributions (QSDs) of these Markov models characterize this meta-stable behavior. For sufficiently large "habitat sizes", QSDs are shown to concentrate on the positive attractors of deterministic models. Moreover, the probability extinction decreases exponentially with habitat size. Alternatively, environmental stochasticity stems from fluctuations in environmental conditions which influence survival, growth, and reproduction. Stochastic difference equations can be used to model the effects of environmental stochasticity on population and community dynamics. For these models, stochastic persistence corresponds to empirical measures placing arbitrarily little weight on arbitrarily low population densities. Sufficient and necessary conditions for stochastic persistence are reviewed. These conditions involve weighted combinations of Lyapunov exponents corresponding to "average" per-capita growth rates of rare species. The results are illustrated with how climatic variability influenced the dynamics of Bay checkerspot butterflies, the persistence

S.J. Schreiber (✉)
Department of Evolution and Ecology, University of California, Davis, CA 95616, USA
e-mail: sschreiber@ucdavis.edu

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_12

of coupled sink populations, coexistence of competitors through the storage effect, and stochastic rock-paper-scissor communities. Open problems and conjectures are presented.

Keywords Random difference equations • Stochastic population dynamics • Coexistence • Quasi-stationary distributions • Demographic noise • Environmental stochasticity • Markov chains

1 Introduction

A long standing, fundamental question in biology is "what are the minimal conditions to ensure the long-term persistence of a population or the long-term coexistence of interacting species?" The answers to this question are essential for guiding conservation efforts for threatened and endangered species, and identifying mechanisms that maintain biodiversity. Mathematical models have and continue to play an important role in identifying these potential mechanisms and, when coupled with empirical work, can test whether or not a given mechanism is operating in a specific population or ecological community [1]. Since the pioneering work of [35] and [56] on competitive and predator–prey interactions, [41, 54] on host–parasite interactions, and [30] on disease outbreaks, nonlinear difference and differential equations have been used to understand conditions for persistence of populations or communities of interacting species. For these deterministic models, persistence or species coexistence is often equated with an attractor bounded away from the extinction states in which case persistence holds over an infinite time horizon [47]. However (with apologies to John Gay), lest biologists believe this theory untrue, the models need to keep probability in view. That is, all natural populations exhibit random fluctuations due to mixture of intrinsic and extrinsic factors known as demographic and environmental stochasticity. The goal of this chapter is to present models that account for these random fluctuations, review some mathematical methods for analyzing these stochastic models, and illustrate how these random fluctuations hamper or facilitate population persistence and species coexistence.

Demographic stochasticity corresponds to random fluctuations due to populations consisting of a finite number of individuals whose fates aren't perfectly correlated. That is, even if all individuals in a population appear to be identical, some undetectable differences between individuals (e.g. in their physiology or microenvironment) result in some individuals dying while others survive. To capture these "unknowable" differences, models can assign the same probabilities of dying to each individuals and treat survival amongst individuals as independent flips of a coin—heads life, tails death. Similarly, surviving individuals may differ in the number of offspring they produce despite appearing to be identical. To capture these unknowable differences, the number of offspring produced by these individuals are modeled as independent draws from the same probability distribution. The resulting stochastic models accounting for these random fluctuations typically correspond to Markov chains on a finite or countable state space[1] e.g. the numbers of individuals, $0, 1, 2, 3, \ldots$, in a population. When these models represent populations or communities whose numbers tend to stay bounded and have no immigration, the populations in these models always go extinct in finite time [10]. Hence, unlike deterministic models, the asymptotic behavior of these stochastic models is trivial: eventually no one is left. This raises the following basic question about the relationship between models accounting for demographic stochasticity and their deterministic counterparts:

> "Any population allowing individual variation in reproduction, ultimately dies out–unless it grows beyond all limits, an impossibility in a bounded world. Deterministic population mathematics on the contrary allows stable asymptotics. Are these artifacts or do they tell us something interesting about quasi-stationary stages of real or stochastic populations?"—Peter [29]

As it turns out, there is a strong correspondence between the quasi-stationary behavior of the stochastic models and the attractors of an appropriately defined mean-field model. Moreover, this correspondence highlights a universal scaling relationship between extinction times and the size of the habitat where the species live. These results and their applications are the focus of the first half of this review.

While demographic stochasticity affects individuals independently, environmental stochasticity concerns correlated demographic responses (e.g. increased survival, growth or reproduction) among individuals. These correlations often stem from individuals experiencing similar fluctuations in environmental conditions (e.g. temperature, precipitation, winds) which impact their survival, growth, or reproduction. Models driven by randomly fluctuating parameters or Brownian motions, such as random difference equations or stochastic differential equations, can capture these sources of random fluctuations. Unlike models for demographic stochasticity, these Markov chains always live on uncountable state spaces where the non-negative reals represent densities of populations of sufficiently large size that one can ignore the effects of being discrete and finite. Consequently, like their deterministic counterparts, extinction in these random difference equations only occurs asymptotically, and persistence is equated with tendency to stay away from low densities [11]. Understanding what this exactly means, reviewing methods for verifying this stochastic form of persistence, and applying these methods to gain insights about population persistence and species coexistence are the focus of the second half of this review.

Of course, all population systems experience a mixture of demographic and environmental stochasticity. While the theoretical biology literature is replete with models accounting for each of these forms of noise separately, I know of no studies that rigorously blend the results presented in this review. Hence, I conclude by discussing some open problems and future challenges at this mathematical interface.

[1] See, however, the discussion for biologically motivated uncountable state spaces.

2 Demographic Stochasticity

To model finite populations and account for demographic stochasticity, we consider Markov chains on a countable state space which usually is the non-negative cone of the integer lattice. Many of these stochastic models have a deterministic counterpart, sometimes called the "deterministic skeleton" or the "mean field model". As I discuss below, these deterministic models can provide some useful insights about the transient behavior of the stochastic models and when coupled with large deviation theory provide insights into the length of these transients.

To get a flavor of the types of models being considered, lets begin with a stochastic counterpart to the discrete-time Lotka-Volterra equations. This example motivates the main results and will illustrate their applicability.

Example 1 (Poisson Lotka-Volterra Processes) The continuous time Lotka-Volterra equations form the bedrock for much of community ecology theory. While there are various formulations of their discrete-time counterparts, a particularly pleasing one that retains several key dynamical features of the continuous-time models was studied by [26]. These models keep track of the densities $x_t = (x_{1,t}, \dots, x_{k,t})$ of k interacting species, where the subscripts denote the species identity i and time t (e.g. year or day). As with the classical continuous time equations, there is a matrix $A = (a_{ij})_{i,j}$ where a_{ij} corresponds to the "per-capita" effect of species j on species i and a vector $r = (r_1, \dots, r_k)$ of the "intrinsic per-capita growth rates" for all of the species. With this notation, the equations take on the form:

$$x_{i,t+1} = x_{i,t} \exp\left(r_i + \sum_j a_{ij} x_{j,t}\right) =: F_i(x_t) \text{ with } i = 1, 2, \dots, k. \qquad (1)$$

The state space for these dynamics are given by the non-negative orthant

$$\mathbb{R}_+^k = \{x \in \mathbb{R}^k : x_i \geq 0 \text{ for all } i\}$$

of the k-dimensional Euclidean space $\mathbb{R}^k$.

To define the Poisson Lotka-Volterra process, let $1/\varepsilon$ be the size of the habitat in which the species live. Let $N_t^\varepsilon = (N_{1,t}^\varepsilon, \dots, N_{k,t}^\varepsilon)$ denote the vector of species abundances which are integer-valued. Then the density of species i is $X_{i,t}^\varepsilon = \varepsilon N_{i,t}^\varepsilon$. Over the next time step, each individual replaces itself with a Poisson number of individuals with mean

$$\exp\left(r_i + \sum_j a_{ij} X_{j,t}^\varepsilon\right).$$

If the individuals update independent of one another, then $N^{\varepsilon}_{i,t+1}$ is a sum of $N^{\varepsilon}_{i,t}$ independent Poisson random variables. Thus, $N^{\varepsilon}_{i,t+1}$ is also Poisson distributed with mean

$$N^{\varepsilon}_{i,t}\exp\left(r_i+\sum_j a_{ij}X^{\varepsilon}_{j,t}\right)=F_i(X^{\varepsilon}_t)/\varepsilon.$$

Namely,

$$\mathbb{P}[X^{\varepsilon}_{i,t+1}=\varepsilon j|X^{\varepsilon}_t=x]=\mathbb{P}[N^{\varepsilon}_{i,t+1}=j|X^{\varepsilon}_t=x]=\exp(-F_i(x)/\varepsilon)\frac{(F_i(x)/\varepsilon)^j}{j!}. \quad (2)$$

The state space for N^{ε}_t is the non-negative, k dimensional integer lattice

$$\mathbb{Z}^k_+=\{(z_1,\dots,z_k):z_i \text{ are non-negative integers}\}$$

while the state space for X^{ε}_t is the non-negative, rescaled integer lattice

$$\varepsilon\mathbb{Z}^k_+=\{(\varepsilon z_1,\dots,\varepsilon z_k):z_i \text{ are non-negative integers}\}.$$

Now consider a solution to deterministic model x_t and the stochastic process X^{ε}_t initiated at the same densities $x_0=X^{\varepsilon}_0=x$. To see how likely X^{ε}_t deviates from x_t, we use Chebyshev's inequality. As the mean and variance of a Poisson random variable are equal, Chebyshev's inequality implies

$$\mathbb{P}\left[|X^{\varepsilon}_{i,1}-x_{i,1}|\geq\delta\middle|X^{\varepsilon}_0=x_0=x\right]\leq\frac{\mathrm{Var}[X^{\varepsilon}_{i,1}]}{\delta^2}=\frac{\varepsilon^2\mathrm{Var}[N^{\varepsilon}_{i,1}]}{\delta^2}=\frac{\varepsilon F_i(x)}{\delta^2} \quad (3)$$

where Var[X] denotes the variance of a random variable X. In words, provided the habitat size $1/\varepsilon$ is sufficiently large, a substantial deviation between X^{ε}_1 and x_1 is unlikely. In fact, one can show that over any finite time interval $[1,T]$, the stochastic dynamics are likely to be close to the deterministic dynamics over the time interval $[1,T]$ provided the habitat size $1/\varepsilon$ is sufficiently large:

$$\lim_{\varepsilon\to0}\mathbb{P}\left[\max_{1\leq i\leq k,1\leq t\leq T}|X^{\varepsilon}_{i,t}-x_{i,t}|\geq\delta\middle|X^{\varepsilon}_0=x_0=x\right]=0. \quad (4)$$

Figure 1 illustrates this fact for a Poisson Lotka-Volterra process with two competing species. Equation (4) is the discrete-time analog of a result derived by [34] for continuous-time Markov chains. [34] also provides "second-order" approximations for finite time intervals using Gaussian processes and stochastic differential equations. While these approximations are also useful for discrete-time models, we do not review them here.

Despite X^{ε}_t stochastically tracking x_t with high probability for long periods of time, eventually their behavior diverges as Poisson Lotka-Volterra processes go extinct in finite time or exhibit unbounded growth.

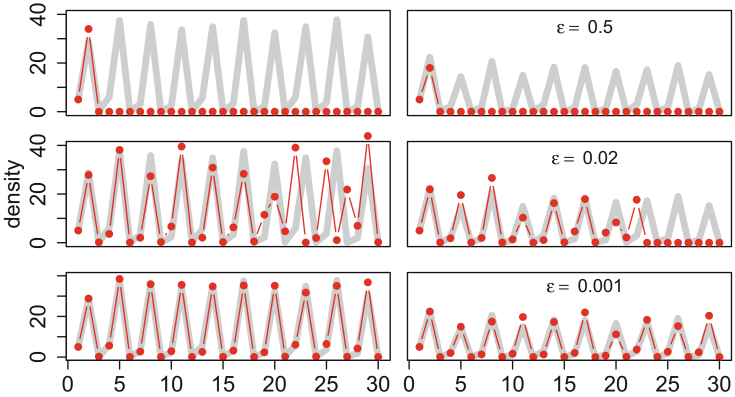


Fig. 1 Realizations of a Poisson Lotka-Voltera process with two competing species (species 1 on the left, species 2 on the right). The deterministic dynamics are shown as a *thick gray line*. Stochastic realizations are shown in *red*. Each row corresponds to a different habitat size $1/\varepsilon$. Parameter values: A is the matrix with rows $(-0.2, 0.1)$, $(-0.15, 0.2)$, and $r = (3.25, 3.25)$ for the model described in Example 1

Proposition 1 *Let X_t^ε be a Poisson Lotka-Volterra process with $\varepsilon > 0$. Then*

$$\mathbb{P}\left[\{X_t^\varepsilon = 0 \, \textit{for some } t\} \cup \{\lim_{t\to\infty} \sum_i X_{i,t}^\varepsilon = \infty\}\right] = 1$$

Furthermore, if F is pre-compact i.e. $F(\mathbb{R}_+^k) \subset [0,m]^k$ for some $m \geq 0$, then

$$\mathbb{P}\left[\{X_t^\varepsilon = 0 \, \textit{for some } t\}\right] = 1$$

The strategy used to prove the first statement of the proposition is applicable to many models of closed populations. The key ingredients are that there is a uniform lower bound to the probability of any individual dying, and individuals die independently of one another [10]. Proving, however, that extinction always occurs with probability one requires additional elements which aren't meet by all ecological models, but is meet for "realistic" models.

Proof For the first assertion, take any integer $m > 0$. Let

$$\beta = \min_{x\in[0,m]^k} \mathbb{P}[X_1^\varepsilon = 0 | X_0^\varepsilon = x] = \min_{x\in[0,m]^k} \exp\left(-\sum_{i=1}^k F_i(x)/\varepsilon\right) > 0.$$

Next we use the following standard result in Markov chain theory [17, Theorem 2.3 in Chapter 5].

Proposition 2 *Let X be a Markov chain and suppose that*

$$\mathbb{P}\left[\bigcup_{s=1}^{\infty}\{X_{t+s}\in C\}\middle|X_t\right]\geq \beta>0 \text{ on } \{X_t\in B\}.$$

Then

$$P\left[\{X_t \text{ enters } B \text{ infinitely often}\}\setminus\{X_t \text{ enters } C \text{ infinitely often}\}\right]=0.$$

Let $\mathscr{B}_m = \{X_t^\varepsilon \text{ enters } [0,m]^k \text{ infinitely often}\}$ and $\mathscr{E} = \{X_t^\varepsilon = 0 \text{ for some } t\}$. Proposition 2 with $B=[0,m]^k$ and $C=\{0\}$ implies that

$$\mathbb{P}\left[\mathscr{B}_m\setminus\mathscr{E}\right]=0. \tag{5}$$

The complement of the event $\cup_m\mathscr{B}_m$ equals the event $\mathscr{A}=\{\lim_{t\to\infty}\sum_i X_{i,t}^\varepsilon=\infty\}$. As $\mathscr{B}_m$ is an increasing sequence of events,

$$\begin{aligned}1&=\mathbb{P}\left[\mathscr{A}\cup\{\cup_m\mathscr{B}_m\}\right]\\&=\lim_{m\to\infty}\mathbb{P}\left[\mathscr{A}\cup\mathscr{B}_m\right]\\&\leq\lim_{m\to\infty}\mathbb{P}\left[\mathscr{A}\cup\mathscr{E}\right]\end{aligned}$$

where the final inequality follows from (5). This completes the proof of the first assertion.

To prove the second assertion, assume that there exists $m>0$ such that $F(x)\in[0,m]^k$ for all $x\in\mathbb{R}_+^k$ i.e. F is pre-compact. Define

$$\begin{aligned}\beta&=\inf_{x\in\mathbb{R}_+^k}\mathbb{P}[X_{t+1}^\varepsilon=0|X_t=x]\\&=\inf_{x\in\mathbb{R}_+^k}\exp\left(-\sum_i F_i(x)/\varepsilon\right)\\&\geq\exp(-k\,m/\varepsilon)\end{aligned}$$

Applying Proposition 2 with $B=\mathbb{R}_+^k$ and $C=\{0\}$ completes the proof of the second assertion. □

Equation (4) and Proposition 1 raise two fundamental questions about these stochastic, finite population models: How long before extinction occurs? Prior to extinction what can one say about the transient population dynamics? To get some insights into both of these questions, we build on the work of [21] and [31] on random perturbations of dynamical systems, and [3] on quasi-stationary distributions.

2.1 Random Perturbations and Quasi-Stationary Distributions

The Poisson Lotka-Volterra process (Example 1) illustrates how Markovian models can be viewed as random perturbations of a deterministic model. To generalize this idea, consider a continuous, precompact[2] map $F : \mathscr{S} \to \mathscr{S}$, where $\mathscr{S}$ is a closed subset of $\mathbb{R}^k$. F will be the deterministic skeleton of our stochastic models. *A random perturbation of* F is a family of Markov chains $\{X^\varepsilon\}_{\varepsilon>0}$ on $\mathscr{S}$ whose transition kernels

$$p^\varepsilon(x,\Gamma) = \mathbb{P}\left[X^\varepsilon_{t+1} \in \Gamma \mid X^\varepsilon_t = x\right] \text{ for all } x \in \mathscr{S} \text{ and Borel sets } \Gamma \subset \mathscr{S}$$

enjoy the following hypothesis:

Hypothesis 2.1 *For any $\delta > 0$,*

$$\lim_{\varepsilon\to 0} \sup_{x\in\mathscr{S}} p^\varepsilon\left(x, \mathscr{S} \setminus N^\delta(F(x))\right) = 0$$

where $N^\delta(y) := \{x \in \mathscr{S} : \|y - x\| < \delta\}$ denotes the δ-neighborhood of a point $y \in \mathscr{S}$.

Hypothesis 2.1 implies that the Markov chains X^ε converge to the deterministic limit as $\varepsilon \downarrow 0$ i.e. the probability of X^ε_1 being arbitrarily close to $F(x)$ given $X^\varepsilon_0 = x$ is arbitrarily close to one for ε sufficiently small. Hence, one can view F as the "deterministic skeleton" which gets clothed by the stochastic dynamic X^ε. The next example illustrates how to verify the Poisson Lotka-Volterra process is a random perturbation of the Lotka-Volterra difference equations.

Example 2 (The Poisson Lotka-Volterra Processes Revisited) Consider the Poisson Lotka-Volterra processes from Example 1 where $F(x) = (F_1(x), \dots, F_k(x))$ and $F_i(x) = x_i \exp(r_i + \sum_j a_{ij}x_j)$ and $\mathscr{S} = \mathbb{R}^k_+$. For many natural choices of r_i and a_{ij}, [26] have shown there exists $C > 0$ such that $F(\mathscr{S}) \subset [0, C]^k$ i.e. F is pre-compact. While the corresponding Lotka-Volterra process X^ε lives on $\varepsilon\mathbb{Z}^k_+$, the process can be extended to all of $\mathscr{S}$ by allowing X^ε_0 to be any point in $\mathscr{S}$ and update with the transition probabilities of (2). With this extension, X^ε_1 always lies in $\varepsilon\mathbb{Z}^k_+$ and p^ε is characterized by the following probabilities

$$p^\epsilon(x,\{y\}) = \prod_{i=1}^k \exp(-F_i(x)/\varepsilon)\frac{(F_i(x)/\varepsilon)^{j_i}}{j_i!} \text{ for } y = \varepsilon(j_1,\dots,j_k) \in \varepsilon\mathbb{Z}^k_+, x \in \mathscr{S}$$

and 0 otherwise. With this extension, Hypothesis 1 for the Lotka-Volterra process follows from equation (3).

[2] Namely, there exists $C > 0$ such that $F(\mathscr{S})$ lies in $[0, C]^k$.

As with the Poisson Lotka-Volterra process, stochastic models of interacting populations without immigration always have absorbing states $\mathscr{S}_0 \subset \mathscr{S}$ corresponding to the loss of one or more populations. Hence, we restrict our attention to models which satisfy the following standing hypothesis:

Hypothesis 2.2 *The state space $\mathscr{S}$ can be written $\mathscr{S} = \mathscr{S}_0 \cup \mathscr{S}_+$, where*

- *$\mathscr{S}_0$ is a closed subset of $\mathscr{S}$;*
- *$\mathscr{S}_0$ and $\mathscr{S}_+$ are positively F-invariant, i.e $F(\mathscr{S}_0) \subseteq \mathscr{S}_0$ and $F(\mathscr{S}_+) \subseteq \mathscr{S}_+$;*
- *the set $\mathscr{S}_0$ is assumed to be absorbing for the random perturbations:*

$$p^\varepsilon(x, \mathscr{S}_+) = 0, \textit{ for all } \varepsilon > 0,\ x \in \mathscr{S}_0. \tag{6}$$

- *absorption occurs in finite time with probability one:*

$$\mathbb{P}\left[X_t^\varepsilon \in \mathscr{S}_0 \textit{ for some } t \geq 1 | X_0^\varepsilon = x\right] = 1$$

for all $x \in \mathscr{S}$ and $\varepsilon > 0$.

The final bullet point implies that extinction of one or more species is inevitable in finite time. For example, Proposition 1 implies this hypothesis for Poisson Lotka-Volterra processes whenever F is pre-compact.

Despite this eventual absorption, the process X^ε may spend exceptionally long periods of time in the set $\mathscr{S}_+$ of transient states provided that $\varepsilon > 0$ is sufficiently small. This "metastable" behavior may correspond to long-term persistence of an endemic disease, long-term coexistence of interacting species as in the case of the Poisson Lotka-Volterra process, or maintenance of a genetic polymorphism. One approach to examining these metastable behaviors are quasi-stationary distributions which are invariant distributions when the process is conditioned on non-absorption.

Definition 2.1 A probability measure μ_ε on $\mathscr{S}_+$ is a *quasi-stationary distribution (QSD)* for p^ε provided there exists $\lambda_\varepsilon \in (0, 1)$ such that

$$\int_{\mathscr{S}_+} p^\varepsilon(x, \Gamma)\mu_\varepsilon(dx) = \lambda_\varepsilon \mu_\varepsilon(\Gamma) \text{ for all Borel sets } \Gamma \subset \mathscr{S}_+.$$

Equivalently, dropping the ε superscript and subscripts, a QSD μ satisfies the identity

$$\mu(\Gamma) = \mathbb{P}_\mu\left[X_t \in \Gamma \mid X_t \in \mathscr{S}_+\right] \text{ for all } t,$$

where $\mathbb{P}_\mu$ denotes the law of the Markov chain $\{X_t\}_{t=0}^\infty$, conditional to X_0 being distributed according to μ.

In the case that the Markov chain has a finite number of states and P is the transition matrix (i.e. $P_{ij} = p(i, \{j\})$), [15] showed that the QSD is given by $\mu(\{i\}) = \pi_i$ where π is the normalized, dominant left eigenvector of the matrix Q given by removing the rows and columns of P corresponding to extinction states in $\mathscr{S}_0$. In this case, λ is the corresponding eigenvalue of this eigenvector. For the Poisson Lotka-Volterra processes in which the unperturbed dynamic F is pre-compact, Proposition 6.1 from [20] implies the existence of QSDs for these processes. Examples of these QSDs for these processes are shown in Figs. 2, 3, and 4. More generally, the existence of QSDs has been studied extensively by many authors as reviewed by [39].

What do these QSD's and λ tell us about the behavior of the stochastic process? From the perspective of metastability, QSDs often exhibit the following property:

$$\mu(\Gamma) = \lim_{t\to+\infty} \mathbb{P}\,[X_t \in \Gamma \mid X_t \in \mathscr{S}_+, X_0 = x]$$

where the limit exists and is independent of the initial state $x \in \mathscr{S}_+$. In words, the QSD describes the probability distribution of X_t, conditioned on non-extinction, far into the future. Hence, the QSD provides a statistical description of the meta-stable behavior of the process. The eigenvalue, λ provides information about the length of the metastable behavior of X_t. Specifically, given that the process is following the QSD (e.g. X_0 is distributed like μ), and λ equals the probability of persisting in the next time step. Thus, the mean time to extinction is $\frac{1}{1-\lambda}$. [22] call $\frac{1}{1-\lambda}$, the "intrinsic mean time to extinction" and, convincingly, argue that it is a fundamental statistic for comparing extinction risk across stochastic models.

2.2 *Positive Attractors, Intrinsic Extinction Risk, and Metastability*

When the habitat size is sufficiently large i.e. ε is small, there is a strong relationship between the existence of attractors in $\mathscr{S}_+$ (i.e. "positive" attractors) for the unperturbed system F and the quasi-stationary distributions of X^ε. This relationship simultaneously provides information about the metastable behavior of the stochastic model and intrinsic probability of extinction, $1 - \lambda_\varepsilon$. To make this relationship mathematically rigorous, we need to strengthen Hypotheses 2.1 and 2.2. [20] presents two ways to strengthen these hypothesis. We focus on their large deviation approach as it is most easily verified. This approach requires identifying a *rate function* $\rho : \mathscr{S} \times \mathscr{S} \to [0, \infty]$ that describes the probability of a large deviation between F and X^ε. That is, for a sufficiently small neighborhood U of a point y, the rate function should have the property

$$\mathbb{P}[X^\varepsilon_{t+1} \in U | X^\varepsilon_t = x] \approx \exp(-\rho(x, y)/\varepsilon).$$

Hypothesis 2.3 provides the precise definition and desired properties of ρ.

Hypothesis 2.3 *There exists a* rate function $\rho : \mathscr{S} \times \mathscr{S} \to [0, +\infty]$ *such that*

(i) ρ *is continuous on* $\mathscr{S}_+ \times \mathscr{S}$*,*

(ii) $\rho(x, y) = 0$ *if and only if* $y = F(x)$,
(iii) *for any* $\beta > 0$,

$$\inf\{\rho(x, y) : \ x \in \mathscr{S},\ y \in \mathscr{S},\ \|F(x) - y\| > \beta\} > 0, \tag{7}$$

(iv) *for any open set U, there is the lower bound*

$$\liminf_{\varepsilon\to 0} \varepsilon \log p^{\varepsilon}(x, U) \geq - \inf_{y\in U} \rho(x, y) \tag{8}$$

that holds uniformly for x in compact subsets of $\mathscr{S}_+$ *whenever U is an open ball in* $\mathscr{S}$*. Additionally, for any closed set C, there is the uniform upper bound*

$$\limsup_{\varepsilon\to 0} \sup_{x\in\mathscr{S}} \varepsilon \log p^{\varepsilon}(x, C) \leq - \inf_{y\in C} \rho(x, y). \tag{9}$$

Equations (7) and (9), in particular, imply that Hypothesis 2.1 holds. Furthermore, as $\mathscr{S}_0$ is absorbing, equation (8) implies that $\rho(x, y) = +\infty$ for all $x \in \mathscr{S}_0, y \in \mathscr{S}_+$. Identifying the rate function ρ typically requires making use of the Gärtner-Ellis theorem [16, Theorem 2.3.6] which provides large deviation estimates for sums of independent random variables. Example 3 below describes how this theorem was used for the Poisson Lotka-Volterra processes.

We strengthen Hypothesis 2.2 as follows:

Hypothesis 2.4 *For any* $c > 0$*, there exists an open neighborhood* V_0 *of* $\mathscr{S}_0$ *such that*

$$\lim_{\varepsilon\to 0} \inf_{x\in V_0} \varepsilon \log p^{\varepsilon}(x, \mathscr{S}_0) \geq -c. \tag{10}$$

Equation (10) implies that

$$\mathbb{P}[X^{\varepsilon}_{t+1} \in \mathscr{S}_0 | X_t \in V_0] \geq \exp(-c/\varepsilon)$$

for $\varepsilon > 0$ sufficiently small. Namely, the probability of absorption near the boundary, at most, decays exponentially with habitat size. The following example discusses why these stronger hypotheses hold for the Poisson Lotka-Volterra process.

Example 3 (Return of the Poisson Lotka-Volterra Process) Using the Gärtner-Ellis theorem [16, Theorem 2.3.6], Faure and Schreiber [20, Proposition 6.4] showed that $\rho(x, y) = \sum_i y_i \log \frac{y_i}{F_i(x)} - y_i$ is the rate function for any Poisson processes with mean $F : \mathbb{R}^k_+ \to \mathbb{R}^k_+$ including the Poisson Lotka-Volterra Process of Example 1. To see why Hypothesis 4 holds for the Poisson Lotka-Volterra process, assume x is such that $x_i \leq \delta$ for some $\delta > 0$ and i. Then

$$\begin{aligned}\varepsilon \log \mathbb{P}[X^{\varepsilon}_{t+1} \in \mathscr{S}_0 | X_t = x] &\geq \varepsilon \log \mathbb{P}[X^{\varepsilon}_{i,t+1} = 0 | X_t = x] \\ &= \varepsilon \log \exp(-F_i(x)/\varepsilon) = -F_i(x)\end{aligned}$$

Hence, for any $c > 0$, choose $\delta > 0$ sufficiently small to ensure that for all i, $F_i(x) \leq c$ whenever $x_i \leq \delta$. In which case, choosing $V_0 = \{x \in \mathbb{R}^k_+ : x_i \leq \delta$ for some $i\}$ satisfies (10).

As many discrete distributions are used in models with demographic stochasticity (e.g. negative binomial, mixtures of Bernoullis and negative binomials), an important open problem is the following:

Problem 1 For which types of random perturbations of an ecological model F do Hypotheses 3 and 4 hold?

To relate QSDs to the attractors of the deterministic dynamics, we recall the definition of an attractor and weak* convergence of probability measures. A compact set $A \subset \mathscr{S}$ is an attractor for F if there exists a neighborhood U of A such that (i) $\cap_{n\geq 1}F^n(U) = A$ and (ii) for any open set V containing A, $F^n(U) \subset V$ for some $n \geq 1$. A weak* limit point of a family of probability measures $\{\mu_\varepsilon\}_{\varepsilon>0}$ on $\mathscr{S}$ is a probability measure μ^0 such that there exists a sequence $\varepsilon_n \downarrow 0$ satisfying

$$\lim_{n\to\infty} \int h(x)\mu^{\varepsilon_n}(dx) = \int h(x)\mu^0(dx)$$

for all continuous functions $h : \mathscr{S} \to \mathbb{R}$. Namely, the expectation of any continuous function with respect to μ^{ε_n} converges to its expectation with respect to μ^0 as $n \to \infty$. The following theorem follows from [20, Lemma 3.9 and Theorem 3.12].

Theorem 2.5 *Assume Hypotheses 2.3 and 2.4 hold. Assume for each $\varepsilon > 0$, there exists a QSD μ_ε for X^ε. If there exists a positive attractor $A \subset \mathscr{S}_+$, then*

- *there exists a neighborhood V_0 of $\mathscr{S}_0$ such that all weak* limit points μ^0 of $\{\mu_\varepsilon\}_{\varepsilon>0}$ are F-invariant and $\mu^0(V_0) = 0$, and*
- *there exists $c > 0$ such that*

$$\lambda_\varepsilon \geq 1 - e^{-c/\varepsilon} \textit{ for all } \varepsilon > 0. \tag{11}$$

Alternatively, assume that $\mathscr{S}_0$ is a global attractor for the dynamics of F. Then any weak-limit point of $\{\mu_\varepsilon\}_{\varepsilon>0}$ is supported by $\mathscr{S}_0$.*

Theorem 2.5 implies the existence of a positive attractor of the deterministic dynamics ensures the stochastic process exhibits metastable behavior for large habitat size, and the probability of extinction $1 - \lambda_\varepsilon$ decreases exponentially with habitat size. Equivalently, the mean time to extinction $1/(1 - \lambda_\varepsilon)$ increases exponential with habitat size. These conclusions are illustrated in Fig. 2 with a one-dimensional Poisson Lotka-Volterra process (the Poisson Ricker process described below in Example 4).

Even if F has no positive attractors, $\mathscr{S}_0$ may not be a global attractor as there might be an unstable invariant set in $\mathscr{S}_+$. For example, single species models with positive feedbacks can have an uncountable number of unstable periodic orbits despite almost every initial condition going to extinction [46]. Hence, the necessary

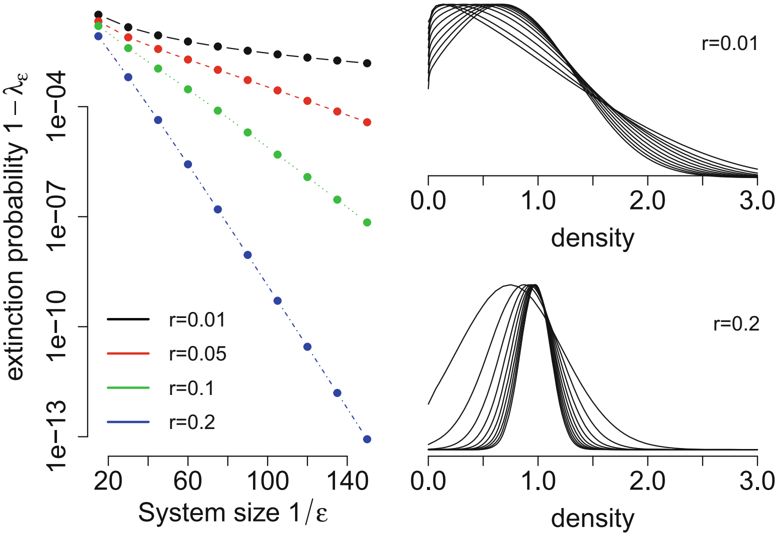


Fig. 2 Extinction probabilities and QSDs for the Poisson Ricker process described in Example 4. In the left panel, the "intrinsic" extinction probability $1 - \lambda_\varepsilon$ plotted as function of the habitat size $1/\varepsilon$ and for different r values. In the right panels, the QSDs plotted for a range of habitat sizes and two r values

and sufficient conditions for metastability in Theorem 2.5 are not equivalent. However, if F has no positive attractors, one can show that all points in $\mathscr{S}_+$ can with arbitrarily small perturbations be "forced" to $\mathscr{S}_0$ [47, 48]. Hence, this raises the following open problem.

Problem 2 If F has no positive attractors, are all the weak*-limit points of the QSDs supported by the extinction set $\mathscr{S}_0$?

While the methodology used to prove Theorem 2.5 provides an explicit expression for $c > 0$, this expression is fairly abstract and only provides a fairly crude lower bound. This suggests the following questions which, if solved, may provide insights into how extinction probabilities depend on the nature of the nonlinear feedbacks within and between populations and the form of demographic stochasticity.

Problem 3 If F has positive attractors, when does the limit

$$\lim_{\varepsilon\to 0} -\frac{1}{\varepsilon}\log(1-\lambda_\varepsilon) =: c$$

exist? If the limit exists, under what circumstances can we derive explicit expressions for c? or good explicit lower bounds for c?

Theorem 2.5 only ensures that the metastable dynamics concentrates on an invariant set for the deterministic dynamics. However, it is natural to conjecture

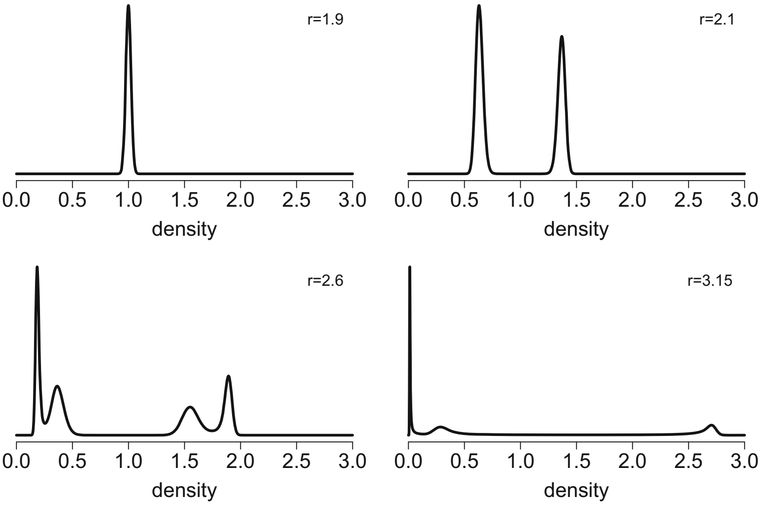


Fig. 3 QSDs for the stochastic Ricker model (see Example 4) for r values where $F(x) = x\exp(r(1-x))$ has a stable periodic orbit. Habitat size $1/\varepsilon$ is $2,500$

that the QSDs μ_ε should concentrate on the positive attractors of F. These positive attractors, however, may coexist with complex unstable behavior. For example, the Ricker equation $F(x) = x\exp(r(1-x))$ can have a stable periodic orbit coexisting with an infinite number of unstable periodic orbits (e.g. the case of a stable period 3 orbit as illustrated in Fig. 3).

To identify when this intuition is correct, a few definitions from dynamical systems are required. For $x \in \mathscr{S}$, let $\omega(x) = \{y$: there exists $n_k \to \infty$ such that $\lim_{k\to\infty} F^{n_k}(x) = y\}$ be the *ω-limit set for x* and $\alpha(x) = \{y$: there exist $n_k \to \infty$ and $y_k \in \mathscr{S}$ such that $F^{n_k}(y_k) = x$ and $\lim_{k\to\infty} y_k = y\}$ be the *α-limit set for x*. Our assumption that F is precompact implies that there exists a global attractor given by the compact, F-invariant set $\Lambda = \cap_{n\geq 0} F^n(\mathscr{S})$. For all $x \in \Lambda$, $\omega(x)$ and $\alpha(x)$ are compact, non-empty, F-invariant sets. A *Morse decomposition* of the dynamics of F is a collection of F-invariant, compact sets $K_1, \dots, K_\ell$ such that

- K_i is isolated i.e. there exists a neighborhood of K_i such that it is the maximal F-invariant set in the neighborhood, and
- for every $x \in \Lambda \setminus \cup_{i=1}^{\ell} K_i$, there exist $j < i$ such that $\alpha(x) \subset K_j$ and $\omega(x) \subset K_i$.

Replacing the invariant sets K_i by points, one can think of F being gradient-like as all orbits move from lower indexed invariant sets to higher indexed invariant sets. Finally, recall that a compact invariant set K is *transitive* if there exists an $x \in K$ such that $\{x, F(x), F^2(x), \dots\}$ is dense in K. Faure and Schreiber [20, Theorem 2.7, Remark 2.8, and Proposition 5.1] proved the following result about QSDs not concentrating on the non-attractors of F. The assumptions of this theorem can be verified for many ecological models.

Theorem 2.6 *Assume Hypotheses 2.3 and 2.4 hold. Let $K_1, \dots, K_\ell$ be a Morse decomposition for F such $K_j, \dots, K_\ell$ are attractors. If*

- *$K_i \subset \mathscr{S}_+$ or $K_i \subset \mathscr{S}_0$ for each i,*
- *$K_i \subset \mathscr{S}_+$ for some $i \geq j$, and*
- *K_i with $i \leq j-1$ is transitive whenever $K_i \subset \mathscr{S}_+$,*

then any weak-limit point of $\{\mu_\varepsilon\}_{\varepsilon>0}$ is F-invariant and is supported by the union of attractors in $\mathscr{S}_+$.*

For random perturbations of deterministic models without absorbing states (e.g. models accounting for immigration or mutations between genotypes), the work of [31] and [21] can be used to show that the stationary distributions often concentrate on a unique attractor. However, due to the singularity of the rate function ρ along the extinction set $\mathscr{S}_0$, the approach used by these authors doesn't readily extend to the stochastic models considered here. This raises the following open problem:

Problem 4 If F has multiple, positive attractors, under what conditions do the QSDs μ_ε concentrate on a unique one of these positive attractors as $\varepsilon \downarrow 0$?

Lets apply some of these results to the Poisson Lotka-Volterra processes from Example 1.

Example 4 (The Ricker Model) The simplest of Poisson Lotka-Volterra processes is the stochastic Ricker model for a single species where $F(x) = x\exp(r(1-x))$ with $r > 0$. [33] proved that for an open and dense set of $r > 0$ values, the Ricker map has a Morse decomposition consisting of a finite number of unstable, intransitive sets (more specifically, hyperbolic sets) and a unique stable period orbit $\{p, F(p), \dots, F^T(p)\}$. As the stable periodic orbit is the only attractor, Theorem 2.6 implies the following result.

Corollary 1 *Consider the Ricker process with $r > 0$ such that $F(x) = x\exp(r(1-x))$ has the aforementioned Morse decomposition. Then any weak*-limit point of $\{\mu_\varepsilon\}_{\varepsilon>0}$ is supported by the unique stable periodic orbit $\{p, F(p), \dots, F^T(p)\}$.*

Figure 3 illustrates this corollary: QSDs concentrating on the stable periodic orbit of period 1 for $r = 1.9$, period 2 for $r = 2.1$, period 4 for $r = 2.6$, and period 3 for $r = 3.15$. Remarkably, in the case of the stable orbit of period 3, there exists an infinite number of unstable periodic orbits which the QSDs do not concentrate on. We note that [27, 32, 42] proved similar results to Corollary 1 using inherently one dimensional methods.

Example 5 (Revenge of the Poisson Lotka-Volterra Processes) For higher dimensional Lotka-Volterra processes, we can use properties of Lotka-Volterra difference equations in conjunctions with Theorems 2.5 and 2.6 to derive two algebraically verifiable results for the stochastic models. First, if the deterministic map $F = (F_1, \dots, F_k)$ with $F_i(x) = x_i \exp(\sum_j A_{ij}x_j + r_i)$ is pre-compact and there is no internal fixed point (i.e. there is no strictly positive solution to $Ax = -r$), then [26] proved that the boundary of the positive orthant is a global attractor. Hence, Theorem 2.5 implies the following corollary.

Corollary 2 *Let X^{ε} be a Poisson Lotka-Volterra process such that F is pre-compact and admits no positive fixed point. Then any weak*-limit point of $\{\mu_{\varepsilon}\}_{\varepsilon>0}$ is supported by $\mathscr{S}_0$, the boundary of the positive orthant of $\mathbb{R}^k_+$.*

On the other hand, [26] derived a simple algebraic condition which ensures that the deterministic dynamics of F has a positive attractor. Namely, there exist $p_i > 0$ such that

$$\sum_i p_i \left(\sum_j A_{ij} x_j^* + r_i \right) > 0 \tag{12}$$

for any fixed point x^* on the boundary of the positive orthant. Hence, Theorem 2.5 implies the following corollary.

Corollary 3 *Assume $F = (F_1, \dots, F_k)$ with $F_i(x) = x_i \exp(\sum_j A_{ij} x_j + r_i)$ is pre-compact and satisfies (12) for some choice of $p_i > 0$. If X^{ε} is the corresponding Poisson Lotka-Volterra process, then any weak*-limit point of $\{\mu_{\varepsilon}\}_{\varepsilon>0}$ is supported by A where $A \subset \mathscr{S}_+$ is the global, positive attractor for F. Moreover, there exists $c > 0$ such that $\lambda_{\varepsilon} \geq 1 - \exp(c/\varepsilon)$ for all $\varepsilon > 0$ sufficiently small.*

Figure 4 illustrates the convergence of the QSDs to the attractor of F for a Lotka-Volterra process of two competing species. Even for populations of only hundreds of individuals ($\varepsilon = 0.01$), this figure illustrates that species can coexist for tens of thousands of generations despite oscillating between low and high densities, a key signature of the underlying deterministic dynamics. However, only at much larger habitat sizes (e.g. $1/\varepsilon = 1,000,000$) do the metastable behaviors clearly articulate the underlying deterministic complexities.

3 Environmental Stochasticity

To understand how environmental fluctuations, in and of themselves, influence population dynamics, we shift our attention to models for which the habitat size is sufficiently large that one can approximate the population state by a continuous variable. Specifically, let $X_t \in \mathbb{R}^k_+$ denote the state of the population or community at time t. The components of $X_t = (X_{1,t}, X_{2,t}, \dots, X_{k,t})$ corresponds to densities or frequencies of subpopulations. To account for environmental fluctuations, let $\mathscr{E} \subset \mathbb{R}^m$ (for some m) be a compact set representing all possible environmental states e.g. all possible precipitation and temperature values. I assume that $E_{t+1} \in \mathscr{E}$ represents the environmental state of the system over the time interval $(t, t+1]$ that determines how the community state changes over that time interval. If the population or community state X_{t+1} depends continuously on E_{t+1} and X_t, then

$$X_{t+1} = F(X_t, E_{t+1}) \tag{13}$$

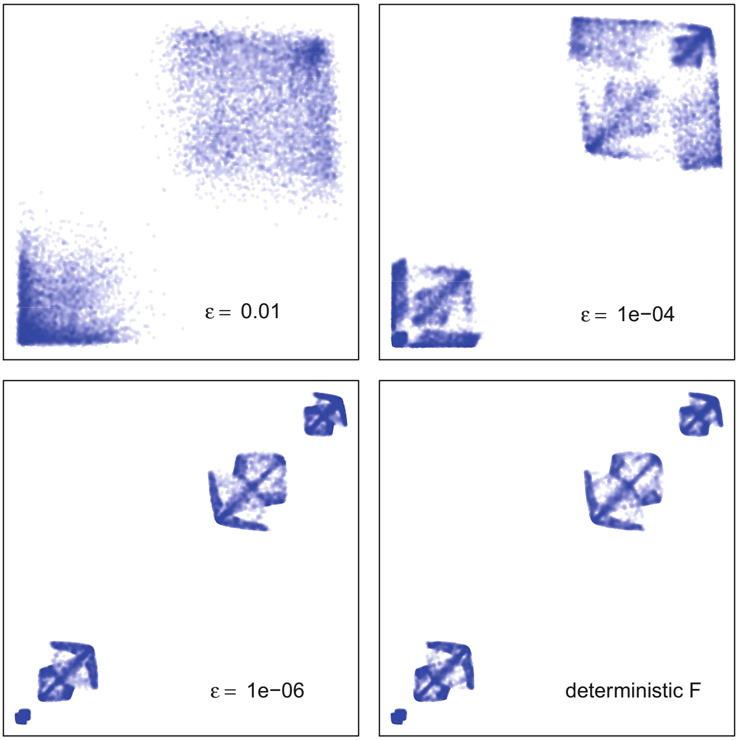


Fig. 4 Numerically estimated QSDs for a Poisson Lotka-Volterra process with two competing species, and the global attractor of the deterministic map $F_i(x) = x_i \exp(\sum_j A_{ij} x_j x + r_i)$. The stochastic and deterministic processes were simulated for 50, 000 time steps and the last 17,500 time steps are plotted in the x_1–x_2 plane. Parameters: A is the matrix with rows $(-0.2, -0.01)$, $(-0.01, -0.2)$ and $r = (2.71, 2.71)$

for a continuous map $F : \mathbb{R}^k_+ \times \mathscr{E} \to \mathbb{R}^k_+$. If the E_t are random variables, then (13) is known as a *continuous, random dynamical system*. [2] provides a thorough overview of the general theory of these random dynamical systems.

To state the main hypotheses about (13), recall that a sequence of random variables, $E_1, E_2, \ldots$, is *stationary* if for every pair of non-negative integers t and s, $E_1, \ldots, E_t$ and $E_{1+s}, \ldots, E_{t+s}$ have the same distribution. The sequence is *ergodic* if with probability one all realizations of the sequence have the same asymptotic statistical properties e.g. time averages (see, e.g., [17] for a more precise definition).

Hypothesis 3.7 *$E_1, E_2, \ldots$ are an ergodic and stationary sequence of random variables taking value in $\mathscr{E}$. Let π be the stationary distribution of this sequence i.e. the probability measure π on $\mathscr{E}$ such that $\mathbb{P}[E_t \in B] = \pi(B)$ for all Borel sets $B \subset \mathscr{E}$.*

This hypothesis is satisfied for a diversity of models of environmental dynamics. For example, E_t could be given by a finite state Markov chain on a finite number of environmental states, say $e_1, e_2, \ldots, e_m \in E$ (e.g. wet and cool, wet and hot, dry and cool, dry and hot) with transition probabilities $p_{ij} = \mathbb{P}[E_{t+1} = e_j | E_t = e_i]$. If the transition matrix $P = (p_{ij})_{i,j}$ is aperiodic and irreducible, then E_t is asymptotically ergodic and stationary. Alternatively, E_t could be given by a sequence of independent and identically distributed random variables or, more generally, an autoregressive process.

Our second hypothesis simply assumes that population densities remain bounded and allows for the possibility of extinction.

Hypothesis 3.8 *There are compact sets* $\mathscr{S} \subset \mathbb{R}^k_+$ *and* $\mathscr{S}_0 \subset \{x \in \mathscr{S} : \prod_i x_i = 0\}$ *such that* $F : \mathscr{S} \times \mathscr{E} \to \mathscr{S}$, $F : \mathscr{S}_0 \times \mathscr{E} \to \mathscr{S}_0$, *and* $F : \mathscr{S}_+ \times \mathscr{E} \to \mathscr{S}_+$ *where* $\mathscr{S}_+ = \mathscr{S} \setminus \mathscr{S}_0$.

For example, $\mathscr{S}$ may equal $[0, M]^k$ where M is the maximal density of a species or $\mathscr{S}$ may be the probability simplex $\Delta = \{x \in \mathbb{R}^k_+ : \sum_i x_i = 1\}$ where $x \in \mathscr{S}$ corresponds to the vector of genotypic frequencies. As in the case of demographic stochasticity, $\mathscr{S}_0$ corresponds to the set where one or more populations have gone extinct. Invariance of $\mathscr{S}_0$ implies that once the population has gone extinct it remains extinct i.e. the "no cats, no kittens" principle. Invariance of $\mathscr{S}_+$ implies that populations can not go extinct in one time step but only asymptotically. This latter assumption is met by most (but not all) models in the population biology literature.

For these stochastic difference equations, there are several concepts of "persistence" which are reviewed in [49]. Here, we focus on the "typical trajectory" perspective. Namely, "how frequently does the typical population trajectory visit a particular configuration of the population state space far into the future?" The answer to this question is characterized by *empirical measures* for X_t:

$$\Pi^x_t(A) = \frac{\#\{0 \leq s \leq t : X_s \in A\}}{t+1}$$

where $X_0 = x$ and A is a Borel subset of $\mathscr{S}$. $\Pi_t(A)$ equals the fraction of time that X_s spends in the set A over the time interval $[0, t]$. Provided the limit exists, the long-term frequency that X_t enters A is given by $\lim_{t\to\infty} \Pi^x_t(A)$. It is important to note that these empirical measures are random measures as they depend on the particular realization of the stochastic process. Figure 5 provides graphical illustrations of empirical measures for a single species model (top row) and a two species model (bottom row). For both models, the empirical measure at time t can be approximated by a histogram describing the frequency X_t spends in different parts (e.g. intervals or hexagons) of the population state space $\mathscr{S}$.

Stochastic persistence corresponds to the typical trajectory spending arbitrarily little time, arbitrarily near the extinction set $\mathscr{S}_0$. More precisely, for all $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$\limsup_{t\to\infty} \Pi^x_t\left(\{x \in \mathscr{S} : \mathrm{dist}(x, \mathscr{S}_0) \leq \delta\}\right) \leq \varepsilon \text{ with probability one for all } x \in \mathscr{S}_+$$

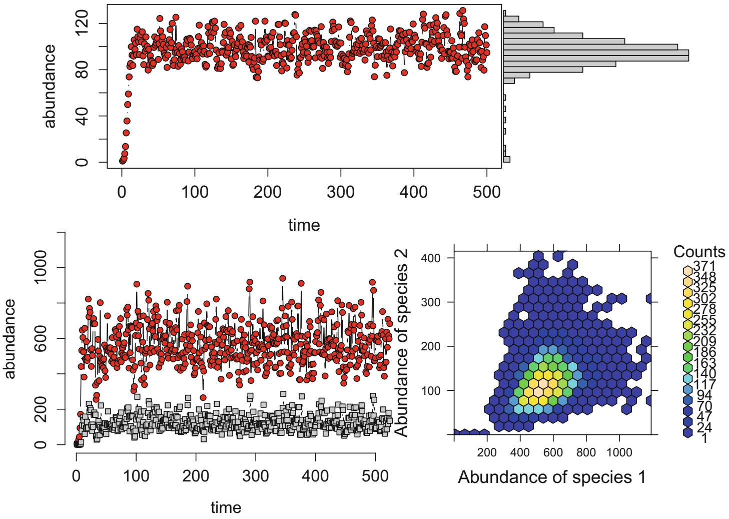


Fig. 5 Visualizing the empirical measures Π_t^x for two models with environmental stochasticity. In top row, the time series of a realization of a stochastic, single species model $X_{t+1} = \frac{\eta_{t+1} X_t}{1+0.01 X_t}$ where η_t is a truncated log normal with log-mean $\log 2$ and log-variance 0.01. Histogram to the right of the time series corresponds to $\Pi_{500}^x([a,b])$ for intervals $[a,b]$ of width 10 from 0 to 140. In the bottom row, the time series of a realization of a stochastic predator-prey model $X_{1,t+1} = X_{1,t} \exp(\eta_{t+1} - 0.001 X_{1,t} - 0.001 X_{2,t}), X_{2,t+1} = 0.5 X_{1,t}(1-\exp(-0.001 X_{2,t}))$ here η_t is a truncated log normal with log-mean $\log 2$ and log-variance 0.04. To the right of the time series, the time spent in each colored hexagon in $\mathbb{R}_+^2$ is shown. $\Pi_{500}^x(H)$ for one of the hexagons $H \subset \mathbb{R}_+^2$ equals the count divided by 500. The truncated normals are used for these models to ensure that dynamics remain in a compact set $\mathscr{S}$

where $\operatorname{dist}(x, S_0) = \min_{y \in \mathscr{S}_0} \|x - y\|$. In contrast to the deterministic notions of uniform persistence or permanence, stochastic persistence allows for trajectories to get arbitrarily close to extinction and only requires the frequency of these events are very small. One could insist that the trajectories never get close to extinction. However, such a definition is too strict for any model where there is a positive probability of years where the population is tending to decline e.g. the models discussed in Sect. 7. Regarding this point, [11] wrote

> “This criterion…places restrictions on the expected frequency of fluctuations to low population levels. Given that fluctuations in the environment will continually perturb population densities, it is to be expected that any nominated population density, no matter how small, will eventually be seen. Indeed this is the usual case in stochastic population models and is not an unreasonable postulate about the real world. Thus a reasonable persistence criterion cannot hope to do better than place restrictions on the frequencies with which such events occur.”

Conditions for verifying stochastic persistence appear in papers by [5, 44, 49, 50]. As the results by [44] are the most general, we focus on them. We begin with single species models and then expand to multi-species models.

3.1 Single Species Models

Consider a single species for which an individual can be in one of k states. For example, these states may correspond to age where k is the maximal age, living in one of k spatial locations or "patches", discrete behavioral states that an individual can move between, different genotypes in an asexual population coupled by mutation, or finite number of developmental stages or size classes. $X_{i,t}$ corresponds to population density of individuals in state i and $X_t = (X_{1,t}, \dots, X_{k,t})$ is the population state. The population state is updated by multiplication by a $k \times k$ matrix $A(X_t, E_{t+1})$ dependent on the population and environmental state:

$$X_{t+1} = A(X_t, E_{t+1})X_t =: F(X_t, E_{t+1}). \tag{14}$$

Assume $A(X, E)$ satisfies the following hypothesis.

Hypothesis 3.9 *A is a continuous mapping from $\mathscr{S} \times \mathscr{E}$ to non-negative $k \times k$ matrices. Furthermore, there exists a non-negative, primitive matrix B such that $A(x, E)$ has the same sign structure as B for all x, E i.e. the i–j-th entry of $A(x, E)$ is positive if and only if the i–j-th entry of B is positive.*

The primitivity assumption implies that there is a time, T, such that after T time steps, individuals in every state contribute to individuals in all other states. Specifically, $A(X_{T-1}, E_T)A(X_{T-2}, E_{T-1}) \dots A(X_0, E_1)$ has only positive entries for any $X_0, \dots, X_{T-1} \in \mathscr{S}$ and $E_1, \dots, E_T \in \mathscr{E}$. This assumption is met for most models.

To determine whether or not the population has a tendency to increase or decrease when rare, we can approximate the dynamics of (14) when $X_0 \approx 0$ by the linearized system

$$Z_{t+1} = B_{t+1}Z_t \text{ where } Z_0 = X_0 \text{ and } B_{t+1} = A(0, E_{t+1}). \tag{15}$$

Iterating this matrix equation gives

$$Z_t = B_t B_{t-1} B_{t-2} \dots B_2 B_1 Z_0.$$

Proposition 3.2 from [45] and Birkhoff's ergodic theorem implies there is a quantity r, the dominant Lyapunov exponent, such that

$$\lim_{t\to\infty} \frac{1}{t} \log \|Z_t\| = r \text{ with probability one}$$

whenever $Z_0 \in \mathbb{R}^k_+ \setminus \{0\}$. Following [7–9], we call r the *low-density per-capita growth* of the population. When $r > 0$, Z_t with probability one grows exponentially and we would expect the population state X_t to increase when rare. Conversely when $r < 0$, Z_t with probability one converges to 0. Consistent with these predictions from the linear approximation, Roth and Schreiber [44, Theorems 3.1,5.1] proved the following result.

Theorem 3.10 *Assume Hypotheses 3.7 through 3.9 hold with $\mathscr{S}_0 = \{0\}$. If $r > 0$, then (14) is stochastically persistent. If $r < 0$ and $A(0, E) \geq A(X, E)$ for all X, E, then*

$$\lim_{t\to\infty} X_t = 0 \textit{ with probability one.}$$

The assumption in the partial converse is a weak form of negative-density dependence as it requires that the best conditions (in terms of magnitude of the entries of A) occurs at low densities. There are cases where this might not be true e.g. models accounting for positive density-dependence, size structured models where growth to the next stage is maximal at low densities.

Example 6 (The Case of the Bay Checkerspot Butterflies) The simplest case for which Theorem 3.10 applies are unstructured models where $k = 1$. In this case, $B_t = A(0, E_t)$ are scalars and

$$r = \mathbb{E}[\log B_t].$$

The exponential e^r corresponds to the geometric mean of the B_t. By Jensen's inequality, the arithmetic mean $\mathbb{E}[B_t]$ is greater than or equal to this geometric mean e^r, with equality only if B_t is constant with probability one. Hence, environmental fluctuations in the low-density fitnesses B_t reduce r and have a detrimental effect on population persistence.

To illustrate this fundamental demographic principle, we visit a study by [38] on the dynamics of Bay checkerspot butterflies, a critically endangered species. In the 1990s, two populations of this species went extinct in Northern California. The population densities for one of these populations is shown in the left hand side of Fig. 6. Both extinctions were observed to coincide with a change in precipitation variability in the 1970s (right hand side of Fig. 6): the standard deviation in precipitation is approximately 50% higher after 1971 than before 1971.

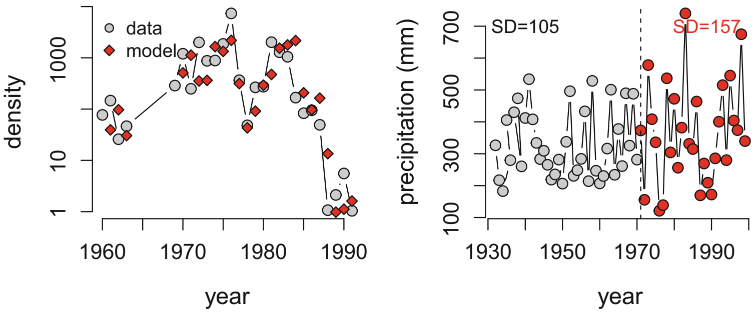


Fig. 6 Checkerspot population dynamics (left) and precipitation (right) from Example 6. Model fit for population dynamics as red diamonds

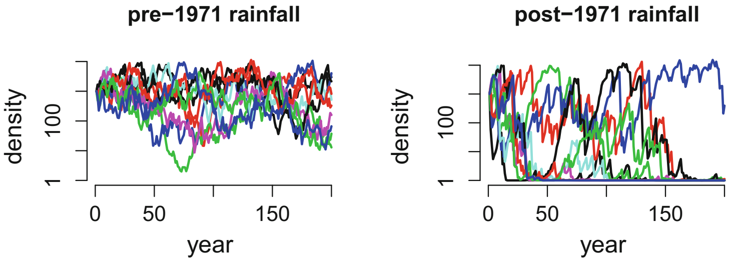


Fig. 7 Simulated checkerspot population dynamics with pre-1971 precipitation data (left) and post-1971 precipitation data (right) from Example 6

To evaluate whether this shift in precipitation variability may have caused the extinction of the checkerspots, [38] developed a stochastic difference equation of the following type

$$n_{t+1} = n_t \exp(a - bn_t + cE_{t+1}^{-2})$$

where E_t is precipitation in year t. Using linear regression on a log-scale yields a model whose fit for one-year predictions are shown as red diamonds in Fig. 6. To compare the pre-1971 and post-1971 population dynamics of the populations, [38] ran their stochastic difference equations with E_t given by independent draws from the corresponding years of precipitation data. The resulting models satisfy all of the assumptions of Theorem 3.10. The model with random draws from the pre-1971 precipitation data yields $r = \mathbb{E}[a + cE_1^{-2}] = 0.04$. Hence, Theorem 3.10 implies stochastic persistence with this form of climatic variability (left hand side of Fig. 7). In contrast, the model with random draws from the post-1971 precipitation data yields $r = -0.049$. Hence, Theorem 3.10 implies the population is extinction bound with this form of climatic variability (right hand side of Fig. 7).

Example 7 (Spatially Structured Populations) To illustrate the application of Theorem 3.10 to structured populations, consider a population in which individuals can live in one of k patches (e.g. butterflies dispersing between heath meadows, pike swimming between the northern and southern basin of a lake, acorn woodpeckers flying between canyons). $X_{i,t}$ is the population density in patch i. Let $E_{t+1} = (E_{1,t+1}, \ldots, E_{k,t+1})$ be the environmental state over $(t, t+1]$ where $E_{i,t}$ be the low-density fitness of individuals in patch i. To account for within-patch competition, let $f_i(X_i, E_i) = E_i/(1 + c_iX_i)$ be the fitness of an individual in patch i where c_i measures the strength of competition within patch i. This fitness function corresponds to the Beverton-Holt model in population biology.

To couple the dynamics of the patches, let d be the fraction of dispersing individuals that go with equal likelihood to any other patch. In the words of Ulysses Everett McGill in *O Brother, Where Art Thou?*

> "Well ain't [these patches] a geographical oddity! Two weeks from everywhere!"

Despite this odd geographic regularity, these all-to-all coupling models have proven valuable to understanding spatial population dynamics. Under these assumptions, we get a spatially structured model of the form

$$X_{i,t+1} = (1-d)f_i(X_{i,t}, E_{i,t+1})X_{i,t} + \frac{d}{k-1}\sum_{j\neq i} f_j(X_{j,t}, E_{j,t+1})X_{j,t}. \tag{16}$$

For this model, $A(X, E)$ is the matrix whose i–j-th entry equals $\frac{d}{k-1}f_j(X_{j,t}, E_{j,t+1})$ for $j \neq i$ and $(1-d)f_i(X_{i,t}, E_{i,t+1})$ for $j = i$.

The low density per-capita growth rate r is the dominant Lyapunov exponent of the random product of the matrices $B_t = A(0, E_t)$. Theorem 3.10 implies this model exhibits stochastic persistence if $r > 0$ and asymptotic extinction with probability one if $r < 0$. In fact, as this spatial model has some special properties (monotonicity and sublinearity), work of Benaïm and Schreiber [5, Theorem 1] implies if $r > 0$, then there is a probability measure m on $\mathscr{S}_+$ such that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} h(X_t) = \int h(x)\, m(dx) \text{ with probability one}$$

for any $x \in \mathscr{S}_+$ and any continuous function $h : \mathscr{S} \to \mathbb{R}$. Namely, for all positive initial conditions, the long-term behavior is statistically characterized by the probability measure m that places no weight on the extinction set. When this occurs, running the model once for sufficiently long describes the long-term statistical behavior for all runs with probability one. The probability measure m corresponds to the marginal of an invariant measure for the stochastic model.

But when is $r > 0$? Finding explicit, tractable formulas for r, in general, appears impossible. However, for sedentary populations ($d \approx 0$) and perfectly mixing populations ($d = 1 - 1/k$), one has explicit expressions for r. In the limit of $d = 0$,

$$r = \max_i \mathbb{E}[\log E_{i,t}]$$

as $f_i(0, E_i) = E_i$. As r varies continuously with d (cf. Benaïm and Schreiber [5, Proposition 3]), it follows that persistence for small d (i.e. mostly sedentary populations) only occurs if $\mathbb{E}[\log E_{i,t}] > 0$. Equivalently, the geometric mean $\exp(\mathbb{E}[\log E_{i,t}])$ of the low-density fitnesses $E_{i,t}$ is greater than one in at least one patch.

When $d = 1 - 1/k$, the fraction of individuals going from any one patch to any other patch is $1/k$. In this case, the model reduces to a scalar model for which

$$r = \mathbb{E}\left[\log\left(\frac{1}{k}\sum_{i=1}^{k} E_{i,t}\right)\right].$$

Namely, e^r is equal to the geometric mean of the spatial means [40]. Applying Jensen's inequality to the outer and inner expressions of r, one gets

$$\log\left(\frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[E_{i,t}]\right) > r > \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}[\log E_{i,t}].$$

Hence, persistence requires that the expected fitness in one patch is greater than one (i.e. $\mathbb{E}[E_{i,t}] > 1$ for some i in the left hand side), but can occur even if all the patches are unable to sustain the population (i.e. $\mathbb{E}[\log E_{i,t}] < 0$ for all i on the right hand side). Hence, local populations which are tending toward extinction (i.e. $\mathbb{E}[\log E_{i,t}] < 0$ in all patches) can persist if they are coupled by dispersal. Even more surprising, [51] shows that stochastic persistence is possible in temporally autocorrelated environments even if $\mathbb{E}[E_{i,t}] < 1$ for all patches.

To better understand how r depends on d, I make raise the following problem which has been proven have an affirmative answer for two-patch stochastic differential equation models by [18].

Problem 5 If $E_{i,t}$ are independent and identically distributed in time and space, then is r an increasing function of d on the interval $(0, 1 - 1/k)$? In particular, if $\mathbb{E}[\log E_{i,t}] < 0 < \mathbb{E}[\log \frac{1}{k}\sum_i E_{i,t}]$, then does there exists a $d^* \in (0, 1-1/k)$ such that the population stochastically persists for $d \in (d^*, 1 - 1/k]$ and goes asymptotically extinct with probability one for $d \in (0, d^*)$?

3.2 Multi-Species Communities

No species is an island as species regularly interact with other species. To account for these interactions, lets extend (14) to account for n species. Within species i, there are k_i states for individuals and $X_{i,t} = (X_{i1,t}, \dots, X_{ik_i,t})$ is the vector of the densities of individuals in these different states. Then $X_t = (X_{1,t}, \dots, X_{n,t})$ is the densities of all species in all of their states and corresponds to the community state at time t. Multiplication by a $k_i \times k_i$ matrix $A_i(X_t, E_{t+1})$ updates the state of species i:

$$X_{i,t+1} = A_i(X_t, E_{t+1})X_{i,t} =: F_i(X_t, E_{t+1}) \text{ with } i = 1, 2, \dots, n. \tag{17}$$

Assume that each of the A_i satisfy Hypothesis 3.9.

To determine whether each species can increase when rare, consider the scenario where a subset of species are absent from the community (i.e. rare) and the remaining species coexist at an ergodic, stationary distribution μ for (17). Then, as in the single species case, we ask: do the rare species have a tendency to increase or decrease in this community context? Before pursuing this agenda, recall that stationarity means that μ is a probability measure on $\mathscr{S}\times\mathscr{E}$ such that (i) the marginal of μ on $\mathscr{E}$ is π i.e. $\pi(B) = \mu(\mathscr{S} \times B)$ for all $B \subset \mathscr{E}$ and (ii) if X_0, E_0 are drawn randomly from this distribution, then E_t, X_t follows this distribution for all time i.e.

$\mathbb{P}[(X_t, E_t) \in B] = \mu(B)$ for all t and Borel sets $B \subset \mathscr{S} \times \mathscr{E}$. Furthermore, ergodicity means that μ is indecomposable i.e. it can not be written as a convex combination of two other stationary distributions. Due to compactness of $\mathscr{E} \times \mathscr{S}$, stationary distributions always exist see, e.g., Arnold [2, Theorem 1.5.8].

By ergodicity, there exists a set of species $I \subset \{1, 2, \dots, n\}$ such that μ is only supported by these species i.e. $\mu(\{x \in \mathscr{S} : \|x_i\| > 0 \text{ if and only if } i \in I\} \times \mathscr{E}) = 1$. Suppose $i \notin I$ is one of the species not supported by μ and the sub-community I follows the stationary dynamics i.e. X_0, E_0 is randomly chosen with respect to μ. To determine whether or not species i has a tendency to increase or decrease when introduced at small densities $x_i = (x_{i1}, \dots, x_{ik_i}) \approx 0$, we can approximate the dynamics of species i with the linearized system

$$Z_{t+1} = B_{t+1} Z_t \text{ where } Z_0 = x_i \text{ and } B_{t+1} = A_i(X_t, E_{t+1}) \tag{18}$$

where X_t, E_t is following the stationary distribution given by μ. Iterating this matrix equation gives

$$Z_t = B_t B_{t-1} B_{t-2} \dots B_2 B_1 Z_0$$

As before, Proposition 3.2 from [45] and Birkhoff's ergodic theorem implies there is a quantity $r_i(\mu)$ such that

$$\lim_{t \to \infty} \frac{1}{t} \log \|Z_t\| = r_i(\mu) \text{ with probability one.}$$

Lets call $r_i(\mu)$ the *per-capita growth rate of species i when the community is in the stationary state given by* μ. For species $i \in I$ in the sub-community I,$r_i(\mu)$ can be defined in the same manner, but it will always equal zero [44, Proposition 8.19]. Intuitively for species not going extinct or growing without bound, the average per-capita growth rate is zero. In the words of [24],

> "a finite world can support only a finite population; therefore, population growth must eventually equal zero."

Using these per-capita growth rates, [44] proved the following theorem.

Theorem 3.11 *Let* $\mathscr{S}_0 = \{x \in \mathscr{S} : \prod \|x_i\| = 0\}$. *If there exist* $p_1, \dots, p_n > 0$ *such that*

$$\sum_i p_i r_i(\mu) > 0 \tag{19}$$

for all ergodic stationary distributions μ *supported by* $\mathscr{S}_0$, *then* (17) *is stochastically persistent.*

The stochastic persistence condition is the stochastic analog of a condition introduced by [25] for ordinary differential equation models. The sum in (19) is effectively only over the missing species as $r_i(\mu) = 0$ for all the species supported by μ. As the reverse of this condition implies that the extinction set $\mathscr{S}_0$ is an attractor for deterministic models, it is natural to raise the following question:

Problem 6 Let $\mathscr{S}_0 = \{x \in \mathscr{S} : \prod \|x_i\| = 0\}$. If there exist $p_1, \ldots, p_n > 0$ such that

$$\sum_i p_i r_i(\mu) < 0$$

for all ergodic stationary distributions μ supported by $\mathscr{S}_0$, then does it follow that for all $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\mathbb{P}[\lim_{t\to\infty} \mathrm{dist}(X_t, \mathscr{S}_0) = 0 | X_0 = x] \geq 1 - \varepsilon$$

whenever $\mathrm{dist}(x, \mathscr{S}_0) \leq \delta$?

For stochastic differential equations on the simplex, Benaïm et al. [6, Theorems 4.2,5.1] proved affirmative answers to this problem for systems with small or large levels of noise. In their case, $\mathscr{S}_0$ was shown to be a global attractor with probability one. This stronger conclusion will not hold in general.

We illustrate Theorem 3.11 with applications to competing species and stochastic Lotka-Volterra differences equations. In both examples, the interacting species are unstructured i.e. $k_i = 1$.

Example 8 (Competing Species and the Storage Effect) One of the fundamental principle in ecology is the competitive exclusion principle which asserts that two species competing for a single limiting resource (e.g. space, nutrients) can not coexist at equilibrium. However, many species which appear to be competing for a single resource do coexist. One resolution to this paradox for competing planktonic species was suggested by [28] who wrote

> "The diversity of the plankton [is] explicable primarily by a permanent failure to achieve equilibrium as the relevant external factors changes."

Intuitively, if environmental conditions vary such that each species has a period in which it does better than its competitors, then coexistence should be possible. Understanding exactly when this occurs is the focus of a series of papers by Peter Chesson and his collaborators [7, 11–14]. We illustrate one of the main conclusions from this work using a model from [12].

Consider two competing species with densities $X_t = (X_t^1, X_t^2)$ in year t. Let $E_{i,t}$ be the low-density per-capita reproductive output of species i, $s_i \in (0, 1)$ the probability of adults surviving to the next year, and $f : [0, \infty) \to (0, \infty)$ a continuously differentiable, decreasing function accounting for negative effects of competition on reproduction. If $C_t = E_{1,t}X_{1,t} + E_{2,t}X_{2,t}$ represents the "intensity of competition among the offspring", then we have the following model of competitive interactions

$$X_{i,t+1} = X_{i,t}\,\underbrace{(E_{i,t+1}f(C_t) + s_i)}_{A_i(X_t,E_{t+1})} \text{ where } C_t = E_{1,t}X_{1,t} + E_{2,t}X_{2,t}. \qquad (20)$$

To ensure that stochastic dynamics eventually enter a compact set $\mathscr{S}$, assume that $\lim_{x\to\infty} f(x) = 0$ and there exists $M > 0$ such that $E_{i,t} \in [0, M]$ for all i and t. The first assumption is satisfied for many models in population biology e.g. $f(x) = \exp(-cx)$ or $\frac{1}{1+cx^b}$ with $c > 0, b > 0$.

To apply Theorem 3.11, we need $p_1, p_2 > 0$ such that $p_1 r_1(\mu) + p_1 r_2(\mu) > 0$ for all ergodic stationary distributions μ supported by $\mathscr{S}_0 = \{x \in S : x_1 x_2 = 0\}$. There are three types of μ to consider: μ supports no species (i.e. $I = \emptyset$), μ only supports species 1 (i.e. $I = \{1\}$), or μ only supports species 2 (i.e. $I = \{2\}$). For μ supported on $\{(0, 0)\} \times \mathscr{E}$ i.e. no species are supported, the persistence condition demands

$$\sum_i p_i r_i(\mu) = \sum_i p_i \mathbb{E}[\log(E_{i,t} f(0) + s_i)] > 0. \tag{21}$$

For μ supported by $\{(x_1, 0) : x_1 > 0\} \times \mathscr{E}$, $r_1(\mu) = 0$ and the persistence criterion requires

$$\sum_i p_i r_i(\mu) = p_2 r_2(\mu) = p_2 \int \log(E_2 f(E_1 X_1) + s_2)\mu(dXdE) > 0. \tag{22}$$

As f is a decreasing function, this condition being satisfied implies

$$\int \log(E_2 f(0) + s_2)\mu(dXdE) = \mathbb{E}[\log(E_{2,t} f(0) + s_2)] > 0.$$

Similarly, for μ supported by $\{(0, x_2) : x_2 > 0\} \times \mathscr{E}$, we need

$$\sum_i p_i r_i(\mu) = p_1 r_1(\mu) = p_1 \int \log(E_1 f(E_2 X_2) + s_1)\mu(dXdE) > 0. \tag{23}$$

which implies

$$\int \log(E_1 f(0) + s_1)\mu(dXdE) = \mathbb{E}[\log(E_{1,t} f(0) + s_1)] > 0.$$

As inequalities (22) and (23) imply inequality (21) for any $p_1, p_2 > 0$, inequalities (22) and (23) imply stochastic persistence. These inequalities correspond to the classical mutual invasibility criterion [55]: if each of the species can increase when rare, the competing species coexist.

To verify whether or not these conditions are satisfied is, in general, a challenging issue. However, [12] developed a formula for the $r_i(\mu)$ when the competition is symmetric. Namely, $s_1 = s_2 =: s$, E_t are independent and identically distributed, and $E_{1,t}, E_{2,t}$ are exchangeable i.e. $P[(E_{1,t}, E_{2,t}) \in B] = P[(E_{2,t}, E_{1,t}) \in B]$ for any Borel $B \subset \mathscr{E} \times \mathscr{E}$. Before describing Chesson's formula, lets examine the dynamics of the deterministic case. Exchangeability and determinism imply there exists a constant $E > 0$ such that $E_{1,t} = E_{2,t} = E$ for all t. Hence, the deterministic model is given by

$$x_{i,t+1} = x_{i,t}\left(Ef(Ex_{1,t} + Ex_{2,t}) + s\right) \text{ with } i = 1, 2.$$

As $x_{1,t+1}/x_{2,t+1} = x_{1,t}/x_{2,t}$ for all t, all radial lines in the positive orthant are invariant. Provided $Ef(0) + s > 1$ (i.e. each species persists in the absence of competition), there exists a line of equilibria connecting the two axes. Regarding these neutral dynamics, [12] wrote

> "Classically, when faced with a deterministic model of this sort ecologists have concluded that only one species can persist when the likely effects a stochastic environment are taken into account. The reason for this conclusion is the argument that environmental perturbations will cause a random walk to take place in which eventually all but one species becomes extinct."

Dispelling this faulty expectation, [12] derived a formula for the $r_i(\mu)$. To describe this formula, assume inequality (21) holds and μ is an ergodic, stationary distribution supporting species 1. As the E_t are independent in time, μ can be written as a product measure $m \times \pi$ on $\mathscr{S} \times \mathscr{E}$ where π is given by Hypothesis 3.7. Define

$$h(E_1, E_2) = \int \log\left(E_2 f(x_1 E_1) + s\right)\, m(dx).$$

[12] showed that

$$r_2(\mu) = -\frac{1}{2}\mathbb{E}\left[\int_{E_{1,t}}^{E_{2,t}} \int_{E_{1,t}}^{E_{2,t}} \frac{\partial^2 h}{\partial E_1 \partial E_2}(E_1, E_2)\, dE_1 dE_2\right].$$

As f is a decreasing function,

$$\frac{\partial^2 h}{\partial E_1 \partial E_2}(E_1, E_2) = \frac{f'(x_1 E_1) x_1 s}{(E_2 f(x_1 E_1) + s)^2} < 0$$

whenever $s > 0$. Hence, $r_2(\mu) > 0$ provided that $\mathbb{P}[E_{1,t} > E_{2,t}] > 0$ (i.e. there is some variation) and $s > 0$. As this holds for any ergodic μ supporting species 1 and a similar argument yields $r_1(\mu) > 0$ for any ergodic μ supporting species 2, it follows that this symmetric version of the model is stochastically persistent (Fig. 8).

The analysis of this model highlights three key ingredients required for environmental fluctuations to mediate coexistence. First, there must periods of time such that each species has a higher birth rate i.e. $E_{1,t}$ and $E_{2,t}$ vary and are not perfectly correlated. Second, year to year survivorship needs to be sufficiently positive (i.e. $s > 0$ in the model) to ensure species can "store" the gains from one favorable period to the next favorable period. Finally, the increase in fitness due to good conditions for one species is greater in years when those conditions are worse for its competitor i.e. $\frac{\partial^2 h}{\partial E_1 \partial E_2}(E_1, E_2) < 0$. These are the key ingredients of the "storage effect" as introduced by [14].

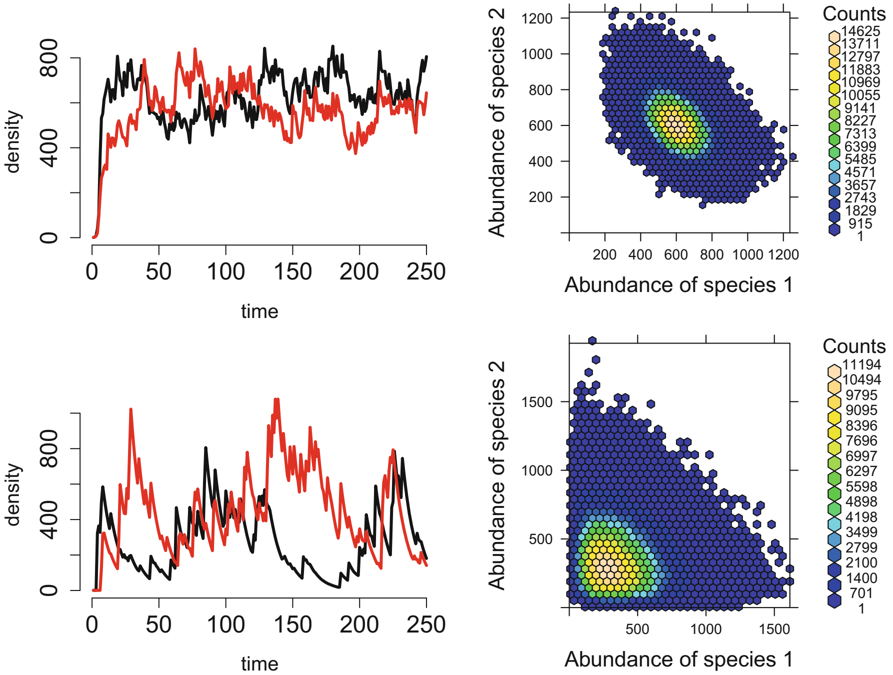


Fig. 8 Stochastic persistence of competing species from Example 8. Two simulations of model (20) with $f(x) = \exp(-0.001x)$, $E_{i,t}$ truncated log normals with log-mean 1 and log-variance 0.25 (upper row) and 25 (lower row). Models were run for $1,000,000$ time steps. Time series on the left show the first 250 time steps. The two dimensional histograms on the right correspond to the last $999,000$ time steps

Example 9 (Stochastic Lotka-Volterra Difference Equations) Previously, we studied the Poisson Lotka-Volterra processes which injected demographic stochasticity into the discrete-time Lotka-Volterra equations (1). Now, we examine the effects of injecting environmental stochasticity into these deterministic equations of n interacting species:

$$X_{i,t+1} = X_{i,t} \exp\left(\sum_{j=1}^{n} A_{ij} X_{j,t} + b_i + E_{i,t}\right) \tag{24}$$

where the matrix $A = (A_{ij})_{i,j}$ describes pairwise interactions between species, $b = (b_1, \ldots, b_n)$ describes the intrinsic rates of growth of each species in the absence of environmental fluctuations, and $E_{i,t}$ describes density-independent fluctuations. [55] used two dimensional versions of (24) to examine niche overlap of competitors in random environments.

The following lemma shows that verifying persistence for these equations reduces to a linear algebra problem. In particular, this lemma implies that the permanence criteria developed by [26] extend to these stochastically perturbed Lotka-Volterra systems.

Lemma 3.1 *Let μ be an ergodic stationary distribution for (24) and $I \subset \{1, \dots, k\}$ be the species supported by μ i.e. $\mu(\{x \in \mathcal{S} : x^i > 0 \text{ iff } i \in I\} \times \mathcal{E}) = 1$. Define $\beta_i = b_i + \mathbb{E}[E_{i,t}]$. If there exists a unique solution $\hat{x}$ to*

$$\sum_{j \in I} A_{ij}\hat{x}_j + \beta_i = 0 \text{ for } i \in I \text{ and } \hat{x}_i = 0 \text{ for } i \notin I \tag{25}$$

then

$$r_i(\mu) = \begin{cases} 0 & \text{if } i \in I \\ \sum_{j \in I} A_{ij}\hat{x}_j + \beta_i & \text{otherwise.} \end{cases}$$

The following proof of this lemma is nearly identical to the proof given by [50] for the case E_t are independent and identically distributed in time.

Proof Let μ and I be as assumed in the statement of the lemma. We have

$$r_i(\mu) = \sum_{j \in I} A_{ij} \int x_j \, \mu(dxdE) + \beta_i$$

for all i. As $r_i(\mu) = 0$ for all $i \in I$,

$$0 = \sum_{j \in I} A_{ij} \int x_j \, \mu(dxdE) + \beta_i$$

for all $i \in I$. Since we have assumed there is a unique solution $\hat{x}$ to this system of linear equations, it follows that $\int x_i \mu(dxdE) = \hat{x}_i$ for all i and the lemma follows.

□

This lemma implies that verifying the stochastic persistence condition reduces to finding $p_1, \dots, p_n > 0$ such that

$$\sum_{i \notin I} p_i \sum_{j \in I} A_{ij}\hat{x}_j + \beta_i > 0$$

for every $I \subset \{1, \dots, n\}$ and $\hat{x} \in \mathcal{S}_0$ satisfying equation (25). The next example illustrates the utility of this criterion.

Example 10 (Rock-Paper-Scissor Dynamics) The Lotka-Volterra model of rock-paper-scissor dynamics is a prototype for understanding intransitive ecological outcomes [37, 52]. Here, a simple stochastic version of this dynamic is given by (24) with X_1, X_2, X_3 corresponding to the densities of the rock, paper, and scissors populations, and the matrixes A and b given by

$$A = -1 + \begin{pmatrix} 0 & -\ell_2 & w_3 \\ w_1 & 0 & -\ell_3 \\ -\ell_1 & w_2 & 0 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

with $1 > w_i > 0$ and $\ell_i > 0$. The $-\ell_i$ correspond to a reduction in the per-capita growth rate of the population losing against population i, and w_i corresponds to the increase in the per-capita growth rate of the population winning against population i. Assume that the $E_{i,t}$ in (24) are compactly supported random variables with zero expectation. Under this assumption, β_i as defined in Lemma 3.1 equal 1.

Our assumptions about A and b imply that in pairwise interactions population 1 is excluded by population 2, population 2 is excluded by population 3, and population 3 is excluded by population 1. Hence, there are only four solutions of (25) that need to be considered: $\hat{x} = (0,0,0), \hat{x} = (1,0,0), \hat{x} = (0,1,0)$, and $\hat{x} = (0,0,1)$. Hence, verifying stochastic persistence reduces to determining whether there exist positive reals p_1, p_2, p_3 such that

$$\begin{aligned} p_1 + p_2 + p_3 &> 0 \\ p_1 \cdot 0 + p_2 w_1 - p_3 \ell_1 &> 0 \\ -p_1 \ell_2 + p_2 \cdot 0 + p_3 w_2 &> 0 \\ p_1 w_3 - p_2 \ell_3 + p_3 \cdot 0 &> 0 \end{aligned}$$

where these equation come from evaluating $\sum_i p_i r_i(\mu)$ at ergodic measures corresponding to $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$. Solving these linear inequalities implies that there is the desired choice of p_i if and only if $w_1 w_2 w_3 > \ell_1 \ell_2 \ell_3$ i.e. the geometric mean of the fitness payoffs to the winners exceeds the geometric mean of the fitness losses of the losers. Figure 9 illustrates the dynamics of coexistence when $w_1 w_2 w_3 > \ell_1 \ell_2 \ell_3$ and exclusion when $w_1 w_2 w_3 < \ell_1 \ell_2 \ell_3$.

4 Parting Thoughts and Future Challenges

The results reviewed here provide some ways to think about species coexistence or population persistence in the face of uncertainty. In the face of demographic uncertainty, species may coexist for exceptionally long periods of time prior to going extinct. I discussed how this metastable behavior may be predicted by the existence of positive attractors for the underlying deterministic dynamics, in which case the times to extinction increase exponentially with habitat size. Alternatively, in the face

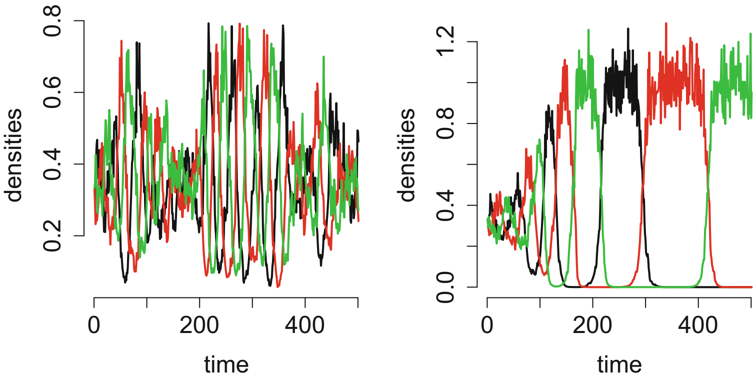


Fig. 9 Stochastic rock-paper-scissor dynamics (Example 10) with stochastic persistence in the left hand panel ($w_i = 0.3 > 0.2 = \ell_i$ for all i) and stochastic exclusion in the right hand panel ($w_i = 0.2 < 0.3 = \ell_i$ for all i)

of environmental stochasticity, species may coexist in the sense of rarely visiting low densities. I discussed how this form of stochastic persistence can be identified by examining species' per-capita growth rates $r_i(\mu)$ when rare. Weighted combinations of these per-capita growth rates can measure to what extent communities move away extinction as one or more species become rare. Despite this progress, many exciting challenges lie ahead.

Many demographic processes and environmental conditions vary continuously in time and are better represented by continuous time models. For continuous-time Markov chains accounting for demographic stochasticity, [36] proved results similar to Theorems 2.5 and 2.6 discussed here. For stochastic differential equations of interacting, unstructured populations in fluctuating environments, there exist some results similar to Theorem 3.11 by [6, 50] and [19]. These stochastic differential equations, however, fail to account for population structure or correlated environmental fluctuations. One step toward temporally correlated environments was recently taken by [4]. They characterized stochastic persistence for continuous-time models of competing species experiencing a finite number of environmental states driven by a continuous-time Markov chain. Generalizing these results to higher dimensional communities and structured populations remains an important challenge. Another exciting possibility is studying stochastic persistence for continuous-time models with stochastic birth or mortality impulses, as often observed in nature.

Biologists often measure continuous traits (e.g. body size or geographical location of an individual) that have important demographic consequences (e.g. larger individuals may produce more offspring and be more likely to survive). Unlike models accounting for discrete traits as considered here, models with continuous traits are infinite-dimensional and, consequently, even stochastic counterparts only accounting for demographic stochasticity correspond to Markov chains on uncountable state spaces (see, e.g., [53]). One form of these models, integral projection models (IPMs), have become exceptionally popular in the ecological literature in the past decade as they interface well with demographic data sets (see, e.g., [43] for a recent discussion). Consequently, there is a need for the development of the infinite-dimensional counterparts to the results presented here (see [23] for results for structured populations facing uncorrelated, environmental stochasticity).

For both forms of stochasticity, there are few results for demonstrating that populations are "extinction-prone" (e.g. limiting QSDs being supported by the extinction set in Theorem 2.5 or Benaïm et al. [6, Theorems 4.2,5.1] for stochastic differential equations). No study of persistence or coexistence is complete without understanding this complementary outcome. Hopefully, answers to Problems 2 and 6 will narrow our gap in understanding these outcomes. Furthermore, even when populations aren't extinction prone in the aforementioned sense, extinction is inevitable as all real population are finite. Answers to Problem 3 and their applications to specific models could provide new insights about how feedbacks between nonlinearities and noise determine the "intrinsic" extinction probabilities, quantities of particular importance for conservation biology.

Finally, there is the elephant in the review: what can one say for models accounting for both forms of stochasticity? At this point, all I have to offer is a natural conjecture which combines the results presented here. Namely, let $x_{t+1} = F(x_t, E_{t+1})$ be a random difference equation and $\{X_t^\varepsilon\}_{\varepsilon>0}$ be a family of Markov chains satisfying the environmental dependent versions of Hypotheses 2.3 and 2.4 e.g. the rate function ρ in Hypothesis 2.3 depends on $E \in \mathscr{E}$ as well as $x, y \in \mathscr{S}$. In light of the results presented here, these models lead to the following challenging problem:

Problem 7 Is it true that stochastic persistence of $x_{t+1} = F(x_t, E_{t+1})$ implies the weak* limit points of the QSDS of $\{X_t^\varepsilon\}_{\varepsilon>0}$ are supported by $\mathscr{S}_+$ and $\lambda_\varepsilon \geq 1 - \exp(-c/\varepsilon)$ for some $c > 0$?

I believe there should be an affirmative answer to this question. Namely, stochastic persistence in the face of environmental fluctuations implies long-term, persistent, metastable behavior for communities of interacting populations of finite size, and the extinction probabilities decay exponentially with community "size." Hopefully, this review will inspire work to address this problem as well as for the other challenges posed here.

Acknowledgements Many thanks to Swati Patel, William Cuello, and two anonymous reviews for providing extensive comments on an earlier version of this manuscript. This work was supported in part by US NSF Grant DMS-1313418 to the author.

References

1. P.B. Adler, S.P. Ellner, and J.M. Levine. Coexistence of perennial plants: an embarrassment of niches. *Ecology letters*, 13:1019–1029, 2010.
2. L. Arnold. *Random dynamical systems*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 1998. ISBN 3-540-63758-3.
3. A.D. Barbour. Quasi-Stationary Distributions in Markov Population Processes. *Advances in Applied Probability*, 8:296–314, 1976.
4. M. Benaïm and C. Lobry. Lotka Volterra in fluctuating environment or "how good can be bad". *arXiv preprint arXiv:1412.1107*, 2014.
5. M. Benaïm and S.J. Schreiber. Persistence of structured populations in random environments. *Theoretical Population Biology*, 76:19–34, 2009.
6. M. Benaïm, J. Hofbauer, and W. Sandholm. Robust permanence and impermanence for the stochastic replicator dynamics. *Journal of Biological Dynamics*, 2:180–195, 2008.
7. P. Chesson. Multispecies competition in variable environments. *Theoretical Population Biology*, 45(3):227–276, 1994.
8. P. Chesson. General theory of competitive coexistence in spatially-varying environments. *Theoretical Population Biology*, 58:211–237, 2000.
9. P. Chesson. Mechanisms of maintenance of species diversity. *Annual Review of Ecology and Systematics*, 31:343–366, 2000. ISSN 00664162.
10. P. L. Chesson. Predator-prey theory and variability. *Annu. Rev. Ecol. Syst.*, 9:323–347, 1978.
11. P. L. Chesson. The stabilizing effect of a random environment. *J. Math. Biol.*, 15(1):1–36, 1982.
12. P.L. Chesson. Interactions between environment and competition: how environmental fluctuations mediate coexistence and competitive exclusion. *Lecture Notes in Biomathematics*, 77:51–71, 1988.
13. P.L. Chesson and S. Ellner. Invasibility and stochastic boundedness in monotonic competition models. *Journal of Mathematical Biology*, 27:117–138, 1989.
14. P.L. Chesson and R.R. Warner. Environmental variability promotes coexistence in lottery competitive systems. *The American Naturalist*, 117(6):923, 1981.
15. J.N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing discrete-time finite markov chains. *Journal of Applied Probability*, 2:88–100, 1965.
16. A. Dembo and O. Zeitouni. *Large Deviation Techniques and Applications*. Applications of Mathematics: Stochastic Modelling and Applied Probability. Springer, 1993.
17. R. Durrett. *Probability: Theory and examples*. Duxbury Press, Belmont, CA, 1996.
18. S.N. Evans, P. Ralph, S.J. Schreiber, and A. Sen. Stochastic growth rates in spatio-temporal heterogeneous environments. *Journal of Mathematical Biology*, 66:423–476, 2013.
19. S.N. Evans, A Hening, and S.J. Schreiber. Protected polymorphisms and evolutionary stability of patch-selection strategies in stochastic environments. *Journal of Mathematical Biology*, 71:325–359, 2015.
20. M. Faure and S. J. Schreiber. Quasi-stationary distributions for randomly perturbed dynamical systems. *Annals of Applied Probability*, 24:553–598, 2014.
21. M. I. Freidlin and A. D. Wentzell. *Random perturbations of dynamical systems*, volume 260 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, New York, second edition, 1998. ISBN 0-387-98362-7. Translated from the 1979 Russian original by Joseph Szücs.
22. V. Grimm and C. Wissel. The intrinsic mean time to extinction: a unifying approach to analysing persistence and viability of populations. *Oikos*, 105:501–511, 2004.
23. D. P. Hardin, P. Takáč, and G. F. Webb. Asymptotic properties of a continuous-space discrete-time population model in a random environment. *Journal of Mathematical Biology*, 26:361–374, 1988a. ISSN 0303-6812.
24. G. Hardin. The tragedy of the commons. *Science*, 162:1243–1248, 1968.

25. J. Hofbauer. A general cooperation theorem for hypercycles. *Monatshefte für Mathematik*, 91:233–240, 1981.
26. J. Hofbauer, V. Hutson, and W. Jansen. Coexistence for systems governed by difference equations of Lotka-Volterra type. *Journal of Mathematical Biology*, 25:553–570, 1987.
27. G. Högnäs. On the quasi-stationary distribution of a stochastic Ricker model. *Stochastic Processes and their Applications*, 70:243–263, 1997.
28. G.E. Hutchinson. The paradox of the plankton. *The American Naturalist*, 95:137–145, 1961.
29. P. Jagers. A plea for stochastic population dynamics. *Journal of Mathematical Biology*, 60:761–764, 2010.
30. W. O. Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal society of London. Series A*, 115:700–721, 1927.
31. Y. Kifer. *Random perturbations of dynamical systems*. Birkhauser, New York, 1988.
32. F.C. Klebaner, J. Lazar, and O. Zeitouni. On the quasi-stationary distribution for some randomly perturbed transformations of an interval. *Annals of Applied Probability*, 8:300–315, 1998. ISSN 1050-5164.
33. O. S. Kozlovski. Axiom A maps are dense in the space of unimodal maps in the C^k topology. *Annals of Mathematics*, 157:1–43, 2003. ISSN 0003-486X.
34. Thomas G. Kurtz. *Approximation of population processes*, volume 36 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1981. ISBN 0-89871-169-X.
35. A. J. Lotka. *Elements of Physical Biology*. Williams and Witkins, Baltimore, 1925.
36. B. Marmet. Quasi-stationary distributions for stochastic approximation algorithms with constant step size. *arXiv preprint arXiv:1303.7081*, 2013.
37. R. M. May and W. Leonard. Nonlinear aspects of competition between three species. *SIAM Journal of Applied Mathematics*, 29:243–252, 1975.
38. J. F. McLaughlin, J. J. Hellmann, C. L. Boggs, and P. R. Ehrlich. Climate change hastens population extinctions. *Proceeding of the National Academy of Sciences USA*, 99:6070–6074, 2002.
39. S. Méléard and D. Villemonais. Quasi-stationary distributions and population processes. *Probability Surveys*, 9:340–410, 2012.
40. J.A.J. Metz, T.J. de Jong, and P.G.L. Klinkhamer. What are the advantages of dispersing; a paper by Kuno extended. *Oecologia*, 57:166–169, 1983.
41. A. J. Nicholson and V. A. Bailey. The balance of animal populations. *Proceedings of the Zoological Society of London*, pages 551–598, 1935.
42. K. Ramanan and O. Zeitouni. The quasi-stationary distribution for small random perturbations of certain one-dimensional maps. *Stochastic Processes and Applications*, 84:25–51, 1999. ISSN 0304-4149.
43. M. Rees, D.Z. Childs, and S.P. Ellner. Building integral projection models: a user's guide. *Journal of Animal Ecology*, 83:528–545, 2014.
44. G. Roth and S.J. Schreiber. Persistence in fluctuating environments for interacting structured populations. *Journal of Mathematical Biology*, 68:1267–1317, 2014.
45. D. Ruelle. Analycity properties of the characteristic exponents of random matrix products. *Advances in Mathematics*, 32:68–80, 1979. ISSN 0001-8708.
46. S. J. Schreiber. Allee effects, chaotic transients, and unexpected extinctions. *Theoretical Population Biology*, 2003.
47. S. J. Schreiber. Persistence despite perturbations for interacting populations. *Journal of Theoretical Biology*, 242:844–52, 2006.
48. S. J. Schreiber. On persistence and extinction of randomly perturbed dynamical systems. *Discrete and Continous Dynamical Systems B*, 7:457–463, 2007.
49. S. J. Schreiber. Persistence for stochastic difference equations: a mini-review. *Journal of Difference Equations and Applications*, 18:1381–1403, 2012.
50. S. J. Schreiber, M. Benaïm, and K. A. S. Atchadé. Persistence in fluctuating environments. *Journal of Mathematical Biology*, 62:655–683, 2011.

51. S.J. Schreiber. Interactive effects of temporal correlations, spatial heterogeneity and dispersal on population persistence. *Proceedings of the Royal Society B: Biological Sciences*, 277:1907–1914, 2010.
52. S.J. Schreiber and T.P. Killingback. Cycling in space: Persistence of rock-paper-scissor metacommunities. *Theoretical Population Biology*, 86:1–11, 2013.
53. S.J. Schreiber and N. Ross. Individual-based integral projection models: The role of size-structure on extinction risk and establishment success. *Methods in Ecology and Evolution*, in press. URL http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12537/abstract.
54. W. R. Thompson. La théorie mathématique de l'action des parasites entomophages et le facteur du hassard. *Annales Faculte des Sciences de Marseille*, 2:69–89, 1924.
55. M. Turelli. Niche overlap and invasion of competitors in random environments i. models without demographic stochasticity. *Theoretical Population Biology*, 20:1–56, 1981.
56. V. Volterra. Fluctuations in the abundance of a species considered mathematically. *Nature*, 118:558–560, 1926.

Part VIII
Number Theory and Algebraic Geometry in Cryptography and Other Applications

Computing Elliptic Curves over $\mathbb{Q}$: Bad Reduction at One Prime

Michael A. Bennett and Andrew Rechnitzer

Abstract We discuss a new algorithm for finding all elliptic curves over $\mathbb{Q}$ with a given conductor. Though based on (very) classical ideas, this approach appears to be computationally quite efficient. We provide details of the output from the algorithm in case of conductor p or p^2, for p prime, with comparisons to existing data.

Keywords Elliptic curves • Cubic forms • Invariant theory

1991 *Mathematics Subject Classification:* Primary 11G05, 11D25, 11D59, Secondary 11E76, 11Y50, 11Y65

1 Introduction

Elliptic curves are ubiquitous objects in pure mathematics, particularly in Number Theory and Algebraic Geometry. It is therefore of some interest to be able to generate or tabulate elliptic curves with desired properties. In this paper, we will describe an algorithm for computing models for all elliptic curves with integer coefficients and given *conductor*. This last quantity is an invariant that provides information about how a given elliptic curve behaves over finite fields $\mathbb{F}_p$, as p ranges over all primes. For the purposes of this paper, we will mostly restrict our attention to the case of conductor p or p^2, for prime p.

If K is a number field and S is a finite set of places of K, containing the infinite places, then a theorem of Shafarevich [42] from 1963 ensures that there are at most finitely many K-isomorphism classes of elliptic curves defined over K with good reduction outside S. In the simplest case, where $K = \mathbb{Q}$, an effective version of this result was proved by Coates [12] in 1970, using bounds for linear forms in p-adic

The authors were supported in part by grants from NSERC.

M.A. Bennett (✉) • A. Rechnitzer
Department of Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2
e-mail: bennett@math.ubc.edu; andrewr@math.ubc.edu

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_13

and complex logarithms. Early attempts to make such results explicit, for fixed sets of “small” primes S, have much in common with the arguments of [12], in that they (often) reduce the problem to one of solving a number of *Thue-Mahler equations*. These are Diophantine equations of the form

$$F(x, y) = u, \tag{1}$$

where F is a binary form (of degree at least 3) and u is an S-*unit*, that is, an integer whose prime factors all lie in S (strictly speaking, for $K = \mathbb{Q}$, we are assuming here that $2 \in S$). In case the form F is reducible in $\mathbb{Z}[x, y]$ (which turns out to be the case when the elliptic curves we are considering have at least one rational 2-torsion point), equation (1) typically is somewhat less challenging to solve. The earliest examples where a complete determination of all elliptic curves $E/\mathbb{Q}$ with good reduction outside a given set S was made were for $S = \{2, 3\}$ (by Coghlan [13] and Stevens (see e.g. [7])), and for $S = \{p\}$ for certain small primes p (by e.g. Setzer [41] and Neumann [35]).

The first case where such a determination was made with corresponding forms in equation (1) irreducible was for $S = \{11\}$, by Agrawal, Coates, Hunt and van der Poorten [1]. The reduction to (1) in this situation is not especially problematical, but subsequent computations (involving the arguments of [12] together with a variety of techniques from computational Diophantine approximation) are quite involved. For whatever reason, there are very few if any subsequent attempts in the literature to find elliptic curves of given conductor via Thue-Mahler equations. Instead, one finds a wealth of results on a completely different approach to the problem, using modular forms. This method relies upon the Modularity theorem of Breuil, Conrad, Diamond and Taylor [9], which was still a conjecture (under various guises) when these ideas were first implemented. Much of the success of this approach can be attributed to Cremona (see e.g. [14, 15]) and his collaborators, who have devoted decades of work to it (and are responsible for the current state-of-the-art). To apply this method to find all $E/\mathbb{Q}$ of conductor N, one computes the space of $\Gamma_0(N)$ modular symbols and the action of the Hecke algebra on it, and then searches for one-dimensional rational eigenspaces. After calculating a large number of Hecke eigenvalues, one is then able to extract corresponding elliptic curves. For a detailed description of how this technique works, the reader is directed to [15]. Via this method (assuming the results of [9]), all $E/\mathbb{Q}$ of conductor $N \leq 380000$ been determined by Cremona, as of April 2016.

In this paper, we will instead return to techniques based upon solving Thue-Mahler equations. Our goal is to provide a treatment that makes the connection between the conductors in question and the corresponding equations (1) straightforward, and the subsequent Diophantine approximation problem as painless as possible. We will rely upon a number of results from classical invariant theory and, for purposes of clarity and simplicity, focus our attention on curves with bad reduction at a single prime (i.e. curves of conductor p or p^2 for p prime). We will unconditionally find all curves of prime conductor up to $2 \cdot 10^9$ (10^{10} in the case of curves of positive discriminant) and conductor p^2 for $p \leq 10^6$. Conditionally, we extend these computations, in the case of prime conductor p, to $p \leq 10^{12}$. The outline of our paper is as follows. In Sect. 2, we will outline some basic facts and notation about elliptic curves. In Sect. 3, we will discuss the invariant theory of cubic forms and state our main theorem which provides our algorithm. Section 4 is devoted to the actual computation of the cubic forms we require. In Sect. 5, we discuss the special cases where $N = p$ or p^2 for p prime while, in Sect. 6, we provide a variety of computational details for these cases and an outline of a heuristic approach to the problem. Finally, in Sect. 7, we give an overview of our output, comparing it to previous results in the literature. In this paper, we concentrate on results specialized to the cases of conductor p and p^2, omitting both more general considerations and any proofs. More general results are described in forthcoming work of the authors [5]. Readers interested in the proofs of a number of results stated here as well as more extensive data should consult that paper. We are in the process of making our data more easily available through the LMFDB. Until this is completed, anyone interested should feel free to contact the authors.

2 Elliptic Curves

Let $S = \{p_1, p_2, \ldots, p_k\}$ be a set of rational primes. Suppose that we wish to find models for isomorphism classes of elliptic curve over $\mathbb{Q}$ with given conductor $N = p_1^{\alpha_1} \cdots p_k^{\alpha_k}$, where the α_i are positive integers. Such a curve has a minimal model

$$E \;:\; y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6$$

with the $a_i \in \mathbb{Z}$ and discriminant $\Delta_E = (-1)^\delta p_1^{\beta_1} \cdots p_k^{\beta_k}$, where the $\beta_i \geq \alpha_i$ are again positive integers and $\delta \in \{0, 1\}$. Writing

$$b_2 = a_1^2 + 4a_2, \quad b_4 = a_1a_3 + 2a_4, \quad b_6 = a_3^2 + 4a_6,$$
$$c_4 = b_2^2 - 24b_4 \text{ and } c_6 = -b_2^3 + 36b_2b_4 - 216b_6,$$

we find that

$$1728\Delta_E = c_4^3 - c_6^2$$

and

$$j_E = c_4^3/\Delta_E.$$

We therefore have

$$c_6^2 = c_4^3 + (-1)^{\delta+1}L, \tag{2}$$

where

$$L = 2^6 \cdot 3^3 \cdot p_1^{\beta_1} \cdots p_k^{\beta_k}.$$

For each prime p, since our model is minimal, we may suppose (via Tate's algorithm; see e.g. Papadopolous [36]), defining $\nu_p(x)$ to be the largest power of a prime p dividing a nonzero integer x, that

$$\min\{3\nu_p(c_4), 2\nu_p(c_6)\} < 12 + 12\nu_p(2) + 6\nu_p(3). \tag{3}$$

In fact, it is equation (2) that lies at the heart of our approach (see also Cremona and Lingham [17] for an approach to the problem that takes as its starting point equation (2), but then heads in a rather different direction).

3 Cubic Forms

Let us suppose that a, b, c and d are integers, and consider the binary cubic form

$$F(x, y) = ax^3 + bx^2y + cxy^2 + dy^3, \tag{4}$$

with discriminant

$$D_F = -27a^2d^2 + b^2c^2 + 18abcd - 4ac^3 - 4b^3d.$$

To such a form we associate a pair of covariants, the Hessian $H = H_F(x, y)$ given by

$$H = H_F(x, y) = -\frac{1}{4}\left(\frac{\partial^2 F}{\partial x^2}\frac{\partial^2 F}{\partial y^2} - \left(\frac{\partial^2 F}{\partial x \partial y}\right)^2\right)$$

and the Jacobian determinant of F and H, a cubic form $G = G_F$ defined via

$$G = G_F(x, y) = \frac{\partial F}{\partial x}\frac{\partial H}{\partial y} - \frac{\partial F}{\partial y}\frac{\partial H}{\partial x}.$$

Note that, explicitly,

$$H = (b^2 - 3ac)x^2 + (bc - 9ad)xy + (c^2 - 3bd)y^2$$

and

$$\begin{aligned}G = {} & (-27a^2d + 9abc - 2b^3)x^3 + (-3b^2c - 27abd + 18ac^2)x^2y \\ & +(3bc^2 - 18b^2d + 27acd)xy^2 + (-9bcd + 2c^3 + 27ad^2)y^3.\end{aligned}$$

These covariants satisfy the syzygy

$$4H(x,y)^3 = G(x,y)^2 + 27D_F F(x,y)^2. \tag{5}$$

We further have

$$\mathrm{Res}(F,G) = -8D_F^3 \text{ and } \mathrm{Res}(F,H) = D_F^2.$$

We can now state our main result, which leads to our algorithm.

Theorem 3.1 *Let $E/\mathbb{Q}$ be an elliptic curve of conductor $N = 2^\alpha 3^\beta N_0$, where N_0 is coprime to* 6. *Then there exists an integral binary cubic form F of discriminant*

$$D_F = (|\Delta_E|/\Delta_E)2^{\alpha_0}3^{\beta_0}N_1,$$

and relatively prime integers u and v with

$$F(u,v) = \omega_0u^3 + \omega_1u^2v + \omega_2uv^2 + \omega_3v^3 = 2^{\alpha_1} \cdot 3^{\beta_1} \cdot \prod_{p|N_0} p^{\kappa_p}, \tag{6}$$

such that E is isomorphic over $\mathbb{Q}$ to $E_{\mathscr{D}}$ where

$$\mathscr{D} = \prod_{p|\gcd(c_4(E),c_6(E))} p^{\min\{[\nu_p(c_4(E))/2],[\nu_p(c_6(E))/3]\}} \tag{7}$$

and

$$E_{\mathscr{D}} \ : \ 3^{[\beta_0/3]}y^2 = x^3 - 27\mathscr{D}^2H_F(u,v)x + 27\mathscr{D}^3G_F(u,v).$$

Here, $N_1 \mid N_0$,

$$(\alpha_0,\alpha_1) = \begin{cases} (2,0) \text{ or } (2,3) \text{ if } \alpha = 0, \\ (3,\geq 3) \text{ or } (2,\geq 4) \text{ if } \alpha = 1, \\ (2,1),(4,0) \text{ or } (4,1) \text{ if } \alpha = 2, \\ (2,1),(2,2),(3,2),(4,0) \text{ or } (4,1) \text{ if } \alpha = 3, \\ (2,\geq 0),(3,\geq 2),(4,0) \text{ or } (4,1) \text{ if } \alpha = 4, \\ (2,0) \text{ or } (3,1) \text{ if } \alpha = 5, \\ (2,\geq 0),(3,\geq 1),(4,0) \text{ or } (4,1) \text{ if } \alpha = 6, \\ (3,0) \text{ or } (4,0) \text{ if } \alpha = 7, \\ (3,1) \text{ if } \alpha = 8, \end{cases}$$

$$(\beta_0, \beta_1) = \begin{cases} (0,0) \ \textit{if} \ \beta = 0, \\ (0, \geq 1) \ \textit{or} \ (1, \geq 0) \ \textit{if} \ \beta = 1, \\ (3,0), (0, \geq 0) \ \textit{or} \ (1, \geq 0) \ \textit{if} \ \beta = 2, \\ (\beta, 0) \ \textit{or} \ (\beta, 1) \ \textit{if} \ \beta \geq 3, \end{cases}$$

and $\kappa_p \in \mathbb{Z}$ *with* $\kappa_p \in \{0,1\}$ *if* $p^2 \mid N_1$. *If* $\beta_0 \geq 3$, *we further have that* $3 \mid \omega_1$ *and* $3 \mid \omega_2$.

A few observations are worth making here. Firstly, there might actually exist a cubic form for which the corresponding Thue-Mahler equation has a solution, where the corresponding $E_{\mathscr{D}}$ has conductor $N_{E_{\mathscr{D}}} \neq N$ (this can occur if certain local conditions at 2 are not satisfied). These local conditions are easy to check and are a minor issue computationally. In practice, for producing tables of elliptic curves of bounded conductor, we will typically apply the above result to find all curves with bad reduction outside a fixed set of primes, working with a number of conductors simultaneously. For such a computation, every twist we encounter will have conductor of interest to us. It is also the case, that the cubic forms arising need not be either primitive (in the sense that $\gcd(\omega_0, \omega_1, \omega_2, \omega_3) = 1$) or irreducible. The former situation (i.e. that of imprimitive forms) can occur if each of the coefficients of F is divisible by 3. The latter occurs precisely when the curve E has at least one rational 2-torsion point. We note that necessarily

$$\mathscr{D} \mid 2^3 \cdot 3^2 \cdot \prod_{p \mid N_0} p, \tag{8}$$

so that, given N, there are a finite set of $E_{\mathscr{D}}$ to consider.

In the event that, for a given binary form $F(x,y) = ax^3 + bx^2y + cxy^2 + dy^3$, we have $3 \mid b$ and $3 \mid c$, say $b = 3b_0$ and $c = 3c_0$, then we have that $27 \mid D_F$ and can write $D_F = 27\widetilde{D}_F$, where

$$\widetilde{D}_F = -a^2d^2 + 6ab_0c_0d + 3b_0^2c_0^2 - 4ac_0^3 - 4b_0^3d.$$

One may observe that the set of forms with both $3 \mid b$ and $3 \mid c$ is closed within the larger set of all binary cubic forms in $\mathbb{Z}[x,y]$, under the action of both $\mathrm{SL}_2(\mathbb{Z})$ and $\mathrm{GL}_2(\mathbb{Z})$. Note that, for such a form, we have

$$\widetilde{H}_F(x,y) = \frac{H_F(x,y)}{9} = (b_0^2 - ac_0)x^2 + (b_0c_0 - ad)xy + (c_0^2 - b_0d)y^2$$

and

$$\begin{aligned} \widetilde{G}_F(x,y) = \tfrac{G_F(x,y)}{27} = (-a^2d + 3ab_0c_0 - 2b_0^3)x^3 + 3(-b_0^2c_0 - ab_0d + 2ac_0^2)x^2y \\ +3(b_0c_0^2 - 2b_0^2d + ac_0d)xy^2 + (-3b_0c_0d + 2c_0^3 + ad^2)y^3, \end{aligned}$$

whereby our syzygy now becomes

$$4\widetilde{H}_F(x,y)^3 = \widetilde{G}_F(x,y)^2 + \widetilde{D}_F F(x,y)^2. \tag{9}$$

Theorem 3.1 is based upon a generalization of a very classical result of Mordell [32] (see also Theorem 3 of Chapter 24 of Mordell [33]), where the Diophantine equation $X^2 + kY^2 = Z^3$ is treated through reduction to binary cubic forms and their covariants, under the assumption that X and Z are coprime. That this last restriction could be eliminated, with some care, was noted by Sprindzuk (see Chapter VI of [44]).

Converting Theorem 3.1 into an algorithm for finding all $E/\mathbb{Q}$ of conductor N is a straightforward exercise. We proceed as follows.

(1) Compute $\mathrm{GL}_2(\mathbb{Z})$-representatives for every binary form F with discriminant

$$\Delta_F = \pm 2^{\alpha_0} 3^{\beta_0} N_1$$

for each divisor N_1 of N_0, and each possible pair (α_0, β_0) given in the statement of Theorem 3.1. The (very efficient) algorithm for carrying this out is described in detail in Sect. 4.

(2) Solve the corresponding Thue-Mahler equations. This is a deterministic procedure (see Tzanakis and de Weger [47, 48]) but not, in general, one that could reasonably be described as routine.

(3) Check "local" conditions and output the elliptic curves that arise.

As we shall see, the first and third of these steps are straightforward (indeed, the third is essentially trivial). All of the real work is concentrated in step (2). In Sect. 5, we will focus our attention on carrying out this procedure in the special case where $N = p$ or $N = p^2$ for p prime. For these conductors, we encounter the happy circumstance that the Thue-Mahler equations (6) reduce to Thue equations (i.e. where the exponents on the right hand side of (6) are all absolutely bounded). In such a situation, there are easily implemented computational routines for solving such equations, available in Pari/GP [37] or in Magma [8]. Further, it is possible to apply a much more computationally efficient argument to find all such elliptic curves heuristically (but not deterministically). We will describe such an approach later in the paper, in Sect. 6.

4 Finding Representative Forms

As we have seen, in order to find elliptic curves over $\mathbb{Q}$ with good reduction outside a given set of primes, it suffices to determine a set of representatives for $\mathrm{GL}_2(\mathbb{Z})$-equivalence classes of binary cubic forms with certain discriminants, and then solve a number of corresponding Thue-Mahler equations. In this section, we will describe how to find distinguished *reduced* representatives for equivalence classes of cubic forms with a given discriminant. In each case, the notion of *reduction* is related to associating to a given cubic form a particular definite quadratic form—in case of positive discriminant, for example, the Hessian H defined earlier. In what follows, we will state our definitions of reduction solely in terms of the coefficients of the given cubic form, keeping the associated Hessian hidden.

4.1 Forms of Positive Discriminant

In the case of positive discriminant forms, we will appeal to a classical reduction theory, dating back to work of Hermite [27, 28] and later used by Davenport (see e.g. [18, 19] and [20]). This procedure allows us to determine a *reduced* element within a given equivalence class of forms. We will assume the forms we are treating are irreducible, (and treat the case of reducible forms somewhat differently). We follow work of Belabas [2] (see also Belabas and Cohen [3] and Cremona [16]), a modern treatment and refinement of Hermite's method.

Definition 4.1 An irreducible binary integral cubic form

$$F(x,y) = ax^3 + bx^2y + cxy^2 + dy^3$$

of positive discriminant is called *reduced* if we have

- $|bc - 9ad| \leq b^2 - 3ac \leq c^2 - 3bd$,
- $a > 0, b \geq 0$, where $d < 0$ whenever $b = 0$,
- if $bc = 9ad, d < 0$,
- if $b^2 - 3ac = bc - 9ad, b < |3a - b|$, and
- if $b^2 - 3ac = c^2 - 3bd, a \leq |d|$, and $b < |c|$ whenever $|d| = a$.

The main value of this notion of reduction is in the following result (Corollary 3.3 of [2]).

Proposition 4.1 *Any irreducible cubic form with positive discriminant is* $GL_2(\mathbb{Z})$*-equivalent to a unique reduced one.*

To determine equivalence classes of reduced cubic forms with bounded discriminant, we will appeal to the following result (Lemma 3.5 of Belabas [2]).

Lemma 4.2 *Let X be a positive real number and*

$$F(x,y) = ax^3 + bx^2y + cxy^2 + dy^3$$

be a reduced form whose discriminant lies in $(0, X]$*. Then we have*

$$1 \leq a \leq \frac{2X^{1/4}}{3\sqrt{3}}$$

and

$$0 \leq b \leq \frac{3a}{2} + \left(\sqrt{X} - \frac{27a^2}{4}\right)^{1/2}.$$

If we denote by P_2 the unique positive real solution of the equation

$$-4P_2^3 + (3a+2b)^2 P_2^2 + 27a^2 Z = 0,$$

then

$$\frac{b^2 - P_2}{3a} \leq c \leq b - 3a.$$

4.2 Forms of Negative Discriminant

In case of negative discriminant, we require a different notion of reduction, as the Hessian is no longer a definite form. We will instead, following Belabas [2], use an idea of Berwick and Mathews [6]. We take as our definition of a reduced form an alternative characterization due to Belabas (Lemma 4.2 of [2]).

Definition 4.2 An irreducible binary integral cubic form

$$F(x,y) = ax^3 + bx^2y + cxy^2 + dy^3$$

of negative discriminant is called *reduced* if we have

- $d^2 - a^2 > bd - ac,$
- $-(a-b)^2 - ac < ad - bc < (a+b)^2 + ac,$
- $a > 0,\ b \geq 0$ and $d > 0$ whenever $b = 0.$

Analogous to Proposition 4.1, we have, as a consequence of Lemma 4.3 of [2] :

Proposition 4.3 *Any irreducible cubic form with negative discriminant is $GL_2(\mathbb{Z})$-equivalent to a unique reduced one.*

To count the number of reduced cubic forms in this case, we use Lemma 4.4 of Belabas [2] :

Lemma 4.4 *Let X be a positive real number and*

$$F(x,y) = ax^3 + bx^2y + cxy^2 + dy^3$$

be a reduced form whose discriminant lies in $[-X, 0)$. Then we have

$$1 \leq a \leq \left(\frac{16X}{27}\right)^{1/4}$$

$$0 \leq b \leq \frac{3a}{2} + \left(\sqrt{X/3} - \frac{3a^2}{4}\right)^{1/2}$$

$$1 - b \leq c \leq \left(\frac{X}{4a}\right)^{1/3} + \begin{cases} b^2/3a & \textit{if } a \geq 2b/3, \\ b - 3a/4 & \textit{otherwise.} \end{cases}$$

It is worth noting here that a different notion of reduction for cubic forms of negative discriminant is described in Cremona [16], arising from classical work of Julia [29]. This definition leads to shorter loops for the coefficient a and a slight improvement in the expected complexity (though the number of (a,b,c,d) one treats still grows linearly in the variable X).

The techniques we have described here provide a computationally efficient way to write down representatives for classes of irreducible cubic form with bounded absolute discriminant. The problem of finding all such forms of a fixed discriminant (without computing those of smaller discriminant) is a slightly different one. One approach would be to loop over the first three coefficients a, b, c of the form as previously, and then solve the corresponding quadratic equation for d. Even a relatively simplistic approach like this makes it computationally feasible to find forms of a desired, fixed discriminant exceeding 10^{15}.

4.3 Reducible Forms

We can define somewhat similar notions of reduction for reducible forms (see e.g. [4]). For our purposes, though, it is enough to recall that we may suppose that a reduced form is equivalent to one of the shape

$$F(x,y) = bx^2y + cxy^2 + dy^3 \text{ with } 0 \leq d \leq c,$$

whereby we have

$$\Delta_F = b^2(c^2 - 4bd).$$

To determine all elliptic curves with good reduction outside $S = \{p_1, p_2, \ldots, p_k\}$, corresponding to reducible cubics in Theorem 3.1 (i.e. those E with at least one rational 2-torsion point), it suffices to find all such triples (b,c,d) for which there exists integers x and y with, writing $S^* = S \cup \{2\}$, both $b^2(c^2 - 4bd)$ and $bx^2y + cxy^2 + dy^3$ S^*-units. For this to occur, it is clearly necessary that $b, c^2 - 4bd, y$ and $\mu = bx^2 + cxy + dy^2$ are S^*-units. Taking the discriminant of this last quadratic as a function of x, we thus require that

$$(c^2 - 4bd)y^2 + 4b\mu = Z^2, \tag{10}$$

for some integer Z. This is an equation of the shape

$$X + Y = Z^2 \tag{11}$$

in S^*-units X and Y. There is an algorithm for solving such equations described in detail in Chapter 7 of de Weger [49] (see also [50]), relying upon bounds for linear forms in p-adic and complex logarithms and various reduction techniques. As of now, we are unaware of any implementation of this algorithm in available computational algebra packages. While a priori equation (10) arises as only a necessary condition for the existence of an elliptic curve of the desired form, given any solution to (10), the curve

$$E \; : \; y^2 = x^3 + Zx^2 + b\mu x$$

has discriminant

$$\Delta_E = 16b^2\mu^2(Z^2 - 4b\mu) = 16b^2\mu^2(c^2 - 4bd)y^2,$$

and hence good reduction outside S^*.

4.4 A Final Note

One last observation which is necessary here before we proceed is that while G_F^2 is $\mathrm{GL}_2(\mathbb{Z})$-covariant, the same is not actually true for G_F (it is, however, an $\mathrm{SL}_2(\mathbb{Z})$-covariant). This may seem like a subtle point, but what it means for us in practice is that, having found our $\mathrm{GL}_2(\mathbb{Z})$-representative forms F and corresponding curves of the shape $E_{\mathscr{D}}$ from Theorem 3.1, we need also check to see if

$$\tilde{E}_{\mathscr{D}} \quad : \quad 3^{[\beta_0/3]}y^2 = x^3 - 27\mathscr{D}^2 H_F(u,v)x - 27\mathscr{D}^3 G_F(u,v),$$

the quadratic twist of $E_{\mathscr{D}}$ by -1, yields a curve of the desired conductor.

5 Conductors $N = p$ and $N = p^2$

In the case where we want to find elliptic curves E of conductor $N = p$ prime, as noted earlier, things are especially simple. Suppose that E is such curve with invariants c_4 and c_6. From Papadopolous [36], we necessarily have

$$\begin{gathered}(\nu_p(c_4), \nu_p(c_6)) = (0,0) \text{ and } \nu_p(L) \geq 1,\\ (\nu_2(c_4), \nu_2(c_6)) = (0,0) \text{ or } (\geq 4, 3), \text{ and } \nu_2(L) = 6,\\ (\nu_3(c_4), \nu_3(c_6)) = (0,0) \text{ or } (1, \geq 3), \text{ and } \nu_3(L) = 3,\end{gathered}$$

and hence $\mathscr{D} = 1$ or 2. Theorem 3.1 thus implies that there is a cubic form of discriminant ± 4 or $\pm 4p$, and integers u, v, with

$$F(u, v) = p^n \text{ or } 8p^n,\ c_4 = \mathscr{D}^2 H_F(u, v) \text{ and } c_6 = -\frac{1}{2}\mathscr{D}^3 G_F(u, v),\ \mathscr{D} \in \{1, 2\},$$

for some integer n. Similarly, if $N = p^2$, we are interested in finding cubic forms of discriminant $\pm 4 \cdot p^\tau$ for $\tau \in \{0, 1, 2\}$, and solving $F(x, y) = 8 \cdot p^n$, where $n \in \{0, 1\}$ if $\tau = 2$. In this situation, we have that $\mathscr{D} \mid 2p$.

If we first consider the case of a curve E of conductor p, appealing to Théorème 2 of Mestre and Oesterlé [30] (and using [9]), we either have $\Delta_E = \pm p$, or our prime $p \in \{11, 17, 19, 37\}$, or we have $p = t^2 + 64$ for some integer $t \equiv 1 \mod 4$ and our curve E is isomorphic to that given by

$$y^2 + xy = x^3 + \frac{t-1}{4}x^2 + 4x + t.$$

In this case, we have a rational point of order 2 given by $(x, y) = (-t/4, t/8)$ and discriminant $(t^2 + 64)^2$. Excluding these latter cases, in the notation of the preceding section, we thus have $\alpha_0 = 2, \alpha_1 \in \{0, 3\}, \beta_0 = \beta_1 = 0, \kappa_p = 0$ and $N_1 \in \{1, p\}$. We are therefore interested in finding all binary cubic forms (reducible and irreducible) F of discriminant ± 4 and $\pm 4p$ and subsequently solving

$$F(x, y) \in \{1, 8\}.$$

Next consider when E has conductor $N = p^2$, so that $p \mid c_4$ and $p \mid c_6$. From (3), we may suppose that $(\nu_p(c_4), \nu_p(c_6), \nu_p(\Delta_E))$ is one of

$$(\geq 1, 1, 2), (1, \geq 2, 3), (\geq 2, 2, 4), (2, 3, \geq 7), (\geq 3, 4, 8), (3, \geq 5, 9) \text{ or } (\geq 4, 5, 10),$$

or we have that $(\nu_p(c_4), \nu_p(c_6), \nu_p(\Delta_E)) = (\geq 2, \geq 3, 6)$. In this last case, the quadratic twist of our curve E by $(-1)^{(p-1)/2}p$ has good reduction at p and hence conductor 1, a contradiction. If we have $(\nu_p(c_4), \nu_p(c_6), \nu_p(\Delta_E)) = (2, 3, \geq 7)$, then E necessarily arises as the $(-1)^{(p-1)/2}p$-twist of a curve of conductor p, say E_1, with corresponding $(\nu_p(c_4(E_1)), \nu_p(c_6(E_1)), \nu_p(\Delta_{E_1})) = (0, 0, \nu_p(\Delta_E) - 6)$. Similarly, curves with $(\nu_p(c_4), \nu_p(c_6), \nu_p(\Delta_E)) = (\geq 3, 4, 8)$ arise as twists of those with $(\nu_p(c_4), \nu_p(c_6), \nu_p(\Delta_E)) = (\geq 1, 1, 2)$, those with $(\nu_p(c_4), \nu_p(c_6), \nu_p(\Delta_E)) = (3, \geq 5, 9)$ come from ones with $(\nu_p(c_4), \nu_p(c_6), \nu_p(\Delta_E)) = (1, \geq 2, 3)$, and those with $(\nu_p(c_4), \nu_p(c_6), \nu_p(\Delta_E)) = (\geq 4, 5, 10)$ from ones with $(\nu_p(c_4), \nu_p(c_6), \nu_p(\Delta_E)) = (\geq 2, 2, 4)$.

Supposing we have already computed all curves of conductor p, it remains therefore, up to twisting, to find $E/\mathbb{Q}$ with minimal discriminant

$$\Delta_E \in \{\pm p^2, \pm p^3, \pm p^4\}$$

(as noted by Edixhoven, de Groot and Top in Lemma 1 of [21]). In particular, from Theorem 3.1, we are led to consider equations of the shape

$$F(x, y) = 8 \quad \text{for } F \text{ a form of discriminant } \pm 4p^2, \tag{12}$$

$$F(x, y) = 8p \quad \text{for } F \text{ a form of discriminant } \pm 4p \tag{13}$$

and

$$F(x, y) = 8p \quad \text{for } F \text{ a form of discriminant } \pm 4p^2, \tag{14}$$

corresponding to $\Delta_E = \pm p^2, \pm p^3$ and $\pm p^4$, respectively.

5.1 *Reducible Forms*

To find all elliptic curves $E/\mathbb{Q}$ with conductor p or p^2 arising (in the notation of Theorem 3.1) from reducible forms, we are led to solve the equation

$$F(x, y) = 8\,p^n, \quad n \in \mathbb{Z}, \quad \gcd(x, y) \mid 2,$$

for reducible binary cubic forms of discriminant ± 4, $\pm 4p$ and $\pm 4p^2$. This is an essentially elementary exercise (if somewhat painful). Alternatively, we may note that the elliptic curves of conductor p or p^2 arising from reducible cubic forms are precisely those with at least one rational 2-torsion point and hence we can appeal to Theorem I of Hadano [24] to the effect that the only such p are $p = 7, 17$ and $p = t^2 + 64$ for integer t.

In any case, after a little work, we can show that the elliptic curves of conductor p or p^2 corresponding to reducible forms, are precisely those given by

(c_4, c_6)	p	Δ_E	N_E
$(273, 4455)$	17	17^2	17
$(33, 12015)$	17	-17^4	17
$(p-256, -t(p+512))$	t^2+64	$-p^2$	p
$(105, 1323)$	7	-7^3	7^2
$(1785, 75411)$	7	7^3	7^2
$(33, -81)$	17	17^3	17
$(4353, 287199)$	17	17	17
$(p-16, -t(p+8))$	t^2+64	p	p

Here, for the sake of concision, we omit quadratic twists by $\pm p$ of conductor p^2.

5.2 Irreducible Forms: Conductor *p*

It is straightforward to show that there are no irreducible cubic forms of discriminant ±4. If we begin by searching for elliptic curves of conductor p coming from irreducible cubics, we thus need to solve equations of the shape $F(x, y) = 8$ for all cubic forms of discriminant $\pm 4p$.

5.3 Irreducible Forms: Conductor p^2

As noted earlier, to find the elliptic curves of conductor p^2 coming from irreducible cubics, we need to find those of conductor p and those of conductor p^2 with $\Delta_F = \pm p^2$, $\pm p^3$ and $\pm p^4$ (and subsequently twist them).

5.3.1 Elliptic Curves of Discriminant $\pm p^3$

For these, we can use the cubic forms of discriminant $\Delta_F = \pm 4p$ we have already found in the course of computing curves of conductor p, and then solve the Thue equation $F(x, y) = 8p$. We can either do this directly, or reduce this problem to one of solving a pair of new Thue equations of the shape $G_i(x, y) = 8$. To see how this "reduction" proceeds, note, since we assume that $p \parallel \Delta_F$, we have, for $F(x, y) = ax^3 + bx^2y + cxy^2 + dy^3$,

$$F(x, y) \equiv a(x - r_0 y)^2 (x - r_1 y) \mod p,$$

where, since we may suppose that F is a reduced form (whereby $1 \leq a < p$), we necessarily have that $p \nmid a$. We thus obtain

$$2r_0 + r_1 \equiv -b/a \mod p,$$
$$r_0^2 + 2r_0 r_1 \equiv c/a \mod p$$

and

$$r_0^2 r_1 \equiv -d/a \mod p.$$

From the first two of these, we have

$$3ar_0^2 + 2br_0 + c \equiv 0 \mod p$$

and so, assuming that $t^2 \equiv b^2 - 3ac \mod p$,

$$(r_0, r_1) \equiv (3a)^{-1} \left(-b \pm t, -b \mp 2t\right) \mod p.$$

Given these two pairs, we are left to check to see which one satisfies $r_0^2 r_1 \equiv -d/a \mod p$.

To list our pairs (r_0, r_1), we need to find a square root of $b^2 - 3ac$ modulo p. There are efficient ways to do this via the Tonelli-Shanks algorithm, for example (and almost trivially if, say, $p \equiv 3 \mod 4$).

Given that we know r_0 and r_1, we thus have, if $F(x, y) = 8p$, either $x \equiv r_0 y \mod p$ or $x \equiv r_1 y \mod p$. In either case, we write $x = r_i y + pu$ so that, from $ax^3 + bx^2y + cxy^2 + dy^3 = 8p$, we are led to solve the two equations $G_i(u, y) = 8$, where

$$G_i(u, y) = ap^2u^3 + (3apr_i + bp)u^2y + (3ar_i^2 + 2br_i + c)uy^2 + \frac{1}{p}(ar_i^3 + br_i^2 + cr_i + d)y^3.$$

We observe that $\Delta_{G_i} = p^2 \Delta_F$.

In practice, for our deterministic approach, we will actually solve the equation $F(x, y) = 8p$ directly. For our heuristic approach (where a substantial increase in the size of the form's discriminant is not especially problematic), we will reduce to consideration of the equations $G_i(x, y) = 8$.

We note that there are (conjecturally infinite) families of primes for which we can guarantee that the equation $F(x, y) = 8p$ has solutions. For example, if we write $p_{r,s} = r^4 + 9r^2s^2 + 27s^4$, then, if $p = p_{r,s}$ for some choice of integers r and s, we have that the cubic form

$$F(x, y) = sx^3 + rx^2y - 3sxy^2 - ry^3$$

has discriminant $4p$. Further, we have a polynomial identity $F(x, y) = 8p$ for $x = 2r^2/s + 6s$ and $y = -2r$, or if $x = 6s$ and $y = -18s^2/r - 2r$. In particular, this provides four one-parameter families of primes for which there exists a cubic form F of discriminant $4p$ and integers x and y such that $F(x, y) = 8p$. Specifically, we have, choosing $s \in \{1, 2\}$, in the first case and $r \in \{1, 2\}$ in the second, i.e.

$$\begin{aligned}(p, x, y) = {} & (r^4 + 9r^2 + 27, 2r^2 + 6, -2r), (r^4 + 36r^2 + 432, r^2 + 12, -2r),\\ & (27s^4 + 9s^2 + 1, 6s, -18s^2 - 2), (27s^4 + 36s^2 + 16, 6s, -9s^2 - 4).\end{aligned}$$

Similar, if $p_{r,s} = r^4 - 9r^2s^2 + 27s^4$, the form

$$F(x, y) = sx^3 + rx^2y + 3sxy^2 + ry^3$$

has discriminant $-4p$. The equation $F(x, y) = 8p$ has solutions

$$(x, y) = (-2r^2/s + 6s, 2r) \text{ and } (6s, -18s^2/r + 2r)$$

and hence we again find (one parameter) families of primes corresponding to either r or s in $\{1, 2\}$:

$$\begin{aligned}(p,x,y) = {} & (r^4-9r^2+27,-2r^2+6,2r),(r^4-36r^2+432,-r^2+12,2r),\\ & (27s^4-9s^2+1,6s,-18s^2+2),(27s^4-36s^2+16,6s,-9s^2+4).\end{aligned}$$

We expect that each of the quartic families described here attains infinitely many prime values, but proving this is beyond current technology.

5.3.2 Elliptic Curves of Discriminant p^2 and p^4

Elliptic curves of discriminant p^2 and p^4 arise from solving the Thue equations $F(x,y)=8$ and $F(x,y)=8p$, respectively, for cubic forms F of discriminant $4p^2$. In order for there to exist a cubic form of discriminant $4p^2$, it is necessary and sufficient that we are able to write $p=r^2+27s^2$ for positive integers r and s, whereby F is equivalent to the form

$$F_{r,s}(x,y)=sx^3+rx^2y-9sxy^2-ry^3.$$

From this we are led to solve

$$F_{r,s}(x,y)=8 \text{ and } F_{r,s}(x,y)=8p.$$

In the latter case, we may, if we choose, reduce the equation to a single Thue equation of the form $G_{r,s}(x,y)=8$. To see this, note that we may suppose that $p\nmid y$. It follows that the congruence

$$su^3+ru^2-9su-r\equiv 0 \mod p$$

has a single solution modulo p (since $p^2\mid\Delta_F$), given (as is readily checked) by $r_0\equiv 9r^{-1}s \mod p$. We thus have $x\equiv r_0y \mod p$, so that, writing $x=r_0y+vp$, we have

$$F_{r,s}(r_0y+vp,y)=p(a_0v^3+b_0v^2y+c_0vy^2+d_0y^3)$$

and hence, renaming v,

$$G_{r,s}(x,y)=a_0x^3+b_0x^2y+c_0xy^2+d_0y^3=8,$$

where

$$a_0=sp^2,\ b_0=(3r_0s+r)p,\ c_0=3r_0^2s+2rr_0-9s \text{ and } d_0=(r_0^3s+rr_0^2-9r_0s-r)/p.$$

We observe that

$$\Delta_{G_{r,s}}=4p^4.$$

Once again, for our deterministic approach, we solve the equation $F_{r,s}(x,y) = 8p$ directly, while, for our heuristic approach, we consider instead the equation $G_{r,s}(x,y) = 8$.

5.3.3 Elliptic Curves of Discriminant $-p^2$ and $-p^4$

Elliptic curves of discriminant $-p^2$ and $-p^4$ arise from again solving the Thue equations $F(x,y) = 8$ and $F(x,y) = 8p$, respectively, this time for cubic forms F of discriminant $-4p^2$. For such form to exist, we require that $p = |r^2 - 27s^2|$ for integers r and s (so that these primes are precisely those of the form $\pm 1 \mod 12$) and find that F is necessarily equivalent to

$$F_{r,s}(x,y) = sx^3 + rx^2y + 9sxy^2 + ry^3.$$

If we wish to solve $F_{r,s}(x,y) = 8p$, as previously, we may note that, if $r_0 \equiv -9r^{-1}s \mod p$, then

$$sr_0^3 + rr_0^2 + 9sr_0 + r \equiv r^{-3}(r^2 - 27s^2)(r^2 + 27s^2) \equiv 0 \mod p.$$

Again write $x = r_0y + vp$, so that, renaming v, we have

$$G_{r,s}(x,y) = a_0x^3 + b_0x^2y + c_0xy^2 + d_0y^3 = 8,$$

where now

$$a_0 = sp^2,\ b_0 = (3r_0s+r)p,\ c_0 = 3r_0^2s+2rr_0+9s \text{ and } d_0 = (r_0^3s+rr_0^2+9r_0s+r)/p.$$

While it is not immediately obvious that, given we know the existence of integers r and s such that $p = |r^2 - 27s^2|$, we can actually find them, it is, in fact, computationally straightforward to do so, via the following result, an almost direct consequence Theorem 112 of Nagell [34] :

Proposition 5.1 *If $p \equiv 1 \mod 12$ is prime, there exist positive integers r and s such that*

$$r^2 - 27s^2 = p$$

and

$$r < \frac{3}{2}\sqrt{6p},\ \ s < \frac{5}{18}\sqrt{6p}.$$

If $p \equiv -1 \mod 12$ is prime, there exist positive integers r and s such that

$$r^2 - 27s^2 = -p$$

and

$$r < \frac{5}{2}\sqrt{2p}, \quad s < \frac{1}{2}\sqrt{2p}.$$

As a final comment, we note that if we have two solutions to the equation $|r^2 - 27s^2| = p$, say (r_1, s_1) and (r_2, s_2), then the corresponding forms

$$s_1x^3 + r_1x^2y + 9s_1xy^2 + r_1y^3 \text{ and } s_2x^3 + r_2x^2y + 9s_2xy^2 + r_2y^3$$

are readily seen to be $\mathrm{GL}_2(\mathbb{Z})$-equivalent.

6 Computational Details

The computations required to generate curves of prime conductor p (and subsequently conductor p^2) fall into a small number of distinct parts.

6.1 Generating the Required Forms

To find the irreducible forms potentially corresponding to elliptic curves of prime conductor $p \leq X$ for some fixed positive real X, arguing as in Sect. 4, we generated all reduced forms $F(x, y) = ax^3 + bx^2y + cxy^2 + d$ with discriminants in $(0, 4X]$ and $[-4X, 0)$, separately, by looping over a finite set of a, b, c, d values as prescribed by Lemmata 4.2 and 4.4, respectively. As each form was generated, we checked to see if it actually satisfied the desired definition of reduction. Of course, this does not only produce forms with discriminant $\pm 4p$—as each form was produced, we kept only those whose discriminant was in the appropriate range, and equal to $\pm 4p$ for some prime p. Checking primality was done using the Miller-Rabin primality test (see [31, 40]; to make this deterministic for the range we require, we appeal to [43]). While it is straightforward to code the above in computer algebra packages such as `sage`, `maple` or `magma`, we instead implemented it in `c++` for speed. To avoid possible numerical overflows, we used the `CLN` library [25] for `c++`.

Constructing all the required positive discriminant forms took approximately 40 days of CPU time on a modern server, and about 300 gigabytes of disc space. Thankfully, the computation is easily parallelised and it only took about 1 day of real time. We split the jobs by running a manager which distributed a-values to the other cores. The output from each a-value was stored as a tab-delimited text file with one tuple of p, a, b, c, d on each line.

Generating all forms of negative discriminant took about 3 times longer and required about 900 gigabytes of disc space. The distribution of forms is heavily weighted to small values of a. To allow us to spread the load across many CPUs we actually split the task into 2 parts. We first ran $a \geq 3$, with the master node distributing a-values to the other cores. We then ran $a = 1, 2$ with the master node distributing b-values to the other cores. The total CPU time was about 3 times longer than for the positive case (there being essentially three times as many forms), but more real-time was required due to these complications. Thus generating all forms took less than 1 week of real time but required about 1.2 terabytes of disc space.

We then sorted the forms into discriminant order, while keeping positive and negative discriminant separated. Sorting a terabyte of data is a non-trivial task, and in practice we did this by first sorting[1] the forms for each a-value and then splitting them into files of discriminants in the ranges $[n \times 10^9, (n+1) \times 10^9)$ for $n \in [0, 999]$. Finally, all the files of each discriminant range were sorted together. This process for positive and negative forms took around 2 days of real time. We found 9247369050 forms of positive discriminant and 27938060315 of negative discriminant, with absolute value bounded by 10^{12}. Of these, 475831852 and 828238359, respectively had $F(x, y) = 8$ solvable, leading to 159552514 and 276341397 elliptic curves of positive and negative discriminant, respectively, with prime conductor up to 10^{12}.

6.2 Complete Solution of Thue Equations: Conductor p

For each form encountered, we needed to solve the Thue equation

$$ax^3 + bx^2y + cxy^2 + dy^3 = 8$$

We approached this in two distinct ways.

To solve the Thue equation rigorously, we appealed to by now well-known arguments of Tzanakis and de Weger [46], based upon lower bounds for linear forms in complex logarithms, together with lattice basis reduction; these are implemented in several computer algebra packages, including magma [8] and Pari/GP [37]. The main computational bottleneck in this approach is typically that of computing the fundamental units in the corresponding cubic fields; for computations p of size up to 10^9 or so, we encountered no difficulties with any of the Thue equations arising (in particular, the fundamental units occurring can be certified without reliance upon the Generalized Riemann Hypothesis).

We ran this computation in magma [8], using its built in Thue equation solver. Due to memory consumption issues, we fed the forms into magma in small batches, restarting magma after each set. We saved the output as a tuple

$$p, a, b, c, d, n, \{(x_1, y_1), \ldots, (x_n, y_n)\},$$

[1] Using the standard unix sort command and taking advantage of multiple cores.

where p, a, b, c, d came from the form, n counts the number of solutions of the Thue equation and (x_i, y_i) the solutions. These solutions can then be converted into corresponding elliptic curves in minimal form using Theorem 3.1 and standard techniques.

For positive discriminant, this approach works without issue for $p < 10^{10}$. For negative discriminant, however, the fundamental units in the associated cubic field can be extremely large (in the neighbourhood of $e^{\sqrt{p}}$). For this reason, finding all negative discriminant curves with prime conductor exceeding $2 \cdot 10^9$ or so proves to be extremely slow. Consequently, for large p, we turned to a non-exhaustive method, which, though it finds solutions to the Thue equation, is not actually guaranteed to find them all.

6.3 Non-exhaustive, Heuristic Solution of Thue Equations

If we wish to find all "small" solutions to a Thue equation (which, subject to various well-accepted conjectures, might actually prove to be all solutions), there is an obvious and very quick computational approach we can take, based upon the idea that, given any solution to the equation $F(x, y) = m$ for fixed integer m, we necessarily either have that x and y are small, or that x/y is a convergent in the infinite simple continued fraction expansion to a root of the equation $F(x, 1) = 0$.

Such an approach was developed in detail by Attila Pethő [38, 39]; in particular, he provides a precise and computationally efficient distinction between "large" and "small" solutions. Following this, for each form F under consideration, we expanded the roots of $F(x, 1) = 0$ to high precision, again using the CLN library for c++. We then computed the continued fraction expansion for each real root, along with its associated convergents. Each convergent x/y was then substituted into $F(x, y)$ and checked to see if $F(x, y) = \pm 1, \pm 8$. Replacing (x, y) by one of $(-x, -y)$, $(2x, 2y)$ or $(-2x, -2y)$, if necessary, then provided the required solutions of $F(x, y) = 8$. The precision was chosen so that we could compute convergents x/y with $|x|, |y| \leq 2^{128} \approx 3.4 \times 10^{38}$. We then looked for solutions of small height using a brute force search over a relatively small range of values.

To "solve" $F(x, y) = 8$ by this method, for all forms with discriminant $\pm 4p$ with $p \leq 10^{12}$, took about 1 week of real time using 80 cores. The resulting solutions files (in which we stored also forms with no corresponding solutions) required about 1.5 terabytes of disc space. Again, the files were split into files of absolute discriminant (or more precisely absolute discriminant divided by 4) in the ranges $[n \times 10^9, (n + 1) \times 10^9)$ for $n \in [0, 999]$.

6.4 Conversion to Curves

Once one has a tuple a, b, c, d, x, y, one then computes $G_F(x, y)$ and $H_F(x, y)$, appeals to Theorem 3.1 and checks twists. This leaves us with a list of pairs (c_4, c_6) corresponding to elliptic curves. It is now straightforward to derive a_1, a_2, a_3, a_4, a_6 for a corresponding elliptic curve in minimal form (see e.g. Cremona [15]). For each curve, we saved a tuple $p, a_1, a_2, a_3, a_4, a_6, \pm 1$ with the last entry being the sign of the discriminant of the form used to generate the curve (which coincides with the sign of the discriminant of the curve). We then merged the curves with positive and negative discriminants and added the curves with prime conductor arising from reducible forms (i.e. of small conductor or for primes of the form $t^2 + 64$). After sorting by conductor, this formed a single file of about 17 gigabytes.

6.5 Conductor p^2

The conductor p^2 computation was quite similar, but was split into parts.

6.5.1 Twisting Conductor p

The vast majority of forms of conductor p^2 are quadratic twists of curves of conductor p. To compute these we took all curves with conductor $p \le 10^{10}$ and computed c_4 and c_6. The twisted curve then has corresponding c-invariants

$$c_4' = p^2 c_4 \quad \text{and} \quad c_6' = (-1)^{(p-1)/2} p^3 c_6.$$

The minimal a-invariants were then computed as for curves of conductor p.

We wrote a simple c++ program to read curves of conductor p and then twist them, recompute the a-invariants and output them as a tuple $p^2, a_1, a_2, a_3, a_4, a_6, \pm 1$. The resulting code only took a few minutes to process the approximately 1.1×10^7 curves.

6.5.2 Solving $F(x, y) = 8p$ with F of Discriminant $\pm 4p$

There was no need to find forms for this computation; we reused the positive and negative forms of discriminant $\pm 4p$ with $p \le 10^{10}$ from the conductor-p computations. We subsequently rigorously solved the corresponding equations $F(x, y) = 8p$ for $p \le 10^8$. To solve the Thue equation $F(x, y) = 8p$ for $10^8 < p \le 10^{10}$, using the non-exhaustive, heuristic method, we first converted the equation to a pair of new Thue equations of the form $G_i(x, y) = 8$ as described in Sect. 5.3.1 and then applied Pethő's solution search method.

The solutions were then processed into curves as for the conductor p case above, and the resulting curves were twisted by $\pm p$ in order to search for more curves of conductor p^2.

6.5.3 Solving $F(x, y) \in \{8, 8p\}$ with F of Discriminant $\pm 4p^2$

To find forms of discriminant $4p^2$ with $p \leq 10^{10}$ we need only check to see which primes are of the form $p = r^2 + 27s^2$ in the desired range. To do so, we simply looped over r and s values and then again checked primality using Miller-Rabin. As each prime was found, the corresponding p, r, s tuple was converted to a form as in Sect. 5.3.2, and the Thue equations $F(x, y) = 8$ and $F(x, y) = 8p$ were solved, using the rigorous approach for $p < 10^6$ and the non-exhaustive method described previously for $10^6 < p \leq 10^{10}$. Again, in the latter situation, the equation $F(x, y) = 8p$ was converted to a new equation $G(x, y) = 8$ as described in Sect. 5.3.2. The process for forms of discriminant $-4p^2$ was very similar, excepting that more care is required with the range of r and s. The non-exhaustive method solving both $F(x, y) = 8$ and $F(x, y) = 8p$ for positive and negative forms took a total of approximately 5 days of real time on a smaller server of 20 cores. The rigorous approach, even restricted to prime $p < 10^6$ was much, much slower.

The solutions were then converted to curves as with the previous cases and each resulting curve was twisted by $\pm p$ to search for other curves of conductor p^2.

7 Data

7.1 Previous Work

The principal prior work on computing table of elliptic curves of prime conductor was carried out in two lengthy computations, by Brumer and McGuinness [10] in the late 1980s and by Stein and Watkins [45] slightly more than ten years later. For the first of these computations, the authors fixed the a_1, a_2 and a_3 invariants (12 possibilities) and looped over a_4 and a_6 chosen to make the corresponding discriminant small. By this approach, they were able to find 311243 curves of prime conductor $p < 10^8$ (representing approximately 99.6% of such curves). In the latter case, the authors looped instead over c_4 and c_6, subject to (necessary) local conditions. They obtained a large collection of elliptic curves of general conductor to 10^8, and 11378912 of those with prime conductor to 10^{10} (which we estimate to be slightly in excess of 99.8% of such curves).

7.2 Counts: Conductor p

By way of comparison, we found the following numbers of isomorphism classes of elliptic curves over $\mathbb{Q}$ with prime conductor $p \leq X$:

X	$\Delta_E > 0$	$\Delta_E < 0$	Ratio^2	Total	Expected	Total / Expected
10^3	33	51	2.3884	84	68	1.2353
10^4	129	228	3.1239	357	321	1.1122
10^5	624	1116	3.1986	1740	1669	1.0425
10^6	3388	5912	3.0450	9300	9223	1.0084
10^7	19605	34006	3.0087	53611	52916	1.0131
10^8	114452	198041	2.9941	312493	311587	1.0029
10^9	685278	1187686	3.0038	1872964	1869757	1.0017
10^{10}	4171055	7226982	3.0021	11398037	11383665	1.0013
10^{11}	25661634	44466339	3.0026	70127973	70107401	1.0003
10^{12}	159552514	276341397	2.9997	435893911	435810488	1.0002

The data above the line is rigorous (in case of positive discriminant); for negative discriminant, we have a rigorous result only up to 2×10^9. For the positive forms this took about 1 week of real time using 80 cores. Unfortunately, the negative discriminant forms took significantly longer, roughly 2 months of real times using 80 cores. Heuristics given by Brumer and McGuinness [10] suggest that the number of elliptic curves of negative discriminant of absolute discriminant up to X should be asymptotically $\sqrt{3}$ times as many as those of positive discriminant in the same range—here we report the square of this ratio in the given ranges. The aforementioned heuristic count of Brumer and McGuinness suggests that the expected number of E with prime $N_E \leq X$ should be

$$\frac{\sqrt{3}}{12}\left(\int_1^\infty \frac{1}{\sqrt{u^3-1}}du + \int_{-1}^\infty \frac{1}{\sqrt{u^3+1}}du\right) \mathrm{Li}(X^{5/6}),$$

which we list (after rounding) in the table above. It should not be surprising that this "expected" number of curves appears to slightly undercount the actual number, since it does not take into account the roughly $\sqrt{X}/\log X$ curves of conductor $p = n^2 + 64$ and discriminant $-p^2$ (counting only curves of discriminant $\pm p$).

7.3 Counts: Conductor p^2

To compile the final list of curves of conductor p^2, we combined the five lists of curves: twists of curves of conductor p, curves from forms of discriminant $+4p$ and $-4p$, curves from discriminant $+4p^2$ and $-4p^2$. The list was then sorted and any duplicates removed. The resulting list is approximately 1 gigabyte. The counts of curves are below.

X	$\Delta_E > 0$	$\Delta_E < 0$	Total	Ratio^2
10^3	53	93	146	3.0790
10^4	191	322	513	2.8421
10^5	764	1304	2068	2.9132
10^6	3764	6356	10120	2.8515
10^7	20539	35096	55635	2.9198
10^8	116894	200799	317693	2.9508
10^9	691806	1195262	1887068	2.9851
10^{10}	4189445	7247980	11437425	2.9931

Subsequently we decided that we should recompute the discriminants of these curves as a sanity check, by reading the curves into `sage` and using its built-in elliptic curve routines to compute and then factor the discriminant. This took about 1 day on a single core.

The only curves of real interest are those that do not arise from twisting, i.e. those of discriminant $\pm p^2$, $\pm p^3$ and $\pm p^4$. In the last of these categories, we found only 5 curves, of conductors 11^2, 43^2, 431^2, 433^2 and 33013^2. The first four of these were found by Edixhoven, de Groot and Top [21] (and are of small enough conductor to now appear in Cremona's tables). The fifth, satisfying

$$(a_1, a_2, a_3, a_4, a_6) = (1, -1, 1, -1294206576, 17920963598714),$$

has discriminant 33013^4. For discriminants $\pm p^2$ and $\pm p^3$, we found the following numbers of curves, for conductors $p \le X$:

X	$\Delta_E = -p^2$	$\Delta_E = p^2$	$\Delta_E = -p^3$	$\Delta_E = p^3$
10^3	12	4	7	4
10^4	36	24	9	5
10^5	80	58	12	9
10^6	203	170	17	15
10^7	519	441	24	23
10^8	1345	1182	32	36
10^9	3738	3203	48	58
10^{10}	10437	9106	60	86

It is perhaps worth observing that the majority of these curves arise from, in the case of discriminant $\pm p^2$, forms with, in the notation of Sects. 5.3.2 and 5.3.3, either r or s in $\{1, 8\}$. Similarly, for $\Delta_E = \pm p^3$, most of the curves we found come from forms in the eight one-parameter families described in Sect. 5.3.1.

7.4 Thue Equations

It is worth noting that all solutions we encountered to the Thue equations $F(x,y) = 8$ and $F(x,y) = 8p$ we treated were with $|x|, |y| < 2^{30}$. The "largest" such solution corresponded to the equation

$$355x^3 + 293x^2y - 1310xy^2 - 292y^3 = 8,$$

with solution

$$(x,y) = (188455233, -82526573).$$

This leads to the elliptic curve of conductor 948762329069,

$$y^2 + xy + y = x^2 - 2x^2 + a_4x + a_6,$$

with

$$a_4 = -11977910249344808133 41$$

and

$$a_6 = 15955840837175565243579564368641.$$

In the following table, we collect data on the number of $\mathrm{GL}_2(\mathbb{Z})$-equivalence classes of irreducible binary cubic forms of discriminant $4p$ or $-4p$ for p in $[0,X]$, denoted $P_3(0,X)$ and $P_3(-X,0)$, respectively. We also provide counts for those forms where the corresponding equation $F(x,y) = 8$ has at least one integer solution, denoted $P_3^*(0,X)$ and $P_3^*(-X,0)$ for positive and negative discriminant forms, respectively.

X	$P_3(0,X)$	$P_3^*(0,X)$	$P_3(-X,0)$	$P_3^*(-X,0)$
10^3	23	22	78	61
10^4	204	163	740	453
10^5	1851	1159	6104	2641
10^6	16333	7668	53202	16079
10^7	147653	49866	466601	97074
10^8	1330934	314722	4126541	582792
10^9	12050910	1966105	36979557	3530820
10^{10}	109730653	12229663	334260481	21576585
10^{11}	1004607003	76122366	3045402451	133115651
10^{12}	9247369050	475831852	27938060315	828238359

Our expectation is that the number of forms for which the equation $F(x, y) = 8$ has solutions with absolute discriminant up to X is $o(X)$ (i.e. this occurs for essentially zero percent of forms).

7.5 Elliptic Curves with the Same Prime Conductor

One might ask how many isomorphism classes of curves of a given prime conductor can occur. If one believes new heuristics that predict that the Mordell-Weil rank of $E/\mathbb{Q}$ is absolutely bounded, then this number should also be so bounded. As noted by Brumer and Silverman [11], there are 13 curves of conductor 61263451. Up to $p < 10^{12}$, the largest number we encountered was for $p = 530956036043$, with 20 isogeny classes, corresponding to $[a_1, a_2, a_3, a_4, a_6]$ as follows :

$$\begin{aligned}
&[0,-1,1,-1003,37465]\,,[0,-1,1,-1775,45957]\,,\\
&[0,-1,1,-38939,2970729]\,,[0,-1,1,-659,-35439]\,,\\
&[0,-1,1,2011,4311]\,,[0,-2,1,-27597,-1746656]\,,\\
&[0,-2,1,57,35020]\,,[1,-1,0,-13337473,18751485796]\,,\\
&[0,0,1,-13921,633170]\,,[0,0,1,-30292,-2029574]\,,\\
&[0,0,1,-6721,-214958]\,,[0,0,1,-845710,-299350726]\,,\\
&[0,0,1,-86411851,309177638530]\,,[0,0,1,-10717,428466]\,,\\
&[1,-1,0,-5632177,5146137924]\,,[1,-1,0,878,33379]\,,\\
&[1,-1,1,1080,32014]\,,[1,-2,1,-8117,-278943]\,,\\
&[1,-3,0,-2879,71732]\,,[1,-3,0,-30415,-2014316]\,.
\end{aligned}$$

Of these 20 curves, 2 have rank 3, 3 have rank 2, 9 have rank 1 and 6 have rank 0. All have discriminant $-p$. The class group of $\mathbb{Q}(\sqrt{3 \cdot 530956036043})$ is isomorphic to

$$\mathbb{Z}/3\mathbb{Z} \oplus \mathbb{Z}/3\mathbb{Z} \oplus \mathbb{Z}/3\mathbb{Z},$$

which, via a classical result of Hasse [26], explains the existence of a large number of cubic forms of discriminant $-4p$. Elkies [22] found examples of rather larger conductor with more curves, including 21 for $p = 14425386253757$ and discriminant p, 24 for $p = 998820191314747$ and discriminant $-p$.

7.6 Rank and Discriminant Records

In the following table, we list the smallest prime conductor with a given Mordell-Weil rank. These were computed by running through our data, using Rubinstein's upper bounds for analytic ranks (as implemented in Sage) to search for candidate curves of "large" rank which were then checked using mwrank.

N	$[a_1,a_2,a_3,a_4,a_6]$	$\operatorname{sign}(\Delta_E)$	$rk(E(\mathbb{Q}))$
37	$[0,0,1,-1,0]$	$+$	1
389	$[0,1,1,-2,0]$	$+$	2
5077	$[0,0,1,-7,6]$	$+$	3
501029	$[0,1,1,-72,210]$	$+$	4
19047851	$[0,0,1,-79,342]$	$-$	5
6756532597	$[0,0,1,-547,-2934]$	$+$	6

It is perhaps noteworthy that the curve listed here of rank 6 has the smallest known minimal discriminant for such a curve (see Table 4 of Elkies and Watkins [23]).

If we are interested in similar records over all curves, including composite conductors, we have

N	$[a_1,a_2,a_3,a_4,a_6]$	$\operatorname{sign}(\Delta_E)$	$rk(E(\mathbb{Q}))$
37	$[0,0,1,-1,0]$	$+$	1
389	$[0,1,1,-2,0]$	$+$	2
5077	$[0,0,1,-7,6]$	$+$	3
234446	$[1,-1,0,-79,289]$	$+$	4
19047851	$[0,0,1,-79,342]$	$-$	5
5187563742	$[1,1,0,-2582,48720]$	$+$	6
382623908456	$[0,0,0,-10012,346900]$	$+$	7

Here, the curves listed above the line are proven to be those of smallest conductor with the given rank. Those listed below the line have the smallest known conductor for the corresponding rank.

References

1. M. K. Agrawal, J. H. Coates, D. C. Hunt and A. J. van der Poorten, *Elliptic curves of conductor* 11, Math. Comp. 35 (1980), 991–1002.
2. K. Belabas. *A fast algorithm to compute cubic fields,* Math. Comp. 66 (1997), 1213–1237.
3. K. Belabas and H. Cohen, *Binary cubic forms and cubic number fields*, Organic Mathematics (Burnaby, BC, 1995), 175–204. CMS Conf. Proc., 20 Amer. Math. Soc. 1997.
4. M. A Bennett and A. Ghadermarzi, *Mordell's equation : a classical approach*, L.M.S. J. Comput. Math. 18 (2015), 633–646.
5. M. A. Bennett and A. Rechnitzer, *Computing elliptic curves over* $\mathbb{Q}$, submitted for publication.
6. W. E. H. Berwick and G. B. Mathews, *On the reduction of arithmetical binary cubic forms which have a negative determinant*, Proc. London Math. Soc. (2) 10 (1911), 43–53.
7. B. J. Birch and W. Kuyk (Eds.), *Modular Functions of One Variable IV*, Lecture Notes in Math., vol. 476, Springer-Verlag, Berlin and New York, 1975.
8. W. Bosma, J. Cannon, and C. Playoust. The Magma algebra system. I. The user language, *J. Symbolic Comput.*, 24 (1997), 235–265. Computational algebra and number theory (London, 1993).

9. C. Breuil, B. Conrad, F. Diamond and R. Taylor, *On the Modularity of Elliptic Curves over* $\mathbb{Q}$ *: Wild 3-adic Exercises*, J. Amer. Math. Soc. 14 (2001), 843–939.
10. A. Brumer and O. McGuinness, *The behaviour of the Mordell-Weil group of elliptic curves*, Bull. Amer. Math. Soc. 23 (1990), 375–382.
11. A. Brumer and J. H. Silverman, *The number of elliptic curves over* $\mathbb{Q}$ *with conductor* N, Manuscripta Math. 91 (1996), 95–102.
12. J. Coates, *An effective p-adic analogue of a theorem of Thue. III. The diophantine equation* $y^2 = x^3 + k$, Acta Arith. 16 (1969/1970), 425–435.
13. F. Coghlan, *Elliptic Curves with Conductor* 2^m3^n, Ph.D. thesis, Manchester, England, 1967.
14. J. Cremona, *Elliptic curve tables*, http://johncremona.github.io/ecdata/
15. J. Cremona, *Algorithms for modular elliptic curves*, second ed., Cambridge University Press, Cambridge, 1997. Available online at http://homepages.warwick.ac.uk/staff/J.E.Cremona/book/fulltext/index.html
16. J. Cremona, *Reduction of binary cubic and quartic forms,* LMS J. Comput. Math. 4 (1999), 64–94.
17. J. Cremona and M. Lingham, *Finding all elliptic curves with good reduction outside a given set of primes*, Experiment. Math. 16 (2007), 303–312.
18. H. Davenport, *The reduction of a binary cubic form. I.*, J. London Math. Soc. 20 (1945), 14–22.
19. H. Davenport, *The reduction of a binary cubic form. II.*, J. London Math. Soc. 20 (1945), 139–147.
20. H. Davenport and H. Heilbronn, *On the density of discriminants of cubic fields. II.*, Proc. Roy. Soc. London Ser. A. 322 (1971), 405–420.
21. B. Edixhoven, A. de Groot and J. Top, *Elliptic curves over the rationals with bad reduction at only one prime*, Math. Comp. 54 (1990), 413–419.
22. N. D. Elkies, *How many elliptic curves can have the same prime conductor?*, http://math.harvard.edu/~elkies/condp_banff.pdf
23. N. D. Elkies, and M. Watkins, *Elliptic curves of large rank and small conductor*, Algorithmic number theory, 42–56, Lecture Notes in Comput. Sci., 3076, Springer, Berlin, 2004.
24. T. Hadano, *On the conductor of an elliptic curve with a rational point of order* 2, Nagoya Math. J. 53 (1974), 199–210.
25. B. Haible, *CLN, a class library for numbers*, available from http://www.ginac.de/CLN/
26. H. Hasse, *Arithmetische Theorie der kubischen Zahlköper auf klassenkörpertheoretischer Grundlage*, Math. Z. 31 (1930), 565–582.
27. C. Hermite, *Note sur la réduction des formes homogènes à coefficients entiers et à deux indétermineées*, J. reine Angew. Math. 36 (1848), 357–364.
28. C. Hermite, *Sur la réduction des formes cubiques à deux indéxtermineées*, C. R. Acad. Sci. Paris 48 (1859), 351–357.
29. G. Julia, *Étude sur les formes binaires non quadratiques à indétermindés rélles ou complexes*, Mem. Acad. Sci. l'Inst. France 55 (1917), 1–293.
30. J.-F. Mestre and J. Oesterlé. *Courbes de Weil semi-stables de discriminant une puissancemième*, J. reine angew. Math 400 (1989), 173–184.
31. G. L. Miller, *Riemann's hypothesis and tests for primality* in Proceedings of seventh annual ACM symposium on Theory of computing, 234–239 (1975).
32. L. J. Mordell, *The diophantine equation* $y^2 - k = x^3$, Proc. London. Math. Soc. (2) 13 (1913), 60–80.
33. L. J. Mordell, *Diophantine Equations,* Academic Press, London, 1969.
34. T. Nagell, *Introduction to Number Theory*, New York, 1951.
35. O. Neumann, *Elliptische Kurven mit vorgeschriebenem Reduktionsverhalten II*, Math. Nach. 56 (1973), 269–280.
36. I. Papadopolous, *Sur la classification de Néron des courbes elliptiques en caractéristique résseulé* 2 et 3, J. Number Th. 44 (1993), 119–152.
37. The PARI Group, Bordeaux. *PARI/GP version* `2.7.1`, 2014. available at http://pari.math.u-bordeaux.fr/.
38. A. Pethő, *On the resolution of Thue inequalities*, J. Symbolic Computation 4 (1987), 103–109.

39. A. Pethő, *On the representation of* 1 *by binary cubic forms of positive discriminant*, Number Theory, Ulm 1987 (Springer LNM 1380), 185–196.
40. M. O. Rabin, *Probabilistic algorithm for testing primality*, J. Number Th. 12 (1980) 128–138.
41. B. Setzer, *Elliptic curves of prime conductor*, J. London Math. Soc. 10 (1975), 367–378.
42. I. R. Shafarevich, *Algebraic number theory*, Proc. Internat. Congr. Mathematicians, Stockholm, Inst. Mittag-Leffler, Djursholm (1962), 163–176.
43. J. P. Sorenson and J. Webster, *Strong Pseudoprimes to Twelve Prime Bases*, arXiv preprint arXiv:1509.00864.
44. V. G. Sprindzuk, *Classical Diophantine Equations*, Springer-Verlag, Berlin, 1993.
45. W. Stein and M. Watkins, *A database of elliptic curve – first report*, Algorithmic Number Theory (Sydney, 2002), Lecture Notes in Compute. Sci., vol. 2369, Springer, Berlin, 2002, pp. 267–275.
46. N. Tzanakis and B. M. M. de Weger, *On the practical solutions of the Thue equation*, J. Number Theory 31 (1989), 99–132.
47. N. Tzanakis and B. M. M. de Weger, *Solving a specific Thue-Mahler equation*, Math. Comp. 57 (1991) 799–815.
48. N. Tzanakis and B. M. M. de Weger, *How to explicitly solve a Thue-Mahler equation*, Compositio Math., 84 (1992), 223–288.
49. B. M. M. de Weger, *Algorithms for diophantine equations*, CWI-Tract No. 65, Centre for Mathematics and Computer Science, Amsterdam, 1989.
50. B. M. M. de Weger, *The weighted sum of two S-units being a square*, Indag. Mathem. 1 (1990), 243–262.

Part IX
Sustainability and Cooperation

Sustainability of Cooperation in Dynamic Games Played over Event Trees

Georges Zaccour

Abstract In this tutorial, we recall the main ingredients of the theory of dynamic games played over event trees and show step-by-step how to build a sustainable cooperative solution.

Keywords Dynamic games • Cooperation • Sustainability

1 Introduction

Many problems in economics, engineering and management science have the following three features in common: (a) They involve only a few agents (players), which have interdependent payoffs, that is, the action of any player affects the payoffs of all. (b) The agents cooperate or compete repeatedly over time, and the problem involves an accumulation process, e.g., production capacity, pollution stock. (c) Some of the parameter values are uncertain. A natural framework to deal with such problems is the theory of dynamic games played over event trees (DGPET). As an illustration of such a setting, consider a region served by a few electricity producers (players) who compete in one or more market segments (peak-load, local market, export market, etc.). At each period, the price in each segment depends on the total available supply and on the realization of some random events (e.g., weather conditions or the state of the economy). Further, producers invest in different production capacities (nuclear, thermal, hydro, etc.) over time. In the terminology of dynamic games, the quantities committed to each market segment, which are constrained by available capacity, and the investments in different production technologies are the player's control variables and the installed production capacities are the state variables. The players must account for uncertainty in demand when they make their decisions.

Now, suppose that the players (firms, countries, individuals) involved in an example of DGPET agree to cooperate, that is, to coordinate their strategies in order

G. Zaccour (✉)
Chair in Game Theory and Management, GERAD, HEC Montréal, Montréal, QC, Canada
e-mail: georges.zaccour@gerad.ca

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2_14

to maximize their joint payoff over a given time interval $[0, T]$. A legitimate question is then how to ensure that each player will indeed fulfill her part of the agreement over time? This is the question we deal with in this paper.

It is useful from the outset to make some clarifying observations regarding the nature of the problem at hand. First, although it may be appealing to favor short-term agreements to keep all options open, long-term commitments cannot be avoided when the contracting cost is high. For instance, it is unthinkable that the government and the civil service union meet every Monday to negotiate that week's employment conditions. Common sense clearly suggests that both parties should avoid costly and time-consuming negotiations and agree on a collective labor agreement that will remain in place for a number of years.

Second, it is an empirical fact that some long-term agreements are abandoned before their maturity. A drastic illustration of this is the high level of divorce observed around the globe. Haurie [19] cites two reasons why an agreement (contract or cooperative solution), which suits everyone at an initial instant of time may not reach its maturity date T: (i) If the players agree to renegotiate the original agreement at time $\tau \in (0, T]$, it is not certain that they will all want to continue with that agreement. In fact, they will not go on with the original agreement if it is not a solution of the cooperative game that starts out at time τ. (ii) If a player obtains a higher payoff by leaving the agreement at time $\tau \in (0, T]$ than by continuing to implement her cooperative strategies, then she will indeed deviate from cooperation. In the parlance of dynamic optimization and dynamic games, such a breakdown means that the agreement is time inconsistent. It is important to mention here that if the cooperative agreement is an equilibrium, then item (ii) above cannot occur because no player would, by definition, find it optimal to deviate from the solution. It is well-known that, except in games having very special structures (see, e.g., Chiarella et al. [6] and Martín-Herrán and Rincón-Zapatero [32]), a Pareto-optimal (or cooperative) solution is not an equilibrium.

The rest of the paper is organized as follows: In Sect. 2, we give a brief account of the literature dealing with the sustainability of cooperation in dynamic games. In Sect. 3, we recall the main ingredients of dynamic games played over event trees. We explain the approach to achieve a node-consistent outcome in DGPET in Sect. 4. In Sect. 5, we briefly conclude.

2 Brief Literature Review

The literature in dynamic games has followed two streams in its quest of sustain cooperation over time, namely, building cooperative equilibria or defining time-consistent solutions.

Through the implementation of some (punishing) strategies, the first stream seeks to make the cooperative solution an equilibrium of an associated noncooperative game. If this is achieved, then the result will be at once collectively optimal and stable, as no player will find it optimal to deviate unilaterally from the equilibrium.

To build a cooperative equilibrium, players can for instance implement trigger strategies, which are strategies based on the history of the game. Loosely speaking, such strategies are defined as follows: At any decision node, if the history of the game has been till now cooperative, then each player will implement the cooperative action; otherwise, which means that a player has cheated, then all the other players implement their punishing strategies, which are set out in a pre-play arrangement. Intuitively, for such punishing strategies to work, they must be: (i) effective, that is, the deviator would lose from cheating on the agreement, and (ii) credible, that is, it is in the best interest of the other players to implement their punishing strategies if a deviation is observed, rather than sticking to cooperation.

Sustaining a Pareto outcome as an equilibrium has a long history in repeated games, and a well-known result in this area is the so-called folk theorem, which (informally) states that if the players are sufficiently patient, then any Pareto-optimal outcome can be achieved as a Nash equilibrium; see, e.g., Osborne and Rubinstein [37]. A similar theorem has been proved for stochastic games by Dutta [10]. Trigger strategies have also been considered in multistage games and in differential games; see the early contributions by Tolwinski et al. [54], Haurie and Pohjola [20] and Haurie et al. [21]. The books by Dockner et al. [7] and Haurie et al. [22] provide a comprehensive introduction to cooperative equilibria in differential games.

Having the same objective of embedding the cooperative solution with an equilibrium property, Ehtamo and Hämäläinen [11–13] proposed the concept of incentive strategies and a corresponding equilibrium in two-player differential games. A player's incentive strategy is a function of the other player's action. In an incentive equilibrium, each player implements her part of the agreement if the other player also does. In terms of computation, the determination of an incentive equilibrium requires solving a pair of optimal-control problems, which is in general relatively easy to do. A main concern with incentive strategies is their credibility, since it may happen that the best response to a deviation from cooperation is to stick to cooperation rather than to also deviating. In such a situation, the threat of punishment for a deviation is an empty one. In applications, one can derive the conditions that the parameter values must satisfy to have credible incentive strategies. For a discussion of the credibility of incentive strategies in differential games with special structures, see Martín-Herrán and Zaccour [34, 35]. A further drawback of incentive equilibrium is that the concept is defined for only two players. Incentive strategies and equilibria have been applied in a number of areas, including environmental economics (see, e.g., Breton et al. [3], de Frutos and Martín-Herrán [8]), marketing (see, e.g., Martín-Herrán and Taboubi [33], Buratto and Zaccour [4]) and in closed-loop supply chains (De Giovanni et al. [9]).

In the second stream, to which this contribution belongs, the idea is to define a *time-consistent* decomposition over time of the total cooperative payoff (allocation) of player $j, j \in M$, over the planning horizon $[0, T]$. An allocation is time consistent if at any intermediate instant of time the cooperative payoff-to-go dominates (at least weakly) the noncooperative payoff-to-go for all players. It is important to mention that the inequality is verified along the cooperative state trajectory, which means that cooperation has prevailed up to the time of comparison. A stronger condition is used in the concept of *agreeability*, where the above payoff dominance must hold along any feasible state trajectory (see Kaitala and Pohjola [28, 29] and Jørgensen et al. [25, 26]). The literature on time consistency in cooperative dynamic games has essentially been in continuous time. The concept was initially proposed in Petrosjan [40] and Petrosjan and Danilov [43–45]. In these publications in Russian, as well as in subsequent books in English (Petrosjan [41], Petrosjan and Zenkevich [47]), and in Petrosjan [40], time consistency was termed dynamic stability. In Yeung and Petrosjan [57] a *proportional* time-consistent solution was investigated, whereas Petrosjan and Zaccour [46] proposed a time-consistent *Shapley value*. Jørgensen and Zaccour [27] and Yeung and Petrosjan [60] derived time-consistent solutions in environmental and joint-venture games, respectively. Yeung and Petrosjan [58, 59] and Yeung et al. [61] studied time consistency in stochastic differential games. For a general discussion of time consistency in differential games, see the book by Yeung and Petrosjan [59] and the survey by Zaccour [63].

Other papers discussed time-consistent solutions (or very close concepts) for deterministic or stochastic discrete-dynamic games; see, e.g., Chandler and Tulkens [5], Filar and Petrosjan [14], Germain et al. [17], Petrosjan et al. [42], Predtetchinski [51], Lehrer and Scarsini [31] and Xu and Veinott [56]. Finally, Avrachenkov et al. [2] established conditions for time consistency for cooperative Markov decision processes.

3 Games Played over Event Trees

In this section, we recall the main elements of DGPET. This class of games was introduced by Zaccour [62] and Haurie et al. [23], and further developed in Haurie and Zaccour [24]. The initial motivation was an analysis of the European natural gas market, and more specifically, the forecasting of long-term deliveries of gas from four producers (Algeria, Netherlands, Norway and the former USSR) to nine consuming European regions. The deliveries and investments are the control variables, and production capacities and reserves of gas are the state variables. Each consuming region is described by a time-varying demand function whose parameter values are uncertain, with the stochasticity represented by an event tree. This is a situation where the three features mentioned in the introduction, that is, strategic interaction, dynamic, and uncertainty, are clearly present. More recently, the class of DGPET has been applied to electricity markets in, e.g., Pineau and Murto [48], Genc et al. [15], Genc and Sen [16] and Pineau et al. [50]. Here, the main objective is to predict equilibrium investments in different generation technologies in deregulated electricity markets. Parilina and Zaccour [39] constructed an ε-cooperative equilibrium for this class of games and illustrated their results using a linear-quadratic game in environmental economics. For a comprehensive introduction to the class of DGPET, see Haurie et al. [22].

Let $\mathscr{T} = \{0, 1, \ldots, T\}$ be the set periods, and denote by $(\xi(t) : t \in \mathscr{T})$ the exogenous stochastic process represented by an event tree, with a root node n^0 in period 0 and a set of nodes $\mathscr{N}^t$ in period $t = 0, 1, \ldots, T$. Let $a(n^t) \in \mathscr{N}^{t-1}$ be the unique predecessor of node $n^t \in \mathscr{N}^t$ for $t = 0, 1, \ldots, T$, and denote by $S(n^t) \in \mathscr{N}^{t+1}$ the set of all possible direct successors of node $n^t \in \mathscr{N}^t$ for $t = 0, 1, \ldots, T-1$. We call *scenario* any path from node n^0 to a terminal node n^T. Each scenario has a probability, and the probabilities of all scenarios sum up to 1. We denote by π^{n^t} the probability of passing through node n^t, which corresponds to the sum of the probabilities of all scenarios that contain this node. In particular, $\pi^{n^0} = 1$, and π^{n^T} is equal to the probability of the single scenario that terminates in (leaf) node $n^T \in \mathscr{N}^T$. Also, $\sum_{n^t \in \mathscr{N}^t} \pi^{n^t} = 1, \forall t$.

Denote by $M = \{1, \ldots, m\}$ the set of players. Denote by $u_j(n_l^t) \in \mathbb{R}^{m_j}$ the decision variables of player j at node n_l^t, and let $\underline{u}(n_l^t) = \left(u_1(n_l^t), \ldots, u_m(n_l^t)\right)$. Let $X \subset \mathbb{R}^p$, with p a given positive integer, be a state set. For each node $n_l^t \in \mathscr{N}^t$, $t = 0, 1, \ldots, T$, let $U_j^{n_l^t} \subset \mathbb{R}^{\mu_j^{n_l^t}}$, with $\mu_j^{n_l^t}$ a given positive integer, be the control set of player j. Denote by $\underline{U}^{n_l^t} = U_1^{n_l^t} \times \cdots \times U_j^{n_l^t} \times \cdots \times U_m^{n_l^t}$ the product control sets. A transition function$f^{n_l^t}(\cdot,\cdot) : X \times \underline{U}^{n_l^t} \mapsto X$ is associated with each node n_l^t. The state equations are given as

$$x(n_l^t) = f^{a(n_l^t)}\left(x\left(a\left(n_l^t\right)\right), \underline{u}\left(a\left(n_l^t\right)\right)\right), \tag{1}$$

$$\underline{u}\left(a\left(n_l^t\right)\right) \in \underline{U}^{a(n_l^t)}, \quad n_l^t \in \mathscr{N}^t, t = 1, \ldots, T. \tag{2}$$

At each node n_l^t, $t = 0, \ldots, T-1$, the reward to player j is a function of the state and of the controls of all players, given by $\phi_j^{n_l^t}(x(n_l^t), \underline{u}(n_l^t))$. At a terminal node n_l^T, the reward to player j is given by the function $\Phi_j^{n_l^T}(x(n_l^T))$.

We assume that player $j \in M$ maximizes her expected stream of payoffs. The state equations and the reward functions define the following multistage game, where we let

$$\tilde{x} = \{x(n_l^t) : n_l^t \in \mathscr{N}^t, t = 0, \ldots, T\},$$
$$\tilde{\underline{u}} = \{\underline{u}(n_l^t) : n_l^t \in \mathscr{N}^t, t = 0, \ldots, T-1\},$$

and $J_j(\tilde{x}, \tilde{\underline{u}})$ be the payoff to player j, that is,

$$J_j(\tilde{x}, \tilde{\underline{u}}) = \sum_{t=0}^{T-1} \sum_{n_l^t \in \mathscr{N}^t} \pi(n_l^t)\phi_j^{n_l^t}(x(n_l^t), \underline{u}(n_l^t))$$
$$+ \sum_{n_l^t \in \mathscr{N}_T} \pi(n_l^T)\Phi_j^{n_l^T}(x(n_l^T)), \quad j \in M, \tag{3}$$

s.t.

$$x(n_l^t) = f^{a(n_l^t)}(x(a(n_l^t)), \underline{u}(a(n_l^t))), \tag{4}$$

$$\underline{u}(a(n_l^t)) \in \underline{U}^{a(n_l^t)}, \quad n_l^t \in \mathscr{N}^t, t = 1, \ldots, T,$$

$$x(n_0) = x^0 \text{given}. \tag{5}$$

Remark 3.1 As we are dealing with a finite horizon, we do not discount future payoffs. Adding discounting would not cause any conceptual difficulty.

Remark 3.2 The DGPET framework can take into account more complicated constraints on the control variables than the ones considered here, e.g., constraints with lags and coupled constraints (see Kanani Kuchesfehani and Zaccour [30]).

As alluded to before, dealing with long-term cooperation involves at intermediate instants of time, a comparison of noncooperative and cooperative payoffs-to-go.

3.1 Noncooperative and Cooperative Outcomes

In DGPET, the control and state variables are node dependent, and each node $n^t \in \mathscr{N}^t$ represents a possible sample value of the history h^t of the ξ (.) process up to time t. Because of this, a strategy in DGPET is referred to as S-adapted strategy, where the S stands for sample.

Definition 3.1 An admissible S-adapted strategy for player j is a vector $\tilde{u}_j = \{u_j(n_l^t) : n_l^t \in \mathscr{N}^t, t = 0, \ldots, T-1\}$, that is, a plan of actions adapted to the history of the random process represented by the event tree.

Denote by $\underline{\tilde{u}} = (\tilde{u}_j : j \in M)$ the S-adapted strategy vector of the m players. We can thus define a game in normal form,[1] with payoffs $W_j(\underline{\tilde{u}}, x^0) = J_j(\tilde{x}, \underline{\tilde{u}}), j \in M$, where $\tilde{x}$ is obtained from $\underline{\tilde{u}}$ as the unique solution of the state equations that emanate from the initial state x^0.

If the game is played noncooperatively, then the players will seek a Nash equilibrium in S-adapted strategies defined as follows:

Definition 3.2 An S-adapted Nash equilibrium is an admissible S-adapted strategy $\underline{\tilde{u}}^N$ such that for every player j the following holds:

$$W_j(\underline{\tilde{u}}^N, x^0) \geq W_j([\tilde{u}_j, \tilde{\mathbf{u}}_{-j}^N], x^0),$$

where $\tilde{\mathbf{u}}_{-j}^N$ is the Nash equilibrium policy vector of all players $i \neq j$.

We make the following remarks.

[1] To define a game in normal form, we need three elements: (a) a finite set of players $M = \{1, \ldots, m\}$, (b) a strategy set S_i of player $i \in M$, and (c) a payoff function $\pi_i : \prod_{i \in M} S_i \to \mathbb{R}$.

Remark 3.3 Although the S-adapted and open-loop equilibria look similar, they differ in the definitions of the state equations and control variables. In an open-loop information structure, the control variables and the state equations are defined over time. Here, as mentioned above, they are defined (indexed) over the set of nodes of the event tree.

Remark 3.4 As a DGPET has a normal-form representation, the conditions for existence and uniqueness of a Nash equilibrium are the same as in classical games with continuous payoffs with constraints as established in Rosen [53].[2]

If the players agree to cooperate, then they will optimize the sum of their payoffs throughout the entire horizon,[3] that is,

$$\max_{\tilde{u}_j, j\in M} W = \sum_{j\in M} W_j \left(\underline{\tilde{u}}, x^0\right).$$

Denote by $\underline{\tilde{u}}^* \left(x^0\right)$ the resulting vector of cooperative controls, i.e.,

$$\underline{\tilde{u}}^* \left(x^0\right) = \arg\max \sum_{j\in M} W_j \left(\underline{\tilde{u}}, x^0\right).$$

Remark 3.5 The vector $\underline{\tilde{u}}^* \left(x^0\right)$ corresponds to the agreement signed by all players at initial date. This is the vector that we would like to see it implemented throughout the duration of the game.

Denote by $\tilde{x}^* = \left\{x^*(n_l^t) : n_l^t \in \mathscr{N}^t,\ t = 0, 1, \ldots, T\right\}$ the cooperative state trajectory generated by $\underline{\tilde{u}}^* \left(x^0\right)$.

4 Node Consistency

Informally speaking, a cooperative solution in DGPET is node consistent, if the cooperative payoff-to-go of player $j, j \in M$, in the subgame starting at any node is at least equal to the noncooperative payoff-to-go in this subgame. We reiterate that this comparison takes place along the cooperative state trajectory, meaning that at node of comparison n_l^t,$n_l^t \in \mathscr{N}^t, t = 1, \ldots, T$, the state value is $\tilde{x}^* \left(n_l^t\right)$. If all players implement the prescribed actions by joint maximization, then they will collectively obtain the following outcome:

[2] For a detailed treatment in the context of this class of games, see Haurie et al. [22].

[3] We can easily extend our framework to the case where the players maximize a weighted sum of payoffs.

$$W^* = \sum_{j \in M} W_j \left(\underline{\tilde{u}}^* \left(x^0 \right) \right).$$

Two questions remain unresolved:

1. How can W^* be divided among the players? Note that $W_j \left(\underline{\tilde{u}}^* \left(x^0 \right) \right)$ is the before side-payment payoff of player j and not what she will actually obtain after side payments have been made.[4]
2. How do we design a node-consistent agreement? That is, how is it possible to allocate each player's after side-payment payoff over nodes such that all players stick to the agreement as time goes by?

In order to address these issues, we need to implement the following steps:

1. Define a cooperative game and compute all characteristic function values.
2. Choose a solution concept. This amounts at selecting an imputation, that is, a vector whose entries correspond to after-side-payment outcomes of the players.
3. Compute for each node of the event tree the cooperative and noncooperative payoffs-to-go.
4. Define an imputation distribution procedure (IDP) that is node consistent.

4.1 Defining the Cooperative Game

A cooperative game is a triplet (M, v, Y), where M is the set of players; v is the characteristic function that assigns to each coalition $G, G \subseteq M$, a numerical value,

$$v(G) : P(M) \to \mathbb{R}, \quad v(\varnothing) = 0,$$

where $P(M)$ is the power set of M; and Y is the set of imputations, that is,

$$Y = \left\{ (y_1, \ldots, y_m) \text{ such that } y_j \geq v(\{j\}) \text{ and } \sum_{j \in M} y_j = v(M) \right\}.$$

The characteristic function measures the power or the strength of a coalition. Its precise definition depends on the assumption made about what the left-out players—that is, the complement subset of players $M \backslash G$—will do (see, e.g., Ordeshook [36] and Osborne and Rubinstein [37]). In their seminal book, von Neumann and Morgenstern [55] interpreted $v(G)$ as the largest joint payoff that a coalition G can guarantee its members. In the absence of externalities, i.e., if the payoffs to the members of a coalition G is independent of the actions of the non-members $(M \backslash G)$,

[4] The implicit assumption here is that players' utilities (gains) are comparable and transferable; otherwise side payments do not make sense.

then $v(G)$ would be the result of an optimization problem. However, in the presence of externalities, a prediction of the actions of the non-members of G plays a central role in the computation of the worth of a coalition. This aspect has led to different definitions of a characteristic function (see Aumann [1] and Chander and Tulkens [5]). Note that the developments to come are valid for any choice of $v(\cdot)$.

The definition of the set of imputations involves two conditions, namely, individual rationality ($y_j \geq v(\{j\})$) and collective rationality $\left(\sum_{j\in M} y_j = v(M)\right)$. Individual rationality means that no player will accept an allocation or imputation that gives her less than what she can secure by acting alone. Collective rationality means that the total collective gain should be allocated, that is, no deficit or subsidies are considered. To make the connection with what was said earlier, observe that $v(M) = W^* = \sum_{j\in M} W_j\left(\underline{\tilde{u}}^*\left(x^0\right)\right)$, and that player j will get some y_j, which is still to be decided (in the next step) and which will not necessarily be equal to $W_j\left(\underline{\tilde{u}}^*\left(x^0\right)\right)$.

4.2 Selecting Imputations

Game theorists have proposed many solutions for sharing the total cooperative gain among the players. These solutions are typically based on a series of axioms or requirements that the allocation(s) must satisfy, e.g., fairness, stability. We distinguish between solution concepts that select a unique imputation in Y, e.g., Shapley value and the nucleolus, and those that select a subset of imputations, e.g., the core and stable set. The two most used solution concepts in applications of cooperative games are the Shapley value and the core. We will use them to illustrate the process of building a node-consistent cooperative solution.

Definition 4.3 The Shapley value is an imputation $\sigma = (\sigma_1, \ldots \sigma_m)$ defined by

$$\sigma_j = \sum_{\substack{G\subset M \\ j\in G}} \frac{(m-g)!(g-1)!}{m!}[v(G) - v(G\backslash\{j\})]. \tag{6}$$

Being an imputation, the Shapley value satisfies individual rationality, i.e., $\sigma_j \geq v(\{j\})$ for all $j \in M$. The term $[v(G) - v(G\backslash\{j\})]$ corresponds to the marginal contribution of player j to coalition G. Thus, the Shapley value allocates to each player the weighted sum of her marginal contributions to all coalitions that she may join. The Shapley value is the unique imputation satisfying three axioms: fairness (identical players are treated in the same way), efficiency $\left(\sum\limits_{j\in M} \sigma_j = v(M)\right)$ and linearity (if v and w are two characteristic functions defined for the same set of players, then $\sigma_j(v+w) = \sigma_j(v) + \sigma_j(w)$ for all $j \in M$).

To define the core, we need to introduce the concept of dominated imputations. Let $y = (y_1, \ldots, y_n)$ and $z = (z_1, \ldots, z_n)$ be two imputations of the cooperative game $< M, v, Y >$.

Definition 4.4 The imputation $y = (y_1, \ldots, y_m)$ dominates the imputation $z = (z_1, \ldots, z_m)$ through a coalition G if the following two conditions are satisfied:

$$\text{feasibility condition}: \sum_{j \in G} y_j \le v(G),$$

$$\text{preferability condition}: y_j > z_j, \quad \forall j \in G.$$

Definition 4.5 The core is the set of all undominated imputations

The following theorem, due to Gillies [18], characterizes the set of imputations belonging to the core of a cooperative game.

Theorem 4.1 *An imputation $y = (y_1, \ldots, y_m)$ is in the core if*

$$\sum_{j \in G} y_j \ge v(G), \forall G \subseteq M.$$

In other words, the above condition states that an imputation is in the core if it allocates to each possible coalition an outcome that is at least equal to what this coalition can secure by acting alone. Consequently, the core is defined by

$$C = \left\{ (y_1, \ldots, y_m), \text{such that} \sum_{j \in G} y_j \ge v(G), \forall G \subset M, \text{ and} \sum_{j \in M} y_j = v\,(M) \right\}.$$

Note that the core may be empty, may be a singleton or may contain many imputations.[5]

[5] The following example illustrates this statement. Consider a three-player cooperative game with characteristic function values given by

$$v(\{1\}) = v(\{2\}) = v(\{3\}) = 0,$$

$$v(\{1,2\}) = v(\{1,3\}) = v(\{2,3\}) = a, \quad v(\{1,2,3\}) = 1$$

where $0 < a \le 1$. It is easy to verify that three cases can occur: (i) If $0 < a < 2/3$, then the core contains all imputations satisfying $y_j \ge 0$, $\sum_{j \in G} y_j \ge a$ and $\sum_{j \in M} y_j = 1$. (ii) If $a = 2/3$, then the core is a singleton, that is, the only imputation belonging to the core is $(1/3, 1/3, 1/3)$. (iii) If $a > 2/3$, then the core is empty.

4.3 Cooperative and Noncooperative Payoffs-to-Go

Introduce the following notation:

$\tilde{u}_j\left(x^*\left(n_l^t\right)\right)$: An admissible strategy for player j in the subgame starting in node n_l^t, with initial state $x^*\left(n_l^t\right)$, $n_l^t \in \mathcal{N}^t, t = 1,\ldots,T$, and $\underline{\tilde{u}}\left(x^*\left(n_l^t\right)\right) = \left(\tilde{u}_j\left(x^*\left(n_l^t\right)\right) : j \in M\right)$.

$\tilde{u}_j^N\left(x^*\left(n_l^t\right)\right)$: S-adapted equilibrium strategy for player j in the subgame starting in node n_l^t, with initial state $x^*\left(n_l^t\right)$, $n_l^t \in \mathcal{N}^t, t = 1,\ldots,T$, and $\underline{\tilde{u}}^N\left(x^*\left(n_l^t\right)\right) = \left(\tilde{u}_j^N\left(x^*\left(n_l^t\right)\right) : j \in M\right)$.

$\tilde{u}_j^N\left(x^*\left(n_l^t\right), \left[n_v^\tau, n_w^T\right]\right)$: The trajectory of $\tilde{u}_j^N\left(x^*\left(n_l^t\right)\right)$ on the path emanating from node $n_v^\tau, n_v^\tau \in \mathcal{N}^\tau, \tau > t$, and terminating at node $n_w^T \in \mathcal{N}^T$.

$\tilde{u}_j^*\left(x^*\left(n_l^t\right)\right)$: Cooperative strategy (control) for player j in the subgame starting in node n_l^t, with initial state $x^*\left(n_l^t\right)$, $n_l^t \in \mathcal{N}^t, t = 1,\ldots,T$, and $\underline{\tilde{u}}^*\left(x^*\left(n_l^t\right)\right) = \left(\tilde{u}_j^*\left(x^*\left(n_l^t\right)\right) : j \in M\right)$.

$\tilde{u}_j^*\left(x^*\left(n_l^t\right), \left[n_v^\tau, n_w^T\right]\right)$: The trajectory of $\tilde{u}_j^*\left(x^*\left(n_l^t\right)\right)$ on the path emanating from node $n_v^\tau, n_v^\tau \in \mathcal{N}^\tau, \tau > t$, and terminating at node $n_w^T \in \mathcal{N}^T$.

$W_j^N\left(\underline{\tilde{u}}\left(x^*\left(n_l^t\right)\right)\right)$: S-adapted equilibrium payoff of player j in the subgame starting in node n_l^t, with initial state $x^*\left(n_l^t\right)$, $n_l^t \in \mathcal{N}^t, t = 1,\ldots,T$.

$W_j^*\left(\underline{\tilde{u}}\left(x^*\left(n_l^t\right)\right)\right)$: Payoff of player j in the cooperative game starting in node n_l^t, with initial state $x^*\left(n_l^t\right)$, $n_l^t \in \mathcal{N}^t, t = 1,\ldots,T$.

Remark 4.6 The trajectories $\tilde{u}_j^N\left(x^*\left(n_l^t\right), \left[n_v^\tau, n_w^T\right]\right)$ and $\tilde{u}_j^N\left(x^*\left(n_v^\tau\right), \left[n_v^\tau, n_w^T\right]\right)$ do not, in general, coincide. One reason is that the trajectory $\tilde{u}_j^N\left(x^*\left(n_l^t\right), \left[n_v^\tau, n_w^T\right]\right)$ has been computed assuming that the players have cooperated only during the time interval $[0, t]$, whereas $\tilde{u}_j^N\left(x^*\left(n_v^\tau\right), \left[n_v^\tau, n_w^T\right]\right)$ is computed under the assumption of a cooperative mode of play on $[0, \tau]$, with $\tau > t$.

If the players adopt the Shapley value, then, in the whole game, player j gets the following outcome:

$$\sigma_j\left(x^0\left(n^0\right)\right) = \sum_{\substack{G \subset M \\ j \in G}} \frac{(m-g)!(g-1)!}{m!}\left[v\left(G; x^0\left(n^0\right)\right) - v\left(G \backslash \{j\}; x^0\left(n^0\right)\right)\right], \quad (7)$$

with

$$\sum_{j \in M} \sigma_j\left(x^0\left(n^0\right)\right) = v\left(M; x^0\left(n^0\right)\right).$$

Similarly, the Shapley value in the subgame starting in node n_l^t and in state $\tilde{x}^*\left(n_l^t\right)$ is given by

$$\sigma_j\left(x^*\left(n_l^t\right)\right) = \sum_{\substack{G\subset M \\ j\in G}} \frac{(m-g)!(g-1)!}{m!}\left[v\left(G;x^*\left(n_l^t\right)\right) - v\left(G\backslash\{j\};x^*\left(n_l^t\right)\right)\right],$$

$$\sum_{j\in M}\sigma_j\left(x^*\left(n_l^t\right)\right) = v\left(M;x^*\left(n_l^t\right)\right). \tag{8}$$

Now, suppose that the players wish to implement an imputation in the core. The set of imputations in the core of the whole game is given by

$$C\left(x^0\left(n^0\right)\right) = \Big\{\left(y_1\left(x^0\left(n^0\right)\right),\ldots,y_m\left(x^0\left(n^0\right)\right)\right) \mid \sum_{j\in G} y_j\left(x^0\left(n^0\right)\right) \geq v(G;x^0),$$
$$\forall G\subset M, \quad \text{and} \quad \sum_{j\in M} y_j\left(x^0\left(n^0\right)\right) = v(M;x^0)\Big\}, \tag{9}$$

and in the subgame starting from node n_l^t, with state value $x^*\left(n_l^t\right)$, given by

$$C(x^*(n_l^t)) = \Big\{\left(y_1\left(x^*(n_l^t)\right),\ldots,y_m\left(x^*(n_l^t)\right)\right) \mid \sum_{j\in G} y_j \geq v(G;x^*(n_l^t))$$
$$\forall G\subset M, \quad \text{and} \quad \sum_{j\in M} y_j\left(x^*(n_l^t)\right) = v(M;x^*(n_l^t))\Big\}. \tag{10}$$

A main difficulty in defining a node-consistent core is that $C(x^0\left(n^0\right))$ and $C(x^*(n_l^t))$ are not singletons. This implies that the players must agree, at each node, on the imputation that they wish to implement in the subgame starting at that node. Further, we assume that the core of any subgame is nonempty.

4.4 Defining a Node-Consistent Allocation

A cooperative solution in DGPET is node consistent at $x^0\left(n^0\right)$, if the cooperative payoff-to-go of player $j, j \in M$, in the subgame starting at node $n_l^t \in \mathcal{N}^t, t = 1,\ldots,T$, is at least equal to the noncooperative payoff-to-go in this subgame. This will be achieved by introducing an imputation distribution procedure (IDP), that is, payment functions $\beta_j\left(x^*\left(n_l^t\right)\right), j \in M, n_l^t \in \mathcal{N}^t, t = 1,\ldots,T$. The specific values of an IDP will of course depend on the chosen imputation. The idea of IDP was originally introduced in Petrosjan and Danilov [43].

4.4.1 Node-Consistent Shapley Value

Let us suppose that the players choose the Shapley value as solution of the cooperative game.

Definition 4.6 An imputation distribution procedure of the Shapley value at $x_0\,(n_0)$ is given by $\left\{\beta_j\left(x^*\left(n_l^t\right)\right)\right\}_{n_l^t \in \mathcal{N}^t, t=1,\ldots,T}, j \in M$, satisfying

$$\sigma_j\left(x^0\left(n^0\right)\right) = \sum_{\theta=0}^{t-1} \sum_{n_k^\theta \in \mathcal{N}^\theta} \pi(n_k^\theta)\beta_j(x^*(n_k^\theta)), \quad \text{for all } j \in M. \tag{11}$$

Clearly, an IDP always exists as it simply requires the satisfaction of an accounting condition stating that any stream of payments to a player is feasible as long as its total expected value is equal to what that player is entitled to in the whole game. Note that the payments $\beta_j(x^*(n_k^\theta))$ are not (necessarily) equal to the realized payoffs, that is, $\phi_j^{n_l^t}\left(x^*(n_l^t), \underline{\tilde{u}}^*(n_l^t)\right)$. Now, we add the node-consistency condition.

Definition 4.7 The Shapley value $\sigma_j\left(x^0\left(n^0\right)\right)$ and the corresponding imputation distribution procedure $\left\{\beta_j\left(x^*\left(n_l^t\right)\right)\right\}_{n_l^t \in \mathcal{N}^t, t=1,\ldots,T}, j \in M$, are node consistent at $x_0\,(n_0)$, if for any $\left(x^*\left(n_l^t\right)\right)$, $n_l^t \in \mathcal{N}^t$, $t = 0, \ldots, T$, it holds that

$$\begin{aligned}\sigma_j\left(x^0\left(n^0\right)\right) &= \sum_{\theta=0}^{t-1} \sum_{n_k^\theta \in \mathcal{N}^\theta} \pi(n_k^\theta)\beta_j(x^*(n_k^\theta)) \\ &\quad + \sum_{n_k^\theta \in \mathcal{N}^t} \pi(n_k^\theta)\sigma_j\left(x^*(n_l^t)\right), \quad \forall\ j \in M.\end{aligned} \tag{12}$$

The definition states that what we allocate till any intermediate node using the IDP, plus the Shapley value payments in the subgame starting in that node must be equal to what player j is entitled to in the whole game, that is, her Shapley value $\sigma_j\left(x^0\left(n^0\right)\right)$. What remains to be done is to show that there exists an IDP satisfying the above definition. The following theorem, due to Reddy et al. [52], gives the result.

Theorem 4.2 *The IDP* $\left(\beta_1\left(x^*\left(n_l^t\right)\right), \ldots, \beta_m\left(x^*\left(n_l^t\right)\right)\right)$ *defined by*

$$\beta_j\left(x^*\left(n_l^t\right)\right) = \sigma_j\left(x^*\left(n_l^t\right)\right) - \sum_{n_k^{t+1} \in \mathcal{S}(n_l^t)} \pi(n_k^{t+1}|n_l^t)\sigma_j\left(x^*\left(n_k^{t+1}\right)\right), t = 0, \ldots, T-1, \tag{13}$$

$$\beta_j\left(x^*\left(n_l^T\right)\right) = \sigma_j\left(x^*\left(n_l^T\right)\right), \tag{14}$$

satisfies (12).

Proof See Reddy et al. [52]. □

The interpretation of this theorem is straightforward. At any terminal node n_l^T, the IDP payment is exactly the Shapley value in the static game at that node. At all other nodes, the IDP allocates to player j her Shapley value in the subgame starting at that node, minus the expected Shapley value in the subgames that are reached in the sequel. Note that $\beta_j\left(x^*\left(n_l^t\right)\right)$ can assume any sign.

4.4.2 Node-Consistent Core

Defining a node-consistent core is more demanding than defining a node-consistent Shapley value for two main reasons. First, the Shapley value in any subgame, including the whole game, always exists and is unique. The core may be empty in some of the subgames, if not in all of them. As we said before, we suppose here that the cores in all subgames are nonempty; otherwise the construction to follow will not be feasible. Second, at each intermediate node $n_l^t \in \mathcal{N}^t, t > 0$, the players need to agree on which imputation to select in $C(x^*(n_l^t))$, whereas there is no selection process in the case of Shapley value because $\sigma_j\left(x^*(n_l^t)\right)$ is uniquely defined. Note that both these issues pertain to cooperative game theory in general and are not specific to what is done here. Dealing with sets of imputations at each node that are not singletons leads to the following definition of an IDP, which is clearly more restrictive than the one stated above.

Definition 4.8 The node payments $\left\{\beta_j\left(x^*\left(n_l^t\right)\right)\right\}_{n_l^t \in \mathcal{N}^t, t=1,\ldots,T}, j \in M$, constitute an IDP of $y\left(x^0\left(n_0\right)\right) \in C(x^0\left(n^0\right))$, if they satisfy the following conditions:

$$y_j(x^0\,(n_0)) = \sum_{\theta=0}^{T} \sum_{n_l^\theta \in \mathcal{N}^\theta} \pi(n_l^\theta)\beta_j(x^*(n_l^\theta)), \tag{15}$$

$$\sum_{j\in M} \beta_j(x^*(n_l^t)) = \sum_{j\in M} \phi_j^{n_l^t}\left(x^*(n_l^t), \underline{\tilde{u}}^*(n_l^t)\right), \tag{16}$$

$$\sum_{j\in M} \beta_j(x^*(n_l^T)) = \sum_{j\in M} \Phi_j^{n_l^T}(x^*(n_l^T)), \tag{17}$$

where the two last conditions are satisfied for any $n_l^t \in \mathcal{N}^t, t = 0, \ldots, T-1$ (for 16), and any $n_l^T \in \mathcal{N}^T$.

The accounting condition (15) that must be satisfied for the whole game is the same as (11). The next two conditions in the above definition state that the sum of payments, at any node, must be equal to the sum of realized cooperative payoffs at that node. In economic terms, banking payoffs for future use, or borrowing from future periods are not allowed.

Definition 4.9 The imputation $y(x^0) \in C(x^0)$ and corresponding imputation distribution procedure

$$\left(\left\{\beta_j\left(x^*\left(n_l^t\right)\right)\right\}_{n_l^t \in \mathcal{N}^t, t=1,\dots,T} : j \in M\right),$$

are called node consistent in the whole game if for any state $x^*(n_l^t)$, $n_l^t \in \mathcal{N}^t$, $t = 0, \dots, T$, there exists $y(x^*(n_l^t)) = (y_1(x^*(n_l^t)), \dots, y_m(x^*(n_l^t))) \in C(x^*(n_l^t))$ satisfying the following condition:

$$y_j(x^0(n_0)) = \sum_{\theta=0}^{t-1} \sum_{n_k^\theta \in \mathcal{N}^\theta} \pi(n_k^\theta)\beta_j(x^*(n_k^\theta)) + \sum_{n_k^t \in \mathcal{N}^t} \pi(n_k^\theta) y_j\left(x^*(n_l^t)\right). \tag{18}$$

If the payoffs in the nodes are allocated according to the imputation distribution procedure, then node-consistency of imputation $y(x^0)$ from the core means that one can define a feasible distribution procedure under which the continuation values at every node are in the core of the continuation game.

Definition 4.10 The core $C(x^0)$ in the whole game is a node-consistent allocation mechanism if any imputation y from the core $C(x^0)$ is node consistent.

Theorem 4.3 *If the core $C(x^0)$ of the whole game and the core $C(x^*(n_l^t))$ of the subgame starting from any node n_l^t are nonempty, then the core $C(x^0)$ is node consistent when the corresponding imputation distribution procedure for each imputation $y(x^0) \in C(x^0)$ satisfies the following conditions*

for $t = 0, \dots, T-1$:

$$\beta_j(x^*(n_l^t)) = y_j(x^*(n_l^t)) - \sum_{n_k^{t+1} \in \mathcal{S}(n_l^t)} \pi(n_k^{t+1}|n_l^t) y_j(x^*(n_k^{t+1})), \tag{19}$$

and for $t = T$:

$$\beta_j(x^*(n_l^T)) = y_j(x^*(n_l^T)), \tag{20}$$

where $y(x^*(n_l^t)) = (y_1(x^*(n_l^t)), \dots, y_m(x^*(n_l^t))) \in C(x^*(n_l^t))$ for any $n_l^t \in \mathcal{N}^t$, $t = 0, \dots, T$ and $\pi(n_k^{t+1}|n_l^t)$ is the conditional probability that node n_k^{t+1} is reached if node n_l^t has already been reached.

Proof See Parilina and Zaccour [38]. □

If the core $C(x^0)$ of the whole game and the core $C(x^*(n_l^t))$ of a subgame starting from any node n_l^t are nonempty, we can always find at least one imputation $y(x^*(n_l^t)) \in C(x^*(n_l^t))$ and, using the given imputations $y(x^*(n_l^t))$ for all nodes $n_l^t \in \mathscr{N}^t$, $t = 0, \ldots, T$, construct the imputation distribution procedure $\left(\left\{\beta_j(x^*(n_l^t))\right\}_{n_l^t \in \mathscr{N}^t, t=0,\ldots,T} : j \in M\right)$, with formulas (19) and (20) for any imputation from the core $C(x^0)$.

The IDP and the realized outcomes at node $n_l^t \in \mathscr{N}^t, t = 0, \ldots, T-1$ are related by the following side payments:

$$\omega_j(n_l^t, x^*(n_l^t)) = \beta_j(x^*(n_l^t)) - \phi_j^{n_l^t}(x^*(n_l^t), \underline{\tilde{u}}^*(n_l^t)), \tag{21}$$

and for $\forall n_l^T \in \mathscr{N}^T$:

$$\omega_j(n_l^T, x^*(n_l^T)) = \beta_j(x^*(n_l^T)) - \Phi_j^{n_l^T}(x^*(n_l^T)), \tag{22}$$

where $\omega_j(n_l^t, x^*(n_l^t))$ is the transfer payment that player j makes in node n_l^t over the cooperative trajectory $x^*(n_l^t)$, such that

$$\sum_{j \in M} \omega_j(n_l^t, x^*(n_l^t)) = 0,$$

for any node n_l^t over cooperative trajectory $x^*(n_l^t)$. Clearly, $\omega_j(n_l^t, x^*(n_l^t))$ can assume any sign depending on the sign of the difference in the right-hand sides of (21)–(22). □

5 Concluding Remarks

We showed in this paper how to decompose over time the Shapley value and an imputation in the core such that cooperation is sustained at any node of the event tree. Many extensions to our framework can be envisioned. First, it should not be complicated to define node consistency for other solution concepts, such as proportional payments and a Nash bargaining procedure. Second, we assumed that the core $C(x^*(n_l^t))$ in any subgame is nonempty. An interesting open question is whether cooperation can still be sustained if the cores in some of the subgames (not the whole game) are empty. Finally, it would be interesting to consider node consistency for DGPET when the end of the horizon is random.

Acknowledgements This paper draws from my previous work on the subject, in particular Reddy, Shevkoplyas and Zaccour [52] and Parilina and Zaccour [38]. I would like to thank my co-authors in these papers, as well as Alain Haurie and Leon Petrosjan for many stimulating discussions over the last two decades or so on dynamic games played over event trees and time consistency in dynamic games. Research supported by SSHRC, Canada.

References

1. Aumann, R.J.: The core of a cooperative game without side payments. Transactions of the American Mathematical Society **98**, 539–552 (1961)
2. Avrachenkov, K., Cottatellucci, L., Maggi, L.: Cooperative Markov decision processes: Time consistency, greedy players satisfaction, and cooperation maintenance. International Journal of Game Theory **42**, 239–262 (2013)
3. Breton, M., Sokri A., Zaccour, G.: Incentive equilibrium in an overlapping-generations environmental game. European Journal of Operational Research **185**, 687–699 (2008)
4. Buratto, A., Zaccour, G.: Coordination of advertising strategies in a fashion licensing contract. Journal of Optimization Theory and Applications **142**, 31–53 (2009)
5. Chander, P., Tulkens, H.: The core of an economy with multilateral environmental externalities. International Journal of Game Theory **23**, 379–401 (1997)
6. Chiarella, C., Kemp, M.C., Long, N.V., Okuguchi, K.: On the economics of international fisheries. International Economic Review **25**, 85–92 (1984)
7. Dockner, E., Jørgensen, S., Van Long, N., Sorger, G.: Differential Games in Economics and Management Science. Cambridge University Press, Cambridge (2000)
8. De Frutos, J., Martín-Herrán, G.: Does flexibility facilitate sustainability of cooperation over time? A case study from environmental economics. Journal of Optimization Theory and Applications **165**, 657–677 (2015)
9. De Giovanni, P., Reddy, P.V., Zaccour, G.: Incentive strategies for an optimal recovery program in a closed-loop supply chain. European Journal of Operational Research **249**, 605–617 (2016)
10. Dutta, P.K.: A folk theorem for stochastic games. Journal of Economic Theory **66**, 1–32 (1995)
11. Ehtamo, H., Hämäläinen, R.P.: On affine incentives for dynamic decision problems. In: Basar, T. (ed.) Dynamic Games and Applications in Economics, pp. 47–63. Springer, Berlin (1986)
12. Ehtamo, H., Hämäläinen, R.P.: Incentive strategies and equilibria for dynamic games with delayed information. Journal of Optimization Theory and Applications **63**, 355–370 (1989)
13. Ehtamo, H., Hämäläinen, R.P.: A cooperative incentive equilibrium for a resource management problem. Journal of Economic Dynamics and Control **17**, 659–678 (1993)
14. Filar, J., Petrosjan, L.: Dynamic cooperative games. International Game Theory Review **2**(1), 47–65 (2000)
15. Genc, T., Reynolds, S.S., Sen, S.: Dynamic oligopolistic games under uncertainty: A stochastic programming approach. Journal of Economic Dynamics & Control **31**, 55–80 (2007)
16. Genc, T., Sen, S.: An analysis of capacity and price trajectories for the Ontario electricity market using dynamic Nash equilibrium under uncertainty. Energy Economics **30**, 173–191 (2008)
17. Germain, M. Toint, P., Tulkens H., de Zeeuw A.: Transfers to sustain dynamic core-theoretic cooperation in international stock pollutant control. Journal of Economic Dynamics & Control **28**(1), 79–99 (2003)
18. Gillies, D.B.: Some Theorems on *N*-Person Games, Ph.D. Thesis, Princeton University (1953)
19. Haurie, A.: A note on nonzero-sum differential games with bargaining solution. Journal of Optimization Theory and Applications **18**, 31–39 (1976)
20. Haurie, A., Pohjola, M.: Efficient equilibria in a differential game of capitalism. Journal of Economic Dynamics and Control **11**, 65–78 (1987)
21. Haurie, A., Krawczyk, J.B., Roche, M.: Monitoring cooperative equilibria in a stochastic differential game. Journal of Optimization Theory and Applications **81**, 79–95 (1994)
22. Haurie, A., Krawczyk, J.B., Zaccour, G.: Games and Dynamic Games. Scientific World, Singapore (2012)
23. Haurie, A., Zaccour. G., Smeers, Y.: Stochastic equilibrium programming for dynamic oligopolistic markets. Journal of Optimization Theory and Applications **66**(2), 243–253 (1990)
24. Haurie, A., Zaccour, G.: S-adapted equilibria in games played over event trees: An overview. Annals of the International Society of Dynamic Games **7**, 367–400 (2005)

25. Jørgensen, S., Martín-Herrán, G., Zaccour, G.: Agreeability and time-consistency in linear-state differential games. Journal of Optimization Theory and Applications **119**, 49–63 (2003)
26. Jørgensen, S., Martín-Herrán, G., Zaccour, G.: Sustainability of cooperation overtime in linear-quadratic differential game. International Game Theory Review **7**, 395–406 (2005)
27. Jørgensen, S., Zaccour, G.: Time consistent side payments in a dynamic game of downstream pollution. Journal of Economic Dynamics and Control **25**, 1973–1987 (2001)
28. Kaitala, V., Pohjola, M.: Economic development and agreeable redistribution in capitalism: Efficient game equilibria in a two-class neoclassical growth model. International Economic Review **31**, 421–437 (1990)
29. Kaitala, V., Pohjola, M.: Sustainable international agreements on greenhouse warming: A game theory study. Annals of the International Society of Dynamic Games **2**, 67–87 (1995)
30. Kanani Kuchesfehani, K., Zaccour, G.: S-adapted equilibria in games played over event trees with coupled constraints. Journal of Optimization Theory and Applications **166**, 644–658 (2015)
31. Lehrer, E., Scarsini, M.: On the core of dynamic cooperative games. Dynamic Games and Applications **3**, 359–373 (2013)
32. Martín-Herrán, G., Rincón-Zapatero, J.P.: Efficient Markov perfect Nash equilibria: Theory and application to dynamic fishery games. Journal of Economics Dynamics and Control **29**, 1073–1096 (2005)
33. Martín-Herrán, G., Taboubi, S.: Shelf-space allocation and advertising decisions in the marketing channel: A differential game approach. International Game Theory Review **7**(03), 313–330 (2005)
34. Martín-Herrán, G., Zaccour, G.: Credibility of incentive equilibrium strategies in linear-state differential games. Journal of Optimization Theory and Applications **126**, 1–23. (2005)
35. Martín-Herrán, G., Zaccour, G.: Credible linear-incentive equilibrium strategies in linear-quadratic differential games. Annals of the International Society of Dynamic Games **10**, 261–292 (2009)
36. Ordeshook, P.C.: Game theory and Political Theory. Cambridge University Press, Cambridge, UK (1986)
37. Osborne, M.J., Rubinstein, A.: A Course in Game Theory. MIT Press, Cambridge, MA (1994)
38. Parilina, E., Zaccour, G.: Node-consistent core for games played over event trees. Automatica **55**, 304–311 (2015a)
39. Parilina, E., Zaccour, G.: Approximated cooperative equilibria for games played over event trees. Operations Research Letters **43**, 507–513 (2015b)
40. Petrosjan, L.: Stable solutions of differential games with many participants. Viestnik of Leningrad University **19**, 46–52 (1977)
41. Petrosjan, L.: Differential Games of Pursuit. World Scientific, Singapore, 270–282 (1993)
42. Petrosjan, L., Baranova, E.M., Shevkoplyas, E.V.: Multistage cooperative games with random duration. Proceedings of the Steklov Institute of Mathematics (Supplementary issues), suppl. 2, S126–S141 (2004)
43. Petrosjan, L., Danilov, N.N.: Stability of solutions in nonzero sum differential games with transferable payoffs. Journal of Leningrad University **N1**, 52–59 (in Russian) (1979)
44. Petrosjan, L., Danilov, N.N.: Cooperative Differential Games and Their Applications. Tomsk University Press, Tomsk (1982)
45. Petrosjan, L., Danilov, N.N.: Classification of dynamically stable solutions in cooperative differential games. Isvestia of High School **7**, 24–35 (in Russian) (1986)
46. Petrosjan, L., Zaccour, G.: Time-consistent Shapley value of pollution cost reduction. Journal of Economic Dynamics and Control **27**, 381–398 (2003)
47. Petrosjan, L., Zenkevich, N.A.: Game Theory. World Scientific, Singapore (1996)
48. Pineau, P.-O., Murto, P.: An oligopolistic investment model of the Finnish electricity market. Annals of Operations Research **121**, 123–148 (2003)
49. Pineau, P.-O., Rasata, H., Zaccour, G.: A Dynamic Oligopolistic Electricity Market Model with Interdependent Segments, Energy Journal **32**(4), 183–217 (2011a)

50. Pineau, P.-O., Rasata, H., Zaccour, G.: Impact of some parameters on investments in oligopolistic electricity markets. European Journal of Operational Research **213**(1), 180–195 (2011b)
51. Predtetchinski, A.: The strong sequential core for stationary cooperative games. Games and Economic Behavior **61**, 50–66 (2007)
52. Reddy, P. V., Shevkoplyas E., Zaccour, G.: Time-consistent Shapley value for games played over event trees. Automatica **49**(6), 1521–1527 (2013)
53. Rosen, J.B.: Existence and uniqueness of equilibrium points for concave *n*-person games. Econometrica **33**(3), 520–534 (1965)
54. Tolwinski, B., Haurie A., Leitmann, G.: Cooperative equilibria in differential games. Journal of Mathematical Analysis and Applications **119**, 182–202 (1986)
55. Von Neumann, J., Morgenstern, O.: Theory of Games and Economic Behavior. Princeton University Press, Princeton, NJ (1944)
56. Xu, N., Veinott Jr., A.: Sequential stochastic core of a cooperative stochastic programming game. Operations Research Letters **41**, 430–435 (2013)
57. Yeung, D.W.K., Petrosjan, L.: Proportional time-consistent solutions in differential games. Proceedings of International Conference on Logic, Game Theory and Applications, Saint Petersburg, 254–256 (2001)
58. Yeung, D.W.K., Petrosjan, L.: Consistent solution of a cooperative stochastic differential game with nontransferable payoffs. Journal of Optimization Theory and Applications **124**, 701–724 (2005a)
59. Yeung, D.W.K., Petrosjan, L.: Cooperative Stochastic Differential Games. Springer, New York, NY (2005b)
60. Yeung, D.W.K., Petrosjan, L.: Dynamically stable corporate joint ventures. Automatica **42**, 365–370 (2006)
61. Yeung, D.W.K., Petrosjan, L., Yeung, P.M.: Subgame consistent solutions for a class of cooperative stochastic differential games with nontransferable payoffs. Annals of the International Society of Dynamic Games **9**, 153–170 (2007)
62. Zaccour, G.: Théorie des jeux et marchés énergétiques: marché européen de gaz naturel et échanges d'électricité, Ph.D. Thesis, HEC Montréal (1987)
63. Zaccour, G.: Time consistency in cooperative differential games: A tutorial. INFOR **46**(1), 81–92 (2008)

Index

R. Melnik et al. (eds.), *Recent Progress and Modern Challenges in Applied Mathematics, Modeling and Computational Science*, Fields Institute Communications 79, DOI 10.1007/978-1-4939-6969-2