

MATRIX Book Series 2

David R. Wood *Editor-in-chief*
Jan de Gier · Cheryl E. Praeger
Terence Tao *Editors*

2017 MATRIX Annals

MATRI 

 Springer

Editors

David R. Wood (*Editor-in-Chief*)

Jan de Gier

Cheryl E. Praeger

Terence Tao

MATRIX is Australia's international and residential mathematical research institute. It facilitates new collaborations and mathematical advances through intensive residential research programs, each lasting 1–4 weeks.

More information about this series at <http://www.springer.com/series/15890>

David R. Wood
Editor-in-Chief

Jan de Gier • Cheryl E. Praeger • Terence Tao
Editors

2017 MATRIX Annals

MATRI 



MONASH
University



THE UNIVERSITY OF
MELBOURNE



ACEMJS

AUSTRALIAN RESEARCH COUNCIL CENTRE OF EXCELLENCE FOR
MATHEMATICAL AND STATISTICAL FRONTIERS



Springer

Editors

David R. Wood (*Editor-in-Chief*)
Monash University
Melbourne, Australia

Jan de Gier
The University of Melbourne
Melbourne, Australia

Cheryl E. Praeger
University of Western Australia
Perth, Australia

Terence Tao
UCLA
Los Angeles, CA, USA

ISSN 2523-3041

ISSN 2523-305X (electronic)

MATRIX Book Series

ISBN 978-3-030-04160-1

ISBN 978-3-030-04161-8 (eBook)

<https://doi.org/10.1007/978-3-030-04161-8>

Mathematics Subject Classification (2010): 05-XX, 11-XX, 14-XX, 35-XX, 35R30, 81-XX, 82-XX, 91Gxx

© Springer Nature Switzerland AG 2019

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

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

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

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

Preface

MATRIX is Australia's international and residential mathematical research institute. It was established in 2015 and launched in 2016 as a joint partnership between Monash University and The University of Melbourne, with seed funding from the ARC Centre of Excellence for Mathematical and Statistical Frontiers. The purpose of MATRIX is to facilitate new collaborations and mathematical advances through intensive residential research programs, which are currently held in Creswick, a small town nestled in the beautiful forests of the Macedon Ranges, 130 km west of Melbourne.

This book is a scientific record of the eight programs held at MATRIX in 2017:

- *Hypergeometric Motives and Calabi–Yau Differential Equations*
- *Computational Inverse Problems*
- *Integrability in Low-Dimensional Quantum Systems*
- *Elliptic Partial Differential Equations of Second Order: Celebrating 40 Years of Gilbarg and Trudinger's Book*
- *Combinatorics, Statistical Mechanics, and Conformal Field Theory*
- *Mathematics of Risk*
- *Tutte Centenary Retreat*
- *Geometric R-Matrices: from Geometry to Probability*

The MATRIX Scientific Committee selected these programs based on scientific excellence and the participation rate of high-profile international participants. This committee consists of: Jan de Gier (Melbourne University, Chair), Ben Andrews (Australian National University), Darren Crowdy (Imperial College London), Hans De Sterck (Monash University), Alison Etheridge (University of Oxford), Gary Froyland (University of New South Wales), Liza Levina (University of Michigan), Kerrie Mengersen (Queensland University of Technology), Arun Ram (University of Melbourne), Joshua Ross (University of Adelaide), Terence Tao (University of California, Los Angeles), Ole Warnaar (University of Queensland), and David Wood (Monash University).

These programs involved organisers from a variety of Australian universities, including Australian National University, Monash University, Queensland Univer-

sity of Technology, University of Newcastle, University of Melbourne, University of Queensland, University of Sydney, University of Technology Sydney, and University of Western Australia, along with international organisers and participants.

Each program lasted 1–4 weeks, and included ample unstructured time to encourage collaborative research. Some of the longer programs had an embedded conference or lecture series. All participants were encouraged to submit articles to the MATRIX Annals.

The articles were grouped into refereed contributions and other contributions. Refereed articles contain original results or reviews on a topic related to the MATRIX program. The other contributions are typically lecture notes or short articles based on talks or activities at MATRIX. A guest editor organised appropriate refereeing and ensured the scientific quality of submitted articles arising from each program. The Editors (Jan de Gier, Cheryl E. Praeger, Terence Tao and myself) finally evaluated and approved the papers.

Many thanks to the authors and to the guest editors for their wonderful work.

MATRIX is hosting eight programs in 2018, with more to come in 2019; see www.matrix-inst.org.au. Our goal is to facilitate collaboration between researchers in universities and industry, and increase the international impact of Australian research in the mathematical sciences.

David R. Wood
MATRIX Book Series Editor-in-Chief

Hypergeometric Motives and Calabi–Yau Differential Equations

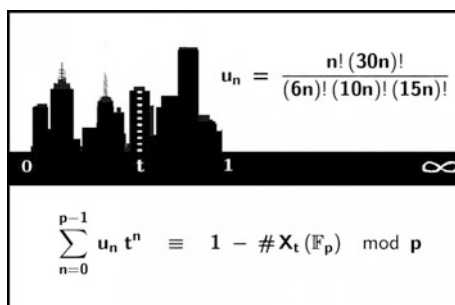
8–27 January 2017

Organisers

Ling Long
Louisiana State Uni

Masha Vlasenko
Institute of Mathematics of the
Polish Academy of Sciences

Wadim Zudilin
Uni Newcastle



The majority of the articles presented below are extended abstracts of the talks given by program participants at the workshop that took place from January 16 to 20, 2017. Some of them present a new perspective or results that appeared due to collaboration following the activity in Creswick.

The two main topics of the program, Calabi–Yau differential equations and hypergeometric motives, provide an explicit approach and experimental ground to such important themes in contemporary arithmetic geometry as the Langlands program, motives and mirror symmetry. Hypergeometric motives are families of motives whose periods are given by generalised hypergeometric functions. Their L -functions are expected to cover a wide range of known L -functions. Due to the recent work of researchers (many of whom were present in Creswick) it is now possible to compute L -functions of hypergeometric motives efficiently. Thus one can test the standard conjectures, e.g. on special values and modularity, for motives of any degree and weight. Many algorithms for computing with the hypergeometric motives are now implemented in the computer algebra system Magma.

Local factors of hypergeometric L -functions can be investigated by the means of finite hypergeometric functions, another topic to which a few articles in this volume are devoted. The techniques developed by the authors allow to transport classical formulas to the finite field setting, count points on algebraic varieties over finite fields, study their congruence properties and Galois representations. Importantly,

finite hypergeometric functions can be viewed as periods of motives over finite fields.

Periods over finite fields form a new angle of understanding the integrality phenomenon arising in mirror symmetry. Originally discovered by physicists in the mid 1980s, mirror symmetry remains one of the central research themes binding string theory and algebraic geometry. Numerous examples show that the expression of the mirror map in so-called canonical coordinates possesses rich arithmetic properties. This expression involves particular solutions to a Picard–Fuchs differential equation of a family of Calabi–Yau manifolds near a singular point. Application of p -adic methods to the study of Calabi–Yau differential equations gives a very promising prospective, as it is announced in the final article by Duco van Straten.

The three weeks at the MATRIX institute were intense and fruitful. To illustrate these words, there was a special lecture by Fernando Rodriguez Villegas scheduled at the very last moment on Thursday afternoon of the workshop week, in which he presented, jointly with David Roberts and Mark Watkins, a new conjecture on motivic supercongruences that was invented in Creswick. This talk influenced what happened in the last week of the workshop. David Broadhurst gave his two lectures on the very first and very last days of the program, reporting in the second talk on the tremendous progress achieved by him in collaboration with David Roberts over the three weeks.

We are confident that ideas and projects that emerged during the program will drive our field of research in the coming years.

Masha Vlasenko
Guest Editor



Participants

James Wan (SUTD, Singapore), Fang-Ting Tu (Louisiana State), Yifan Yang (National Chiao Tung University), Éric Delaygue (Institut Camille Jordan, Lyon), John Voight (Dartmouth), Adriana Salerno (Bates College), Alex Ghitza (Melbourne), Mark Watkins (Sydney), Piotr Achinger (IHES) with Helena, Jan de Gier (Melbourne), David Broadhurst (Open University), Ole Warnaar (Queensland), Ravi Ramakrishna (Cornell), Fernando Rodriguez Villegas (ICTP, Trieste), Sharon Frechette (College of the Holy Cross), Robert Osburn (University College Dublin), Frits Beukers (Utrecht), Paul Norbury (Melbourne), David Roberts (Minnesota Morris), Duco van Straten (Johannes Gutenberg), Holly Swisher (Oregon State), Abdellah Sebbar (Ottawa)

Computational Inverse Problems

11–23 June 2017

Organisers

Tiangang Cui
Monash Uni

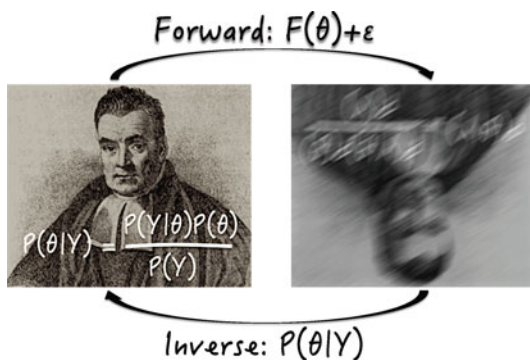
Hans De Sterck
Monash Uni

Markus Hegland
Australian National Uni

Youssef Marzouk
Massachusetts Inst Tech

Ian Turner
Queensland Uni Tech

Karen Willcox
Massachusetts Inst Tech



The integration of complex data sets into large-scale computational models is one of the central challenges of modern applied mathematics. This challenge is present in almost every application area within science and engineering, e.g., geosciences, biological systems, astrophysics, meteorology, aerospace, and subsurface flow. At the heart of this challenge often lies an inverse problem: we seek to convert indirect data into useful characterisations of the unknown model parameters including source terms, initial or boundary conditions, model structure, physical coefficients, etc. Solution of the inverse problem, along with model prediction and uncertainty assessment, can be cast in a Bayesian setting and thus naturally characterised by the posterior distribution over unknown parameters conditioned on the data. Unfortunately, solution of such statistical inverse problems for systems governed by large-scale, complex computational models has traditionally been intractable: models are complicated and computationally expensive to evaluate; available indirect data are

often limited, noisy, and subject to natural variation; inversion algorithms often scale poorly to high-dimensional, or in principle infinite-dimensional, model parameters.

Our program contributed to the active international research effort in computational mathematics to connect theoretical developments with algorithmic advancements, buoyed by a range of cutting-edge applications. The program attracted a total of 47 attendees from a diverse range of highly relevant fields. Our attendees include renowned researchers in numerical analysis, scientific computing, optimisation, and stochastic computation, as well as high profile domain experts working in meteorology, super-resolution imaging, aerospace, and subsurface. The program began with a week of mini-conference. Seven 45-min plenary presentations and twenty 30-min invited presentations were scheduled during the mini-conference. In the second week, we organised thirteen 45-min presentations in the mornings and reserved afternoons for collaboration.

During the program, our attendees presented and extensively collaborated the following key topics in computational inverse problems:

- Deterministic and statistical methods for inverse problems.
- Advanced Markov chain Monte Carlo and quasi Monte Carlo methods.
- Optimal transport theory and its current and potential applications in inverse problems.
- Model reduction methods and multi-scale methods.
- Scalable experimental design methods.
- High performance numerical solvers, including multilevel methods.
- Applications in geothermal engineering, additive manufacturing, aeronautics, remote sensing, and super-resolution imaging.

The articles in this proceedings represent different aspects of the program. For example, Bardsley and Cui describe an optimisation-based methods for nonlinear hierarchical Bayesian inverse problem, Fox et al. presents a novel methods for sequential inverse problems using the Frobenius-Perron operator, MacNamara, McLean and Burrage present an adaptive contour integration methods for solving master equations, Guo, Loeper, and Wang present initial investigations of using optimal transport to solve inverse problems in finance, Harrach and Rieger present a set optimisation technique for reconstructing electrical impedance tomography image using single-measurement, Haario et al. investigates new ideas on characterising chaotic stochastic differential equations, Lamminpää et al. presents a case study on the atmospheric remote sensing, Ye, Roosta-Khorasani, and Cui present an extensive survey on optimisation methods used in inverse problems.

We would like to thank all of the authors who took the time to contribute to this volume. We would also like to thank the MATRIX staff and officials for hosting and facilitating this wonderful event and giving us the opportunity to share our work with this volume.

Tiangang Cui and Hans De Sterck
Guest Editors



Participants

Bart van Bloemen Waanders, Benjamin Peherstorfer, Colin Fox, Gregoire Loeper, Habib N. Najm, Harriet LI, Heikki Haario, Janosch Rieger, Jinglai LI, John Bardsley, Josef Dick, Kate Lee, Kody Law, Lutz Gross, Marko Laine, Nan Ye, Oliver Maclaren, Olivier Zahm, Omar Ghattas, Qinian Jin, Tianhai Tan, Tim Garoni, Zheng Wang, Elizabeth Qian, Gianluca Detommaso, Ruanui Nicholson, Elvar Bjarkason, Fred Roosta, Shev MacNamara, Alessio Spantini, Amani Alahmadi, Hoang Viet Ha, Mia (Xin) Shu, Carl (Ao) Shu, Thien Binh Nguyen, Oliver Krzysik, Brad Marvin, Ellen B Le, Jesse Adams, Hongbo Xie, Hans Elmlund, Cyril Rebol

Integrability in Low-Dimensional Quantum Systems

26 June–21 July 2017

Organisers

Murray Batchelor
Australian National Uni

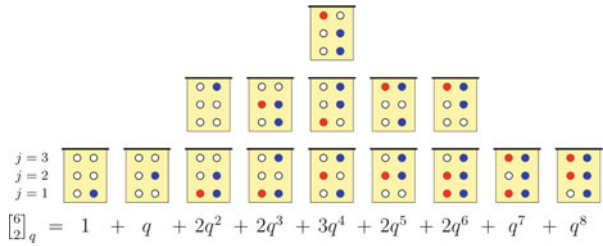
Patrick Dorey
Uni Durham

Giuseppe Mussardo
SISSA Trieste

Paul Pearce
Uni Melbourne

Chaiho Rim
Sogang, Seoul

Clare Dunning
Uni Kent



This MATRIX program focused on aspects of integrability in low-dimensional quantum systems and areas of application. It was organized around currently active hot topics and open problems. The emphasis was on focused research and interaction in small groups to achieve real collaboration. The research topics included:

- AdS/CFT
- Bethe ansatz and quantum spin chains
- Bulk and boundary conformal and quantum field theory
- Cold atoms, strongly correlated systems
- Integrability in models of matter-light interaction
- Logarithmic CFT
- ODE/IM and its massive variants
- Quantum quenches and quantum entanglement
- Random matrix approach to CFT and integrability

Among the integrability community, this workshop was a major event on the international scene and enabled us to bring together scientists at the leading edge of research in integrable quantum systems in low-dimensions. Indeed, with 59 participants over 3 weeks, a significant proportion of the active world-wide community working on quantum integrability was in attendance.

Classical integrability of two-dimensional systems and the related quantum integrability of one-dimensional systems are finding areas of application in statistical physics, condensed matter physics and particle physics in addition to contributing to highly mathematical topics such as Yang-Baxter algebras, quantum groups, cluster algebras, affine Lie algebras and combinatorial representation theory. With a series of Introductory Lectures on Hot Topics and advanced seminars, this workshop offered extensive training to graduate students and Early Career Researchers working in integrability and related topics.

Highlights, among many of the meeting, include the announcement of (1) the analytic calculation of the conformal partition functions of two-dimensional critical percolation, (2) the demonstration of the quantum toroidal integrability behind the AGT correspondence as well as (3) some striking progress on the mathematical description of fusion within the affine Temperley-Lieb algebra. Contributed articles included in these MATRIX Annals cover the topics of (1) form factors, (2) the combinatorics and generating functions of RNA structures, (3) supersymmetric quantum chains and (4) proofs of factorization and sum-to-1 properties of the $A_n^{(1)}$ face models. During the program there were also several groups of collaborators informally reporting rapid progress including (1) a collaboration explaining the mysteries of Baxter's Q -matrix for $sl(2)$ models at roots of unity and (2) a collaboration deriving analytically the correlation functions and conformal weights of critical dense polymers. Many physical applications to quantum quenches, ultracold atoms and matter-light interaction were also showcased during the meeting. All of these represent significant advancement in our discipline.

We gratefully acknowledge the generous support of our sponsors—MATRIX, the Australian Mathematical Sciences Institute (AMSI), the Australian Mathematical Society (AustMS) and the Asia Pacific Center for Theoretical Physics (APCTP). We particularly thank Jan de Gier for his encouragement in bringing this program together. We also thank the very helpful MATRIX staff at Melbourne and Creswick campuses, as well as our outstanding chef Adam, for their many significant contributions to the success of this meeting. Lastly, we thank the authors who kindly took the time and made the effort to contribute to this volume.

Chaiho Rim and Paul Pearce
Guest Editors



Participants

Changrim Ahn (Ewha, Seoul, Korea), Zoltan Bajnok (Wigner, Hungary), Jean-Emile Bourgine (KIAS, Seoul, Korea), Daniel Braak (Augsburg, Germany), Jun-peng Cao (CAS, Beijing, China), Sang Kwan Choi (Sichuan University, China), Ed Corrigan (York, UK), György Fehér (Budapest, Hungary), Angela Foerster (Rio Grande do Sul, Brazil), Holger Frahm (Hannover, Germany), Azat Gainutdinov (Tours, France), Frank Gühmann (Wuppertal, Germany), Xiwen Guan (CAS, Wuhan, China), Jesper Jacobsen (ENS, Paris, France), Shashank Kanade (Alberta, Canada), Andreas Klümpe (Wuppertal, Germany), Karol Kozłowski (Lyon, France), Atsuo Kuniba (Tokyo, Japan), Masahide Manabe (Warsaw, Poland), Chihiro Matsui (Tokyo, Japan), Yutaka Matsuo (Tokyo, Japan), Jianin Mei (Dalian, China), Alexi Morin-Duchesne (Louvain, Belgium), Rafael Nepomechie (Miami, USA), Ovidiu Patu (Bucharest, Romania), Francesco Ravanini (Bologna, Italy), Yvan Saint-Aubin (Montréal, Canada), Kareljan Schoutens (Amsterdam, Netherlands), Junji Suzuki (Shizuoka, Japan), Gabor Takacs (Budapest, Hungary), Masato Wakayama (Kyushu, Japan), Yupeng Wang (CAS, Beijing, China), Waltraut Wustmann (Maryland, USA), Wen-Li Yang (Xian, China), Hong Zhang (ITP, Beijing, China), Huan-Xiang Zhou (Chongqing, China), Rui-Dong Zhu (Tokyo, Japan), Zeying Chen (Uni Melbourne), Jan de Gier (Uni Melbourne), Omar Foda (Uni Melbourne), Alexandr Garbali (Uni Melbourne), Phil Isaac (Uni Queensland), Kazuya Kawasetsu (Uni Melbourne), Sergii Koval (Australian National Uni), Jon Links (Uni Queensland), Tianshu Liu (Uni Melbourne), Vladimir Mangazeev (Australian National Uni), Thomas Quella (Uni Melbourne), Jorgen Rasmussen (Uni Queensland), David Ridout (Uni Melbourne), Boris Runov (Australian National Uni), William Stewart (Uni Melbourne), Michael Wheeler (Uni Melbourne), Paul Zinn-Justin (Uni Melbourne)

Elliptic Partial Differential Equations of Second Order: Celebrating 40 Years of Gilbarg and Trudinger's Book

16–28 October 2017

Organisers

Lucio Boccardo
Sapienza Uni Roma

Florica-Corina Cirstea
Uni Sydney

Julie Clutterbuck
Monash Uni

L. Craig Evans
Uni California Berkeley

Enrico Valdinoci
Uni Melbourne

Paul Bryan
Macquarie Uni



Our program celebrated the 40th anniversary of the publication of Gilbarg and Trudinger's highly influential "Elliptic Partial Differential Equations of Second Order", one of the most highly cited texts in mathematics (over 10,000 citations). We sought to link past research with future perspectives, by discussing what the important developments in the area during these 40 years have been and what are the new trends of contemporary research. Particular attention was given to some of the topics in which the book served as a great source of inspiration, such as fully nonlinear PDEs, viscosity solutions, Hessian equations, optimal transport, stochastic point of view, geometric flows, and so on.

The first week of the program consisted of a series of introductory lectures aimed at Ph.D. students and postdocs, featuring in particular lectures given by Neil Trudinger himself. Special thanks go to Connor Mooney who gave a beautiful series

of lectures with only 24 h notice after a late cancellation by the original lecturer due to illness. The lectures were:

- **Estimates for fully nonlinear equations**, Neil Trudinger
- **Mean Curvature Flow with free boundary**, Valentina Wheeler
- **Optimal regularity in the Calculus of Variations**, Connor Mooney

The second week was devoted to research. During this week, three to four research lectures were held per day with the remainder of the time devoted to research collaboration and the general enjoyment of the beautiful Australian bushland surrounding the institute. Arising from the research workshop were several submissions included in these proceedings. The papers deal with topics such as variational problems, particularly non-linear geometric problems, optimal transport, regularity properties and spectral properties of elliptic operators. Neil Trudinger, one of the two authors for whom are program honoured is famous for his work on such problems. As such, each submission represents a continuation of the legacy of the book “Elliptic partial differential equations of second order” and its continuing influence on mathematics.

- “Boundary regularity of mass-minimizing integral currents and a question of Almgren” by Camillo De Lellis, Guido De Philippis, Jonas Hirsch and Annalisa Massaccesi: This paper is an announcement of results to be published in detail in a forthcoming article. The results describe boundary regularity of area minimizing currents in high codimension ensuring that regular points are dense on the boundary and leading to a structure theorem answering in particular a question of Almgren, implying singular points on the boundary have low dimension, and yielding a monotonicity formula. The announced results, representing a continuation of Allard’s regularity theorem and the monumental “Big Regularity Paper” of Almgren were completed during the second week of our program.
- “Optimal transport with discrete mean field interaction” by Jiakun Liu and Grégoire Loeper: This paper is also an announcement of ongoing work motivated by the motion of self-gravitating matter governed by the Euler-Poisson system. Here the authors build on the first author’s formulation of the problem as a variational problem which is then solved using optimal transport techniques exploiting Monge-Kantorovich duality. The result considers a time-discretisation of more general variational problems obtaining regularity results.
- “A sixth order curvature flow of plane curves with boundary conditions” by James McCoy, Glen Wheeler and Yuhan Wu: This submission announces results for a high order curvature flow. Curvature flows, particularly those arising via variational principles have been extensively studied over the past 30 years. A prototypical example is the Mean Curvature Flow, a second order gradient flow, whilst the Wilmore flow is perhaps the most well known example of a higher gradient flow. The authors describe a sixth order gradient flow with free boundary arising in elasticity theory. The maximum principle arguments that feature heavily in second order flows are of no utility in higher order flows and must be replaced by other techniques. The authors employ a Poincaré inequality

and interpolation inequalities to develop suitable integral estimates leading to the conclusion that the flow smoothly converges to the optimal configuration.

- “Quasilinear parabolic and elliptic equations with singular potentials” by Maria Michaela Porzio: This paper considers quasilinear equations with singular, Hardy potentials. A well known result is that solutions to the heat equation with positive driving term and positive initial data do not exist for Hardy parameters larger than the optimal constant in Hardy’s inequality. This particular paper considers divergence form operators and in particular the asymptotic behaviour of solutions to such equations provided the Hardy parameter is sufficiently small. Unique global existence of solutions is obtained with decay estimates implying the asymptotic behaviour as $t \rightarrow \infty$ is independent of the initial data and hence uniqueness of the associated elliptic problem is assured.
- “How to hear the corners of a drum” by Medet Nursultanov, Julie Rowlett, and David Sher: This paper is a detailed announcement of ongoing work. A well known question asks if one can “hear the shape of a drum?”. The answer in general is no, there exists non-congruent domains for which the Laplacian has the same spectrum. The result of this paper says that smooth domains may be distinguished from domains with corners by the spectrum of the Laplacian. More precisely, the spectrum of the Laplacian on a smooth domain with either Neumann or Robin boundary conditions is never equal to that of the Laplacian on a domain with corners and either Neumann or Robin boundary conditions. The result hinges on a generalisation of Kac’s locality principle.

Special thanks go to Julie Clutterbuck for her wonderful depiction of a well worn copy of Gilbarg’s and Trudinger’s book featured on our program materials.

Paul Bryan
Guest Editor



Participants

Yann Bernard (Monash Uni), Norman Dancer (Uni Sydney), Camillo De Lellis (ETH Zurich), Guido de Philippis (SISSA Trieste), Serena Dipierro (Uni Melbourne), Yihong Du (Uni New England), Nicola Fusco (Uni Napoli), Jesse Gell-Redman (Uni Melbourne), Joseph Grotowski (Uni Queensland), Feida Jiang (Tsinghua Uni), Gary Lieberman (Iowa State Uni), Jiakun Liu (Uni Wollongong), Gregoire Loeper (Monash Uni), Connor Mooney (Uni Texas Austin), Aldo Pratelli (Erlangen-Nürnberg), Maria Michaela Porzio (Roma), Frédéric Robert (Uni Lorraine), Julie Rowlett (Chalmers Uni), Mariel Sáez (Pontificia Uni, Chile), Neil Trudinger (Australian National Uni), John Urbas (Australian National Uni), Jerome Vetois (McGill Uni), Xu-Jia Wang (Australian National Uni), Valentina Wheeler (Uni Wollongong), Bin Zhou (Australian National Uni), Graham Williams (Uni Wollongong), James McCoy (Uni Wollongong)

Combinatorics, Statistical Mechanics, and Conformal Field Theory

29 October–18 November 2017

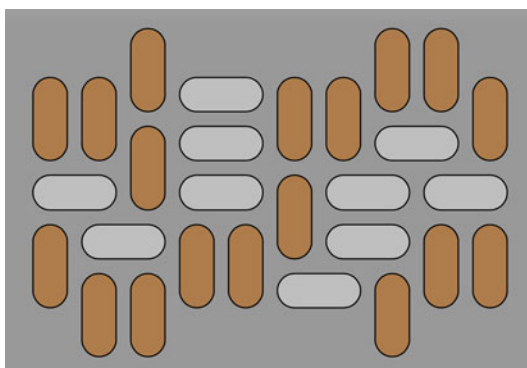
Organisers

Vladimir Korepin
Stony Brook Uni

Vladimir Mangazeev
Australian National Uni

Bernard Nienhuis
Uni Amsterdam

Jorgen Rasmussen
Uni Queensland



This program brought together leading experts in and around the area where statistical mechanics, integrability, conformal field theory, and combinatorics meet and in some sense overlap. A primary goal was to encourage research collaborations across the traditional divides between the research communities interested in the separate disciplines. Significant recent developments stem from this kind of cross fertilisation, and the aim was to cultivate further such collaborations and widen the scope of their successes.

The scientific presentations and discussions were largely centred around Yang-Baxter integrable models; the Razumov-Stroganov conjecture and generalisations thereof; combinatorial points and the role of supersymmetry in integrable lattice models and quantum chains; the combinatorics of spanning trees and pattern-avoiding permutations; and logarithmic conformal field theory.

With the strong emphasis on collaborations and discussions, there were only a couple of seminars per day in the first and third week. The embedded AMSI Workshop took place in the second week and included talks by Bazhanov, Guttman, Hagendorf, Mangazeev, Nienhuis, Pearce, Ridout, Ruelle, Tartaglia, Weston and Wheeler. Sessions with informal and brief presentations were held throughout the

program, with the aim to expand collaborative work on existing research projects and to foster new ideas and collaborations.

The contribution to this volume by Bernard Nienhuis and Kayed Al Qasimi was directly stimulated by discussions with Christian Hagendorf at the MATRIX Workshop. It provides a proof of a conjecture on certain one-point functions related to the Razumov-Stroganov conjecture.

Jorgen Rasmussen
Guest Editor



Participants

Georgy Feher (Budapest Uni Technology and Economics), Christian Hagendorf (Catholic Uni Louvain), Philippe Ruelle (Catholic Uni Louvain), Elena Tartaglia (SISSA Trieste), Murray Batchelor (Australian National Uni), Vladimir Bazhanov (Australian National Uni), Zeying Chen (Uni Melbourne), Omar Foda (Uni Melbourne), Jan de Gier (Uni Melbourne), Alexandr Garbali (Uni Melbourne), Tony Guttmann (Uni Melbourne), Jon Links (Uni Queensland), Paul Pearce (Uni Melbourne), Thomas Quella (Uni Melbourne), David Ridout (Uni Melbourne), Alessandra Vittorini-Orgeas (Uni Melbourne), Michael Wheeler (Uni Melbourne), Paul Zinn-Justin (Uni Melbourne), Robert Weston (Heriot-Watt Uni), Atsuo Kuniba (Tokyo Uni)

Mathematics of Risk

20 November–8 December 2017

Organisers

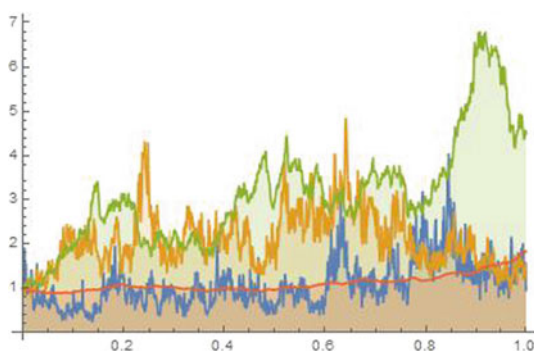
Kostya Borovkov
Uni Melbourne

Kais Hamza
Monash Uni

Masaaki Kijima
Tokyo Metropolitan Uni

Alexander Novikov
Uni Technology Sydney

Peter Taylor
Uni Melbourne



The mathematical modelling of the various types of risk modern society encounters at different levels of its operation has become an important part of applied mathematics as well as a source of challenging theoretical problems. The main need in modelling risk is where the latter refers to a serious danger to society and nature. As illustrated by the recent Global Financial Crisis of 2007–2008, the finance industry is one of the most serious sources of risk. Since the finance industry tried to (at least, partly) blame mathematical models for what had happened, it is all the more important for mathematicians to address the issue of financial risk and use mathematics to find ways to mitigate it.

The need for quantitative risk modelling has, in recent years, attracted enormous worldwide attention. The risk related to both extreme and non-extreme events is generating a vast research activity, which is international by its very nature. Moreover, there is an international regulatory aspect concerning mathematical modelling of financial risks. One of the key elements of the current versions of the Basel accord (a global regulatory framework for bank capital adequacy,

stress testing, and market liquidity risk) is the emphasis on responsible use of mathematical models.

Our program mostly addressed various aspects of mathematical modelling and subsequent analysis of risks related to activities in the finance industry and, more generally, economics. Major attention was also paid to studying the mathematical theory that can be used to model more general types of risk, related to chronic and long-term hazards and extremes, as well as the interplay between them.

The key themes of the program included:

- the modelling of uncertainty and risk events using the theory of stochastic processes, in particular, the evaluation of the distributions of boundary functionals of random processes, including the computation of boundary crossing probabilities;
- new methods for and approaches to computing the prices of financial derivatives;
- the systemic risk, including the stability of national and world financial systems, consequences for the markets from the wide use of algorithmic trading, network modelling of relevant real-life systems;
- risk modelling and quantification, including risk measures;
- the analysis of model risk, i.e. the type of risk that arises due to using inappropriate mathematical models for asset price dynamics etc.;
- mathematical modelling of extreme events due to factors such as natural disasters, human errors, infrastructure and computer control systems' failures.

Our program included two 'embedded events'. In the first week of the program, we ran four 5-h workshops for PhD students, research workers and industry specialists on the following topics:

- Extreme Value Theory—Applications to risk analysis (M. Kratz);
- Financial measures of risk and performance (M. Zhitlukhin);
- Ruin probabilities: exact and asymptotic results (Z. Palmowski);
- Clearing in financial networks (Yu. Kabanov).

In the second week of the program, we hosted a research conference where about 20 talks were given. The slides used by both the workshop presenters and conference speakers are available at the program web-site, <https://www.matrix-inst.org.au/events/mathematics-of-risk>. For the present volume, two of the workshop presenters (M. Kratz and M. Zhitlukhin) prepared more detailed expositions of the material from their workshops. We are most grateful to them for their time and effort required to write these very interesting and instructive papers. Our thanks also go to the other program participants who took time to contribute to this volume. Finally, we would like to thank the MATRIX and Creswick campus staff for facilitating and hosting this event. The participants enjoyed it tremendously. We had an excellent opportunity to engage in joint research work and exchange our ideas, both during the conference week and outside it.

Alexander Novikov and Kostya Borovkov
Guest Editors



Participants

Martin Larsson, Michael Zhitlukhin, Yuri Kabanov, Zbigniew Palmowski, Marie Kratz, Ljudmila Vostrikova, Budhi Surya, Jie Xiong, Kazutoshi Yamazaki, Eugene Feinberg, Jun Sekine, Takashi Shibata, Katsumasa Nishide, Daniel Dufresne, Aihua Xia, Yan Dolinsky, Thomas Taimre, Boris Buchmann, Gregoire Loeper, Alexander Buryak, Koji Anamizu, Hitomi Ito, Nakamura Eri, Yusuke Komatsu, Yasuhiro Shimizu, Kyoko Yagi, Tadao Oryu, Evgeny Prokopenko, Peter Spreij, Jeremy Baker, Aaron Chong, Tianshu Cong, Gurtek Ricky Gill, Qingwei Liu, Jin Sun, Priyanga Dilini Talagala, Hui Yao (Alice), Junyu (Adam) Nie, Eduard Biche, Kaustav Das, Meng Shi, Yunxuan Liu, Yuqing Pan, Wei Ning, Jason Leung, Igor Vladimirov, Libo Li, Peter Straka, Michael Callan, Jaden Li, Nino Kordzakhia, Juri Hinz, Fima Klebaner, Nicholas Read, Kevin Vo, Zhehao Zhang

Tutte Centenary Retreat

26 November–2 December 2017

Organisers

Graham Farr (Chair)
Monash Uni

Marston Conder
Uni Auckland

Dillon Mayhew
Victoria Uni Wellington

Kerri Morgan
Monash Uni

James Oxley
Louisiana State Uni

Gordon Royle
Uni Western Australia



FIG. 9



The year 2017 marked the centenary of the birth of W.T. (Bill) Tutte (1917–2002), the great Bletchley Park cryptologist and pioneering graph theorist. This Retreat was part of a worldwide programme of Tutte Centenary events, see <https://billtuttememorial.org.uk/centenary/>, including events at Bletchley Park, Waterloo, Cambridge, and Monash. It was scheduled for the week preceding the 5th International Combinatorics Conference (5ICC) (<http://www.monash.edu/5icc/>) at Monash.

The Retreat programme focused on three topics that have grown out of seminal contributions made by Tutte at the very start of his career:

Tutte-Whitney Polynomials. These count a wide variety of structures associated with a graph, and are related to network reliability, coding theory, knot theory and statistical physics. They were introduced by Whitney (1932) and Tutte (1947,

1954), and now play a central role in enumerative graph theory. They extend readily to matroids, which were the focus of the second topic.

Matroid Structure Theory. Many aspects of graph theory are especially natural and elegant when viewed in the broader setting of *matroids*, which are combinatorial abstractions of sets of vectors under linear independence. Tutte developed the theory of matroid connectivity, and characterised several important matroid classes in terms of forbidden substructures (*excluded minors*). His work continues to inspire developments in the field.

Symmetric Graphs. A lot of recent research on symmetric graphs builds on ground-breaking theory by Tutte on the trivalent case. Tutte developed the theory of arc-transitive graphs, and the techniques he used formed the foundations of a large and growing branch of discrete mathematics.

The Retreat emphasised collaborative research supported by problem sessions. There were three introductory talks: an overview of Tutte's contributions to mathematics, by James Oxley; Tutte-Whitney polynomials, by Gordon Royle; and symmetric graphs, by Marston Conder and Michael Giudici. Oxley's talk led to the paper 'The contributions of W.T. Tutte to matroid theory' by Graham Farr and James Oxley, included in this volume.

We had a total of 32 participants (including organisers). Participants found the workshop to be an exceptionally stimulating event precisely because, instead of hearing a long sequence of talks about the work of others, they got to work for extended periods on interesting problems with a variety of collaborators. They were able to develop new ideas and learn a lot about other recent work. A number of questions were answered relatively quickly, simply through sharing knowledge among participants. The more substantial research collaborations in each of the three themes dealt with the following.

Tutte-Whitney Polynomials

- An old question of Hassler Whitney (1932) about an elegant extension of the four-colour theorem using duality and Tutte-Whitney polynomials.
- Chromatic polynomial of hypergraphs (with the research having been done and largely written up during the workshop).
- A notion of "rank function" for certain algebraic structures which exhibit a pleasing duality, with the potential to generalise matroid representability and shed light on Tutte polynomials of other combinatorial objects.
- One of the Merino-Welsh conjectures (correlation inequalities relating numbers of acyclic orientations and totally cyclic orientations of a graph).
- The complexity of counting bases in binary matroids.

Matroid Structure Theory

- A problem of connectivity in frame matroids, a generalisation of the matroids that arise from graphs.

- A problem about the relationship between size and rank in matroids, inspired by an old theorem of Edmonds.
- Analysis and generalisation of a distinctive property of wheels and whirls, matroids Tutte identified as playing a fundamental role in his theory of 3-connectivity for matroids.
- Characterisation of the members of a natural class of matroids where the interaction of elements, circuits and cocircuits is particularly elegant and symmetric.

Symmetric Graphs

- Normal quotient analysis for finite edge-transitive oriented graphs of valency four. This led to the paper ‘Biquasiprimitive oriented graphs of valency four’ by Nemanja Poznanovic and Cheryl Praeger in this volume.
- A question of Caprace (at Groups St Andrews, Birmingham, August 2017) on whether there exists a 2-transitive permutation group P such that only finitely many simple groups act arc-transitively on a connected graph X with local action P . Marston Conder gave a partial answer in his paper ‘Simple group actions on arc-transitive graphs with prescribed transitive local action’ in this volume.
- Answer to a question by Folkman (1967) about (bipartite) semi-symmetric graphs of order $2n$ and valency $d \geq n/2$, to be published by Marston Conder and Gabriel Verret.
- Development of the theory of LR-structures, which in some sense extend Tutte’s work on symmetric 3-valent graphs, and the answer to an open question on them.
- A question related to determining the “graph-type” of a larger number of transitive groups. This proved quite difficult, but was solved soon after the Retreat in joint work with someone who could not attend.

It is expected that about fifteen papers will result from research that began during the workshop week, including the three in this volume of *MATRIX Annals*.

We gratefully acknowledge sponsorship by the Faculty of Information Technology, Monash University.

Graham Farr, Marston Conder, Dillon Mayhew and Gordon Royle
Guest Editors



Participants

Joanna Ellis-Monaghan (Saint Michael’s College, Vermont), Johann A. Makowsky (Technion, Israel), Steven Noble (Birkbeck, Uni London), Deborah Chun (West Virginia Uni Technology), Charles Semple (Uni Canterbury), Geoff Whittle (Victoria Uni Wellington), Nick Brettell (Victoria Uni Wellington), Joanna Fawcett (Uni Cambridge), Michael Giudici (Uni Western Australia), Cheryl Praeger (Uni Western Australia), Cai Heng Li (South Uni Science and Technology, Shenzhen), Gabriel Verret (Uni Auckland), Luke Morgan (Uni Western Australia), Primož Potočnik (Uni Ljubljana), Sanming Zhou (Uni Melbourne), Kevin Grace (Louisiana State Uni), Iain Moffatt (Royal Holloway, Uni London), Xian’an Jin (Xiamen Uni), Thomas Britz (UNSW Sydney), Tara Fife (Louisiana State Uni), Nemanja Poznanovic (Uni Melbourne), William Whistler (Durham Uni), Georgina Liversidge (Uni Newcastle), Rutger Campbell (Uni Waterloo), Keisuke Shiromoto (Kumamoto Uni), Ruixue Zhang (Nanyang Technological Uni), Ben Jones (Monash Uni)

Geometric R-Matrices: From Geometry to Probability

17–22 December 2017

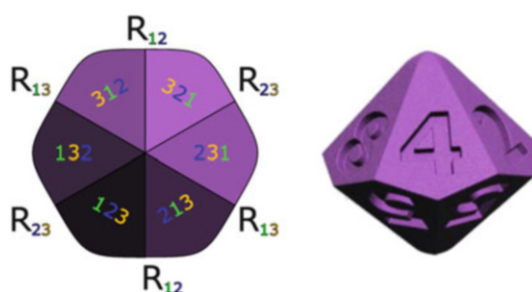
Organisers

Rick Kenyon
Brown Uni

Andrei Okounkov
Columbia Uni

Anita Ponsaing
Uni Melbourne

Paul Zinn-Justin
Uni Melbourne



The focus of this workshop was originally, as the title “Geometric R-matrices” suggests, to discuss the interaction of quantum integrable systems, a topic in mathematical physics, with algebraic geometry and representation theory. As the subtitle “from geometry to probability” indicates, it was quickly expanded to include interactions with other branches of mathematics, in particular combinatorics and probability. Here is a brief sketch of these interactions:

Algebraic Geometry and Representation Theory In the 2000s, the idea emerged that the theory of quantum integrable systems could be used to study the (quantum, equivariant) cohomology of certain varieties that appear naturally in algebraic geometry and representation theory, such as Grassmannians, flag varieties, and related Schubert varieties, orbital varieties, etc. This was reformulated as a beautiful, coherent program by Maulik and Okounkov in the early 2010s, combining ideas from geometric representation theory and in particular Nakajima’s quiver varieties, other ideas from geometry (Gromov–Witten invariants, etc.), with concepts coming from the study of supersymmetric gauge theories (cf. the work of Nekrasov and Shatashvili). The quantum integrable system is defined out of the geometry by starting with its building block which is the R-matrix of our title.

This area was represented during the week by the two mini-courses “Schubert calculus and integrability” by A. Knutson and “Geometric representation theory and quantum integrable systems” by A. Okounkov, as well by several talks.

Combinatorics and Probability Theory There is a large literature, which is still rapidly expanding, on the use of quantum integrable methods to study problems of a combinatorial or probabilistic nature. Some relevant milestones are:

- Kuperberg’s proof of the Alternating Sign Matrix conjecture using the integrability of the six-vertex model in the late 90s, followed by numerous generalizations and variations in the 2000s;
- the enumeration of Plane Partitions, and more the study of the related lozenge tilings as a probabilistic model, using free fermionic techniques (the simplest “integrable” model) in the 2000s, in particular in the work of R. Kenyon, and their extension to non free-fermionic integrable models;
- the connection of integrability and cluster algebras, which remains partially mysterious, though much progress has been done recently in that area; and
- the field of “integrable probability”, closely connected to the subject of this workshop but which has become sufficiently wide in itself that it had its own separate MATRIX program in January 2018.

These topics were all present during this workshop via talks of our participants.

The two contributions in this volume reflect this diversity of subjects. On the one hand, Y. Yang and G. Zhao’s lecture notes “How to sheafify an elliptic quantum group” belong to the first type of interaction, emphasizing the use of cohomological Hall algebras to build geometrically the underlying “symmetry algebras” of quantum integrable systems, namely quantum groups, and applying this construction to elliptic cohomology. On the other hand, G. Koshevoy’s article “Cluster decorated geometric crystals, generalized geometric RSK-correspondences, and Donaldson-Thomas transformations” is of a more combinatorial nature, developing interesting new concepts in the theory of cluster algebras and geometric crystals.

Paul Zinn-Justin
Guest Editor



Participants

Allen Knutson, Gus Schrader, Ole Warnaar, Vassily Gorbounov, Jan de Gier, Alisa Knizel, Olya Mandelshtam, Greta Panova, Hitoshi Konno, Yaping Yang, Gleb Koshevoy, Travis Scrimshaw, Alexandr Garbali, Pak-Hin Li, Hannah Keese, Alessandra Vittorini Orgeas

Contents

PART I REFEREED ARTICLES

Computational Inverse Problems

A Metropolis-Hastings-Within-Gibbs Sampler for Nonlinear Hierarchical-Bayesian Inverse Problems	3
Johnathan M. Bardsley and Tiangang Cui	
Sequential Bayesian Inference for Dynamical Systems Using the Finite Volume Method	13
Colin Fox, Richard A. Norton, Malcolm E. K. Morrison, and Timothy C. A. Molteno	
Correlation Integral Likelihood for Stochastic Differential Equations	25
Heikki Haario, Janne Hakkarainen, Ramona Maraia, and Sebastian Springer	
A Set Optimization Technique for Domain Reconstruction from Single-Measurement Electrical Impedance Tomography Data	37
Bastian Harrach and Janosch Rieger	
Local Volatility Calibration by Optimal Transport	51
Ivan Guo, Grégoire Loeper, and Shiyi Wang	
Likelihood Informed Dimension Reduction for Remote Sensing of Atmospheric Constituent Profiles	65
Otto Lamminpää, Marko Laine, Simo Tukiainen, and Johanna Tamminen	
Wider Contours and Adaptive Contours	79
Shev MacNamara, William McLean, and Kevin Burrage	
Bayesian Point Set Registration	99
Adam Spannaus, Vasileios Maroulas, David J. Keffer, and Kody J. H. Law	

Optimization Methods for Inverse Problems	121
Nan Ye, Farbod Roosta-Khorasani, and Tiangang Cui	
Integrability in Low-Dimensional Quantum Systems	
Diagonal Form Factors from Non-diagonal Ones	141
Zoltan Bajnok and Chao Wu	
Narayana Number, Chebyshev Polynomial and Motzkin Path on RNA Abstract Shapes	153
Sang Kwan Choi, Chaiho Rim, and Hwajin Um	
A Curious Mapping Between Supersymmetric Quantum Chains	167
Gyorgy Z. Feher, Alexandr Garbali, Jan de Gier, and Kareljan Schoutens	
Remarks on $A_n^{(1)}$ Face Weights	185
Atsuo Kuniba	
Elliptic Partial Differential Equations of Second Order: Celebrating 40 Years of Gilbarg and Trudinger's Book	
Boundary Regularity of Mass-Minimizing Integral Currents and a Question of Almgren	193
Camillo De Lellis, Guido De Philippis, Jonas Hirsch, and Annalisa Massaccesi	
Optimal Transport with Discrete Mean Field Interaction	207
Jiakun Liu and Grégoire Loeper	
A Sixth Order Curvature Flow of Plane Curves with Boundary Conditions	213
James McCoy, Glen Wheeler, and Yuhan Wu	
Quasilinear Parabolic and Elliptic Equations with Singular Potentials ...	223
Maria Michaela Porzio	
How to Hear the Corners of a Drum	243
Medet Nursultanov, Julie Rowlett, and David Sher	
Mathematics of Risk	
Nonparametric Bayesian Volatility Estimation	279
Shota Gugushvili, Frank van der Meulen, Moritz Schauer, and Peter Spreij	
The Exact Asymptotics for Hitting Probability of a Remote Orthant by a Multivariate Lévy Process: The Cramér Case	303
Konstantin Borovkov and Zbigniew Palmowski	

**Parisian Excursion Below a Fixed Level from the Last Record
Maximum of Lévy Insurance Risk Process**..... 311
Budhi A. Surya

Tutte Centenary Retreat

**Simple Group Actions on Arc-Transitive Graphs with Prescribed
Transitive Local Action**..... 327
Marston Conder

Biquasiprimitive Oriented Graphs of Valency Four 337
Nemanja Poznanović and Cheryl E. Praeger

The Contributions of W.T. Tutte to Matroid Theory..... 343
Graham Farr and James Oxley

Geometric R-Matrices: from Geometry to Probability

**Cluster Decorated Geometric Crystals, Generalized Geometric
RSK-Correspondences, and Donaldson-Thomas Transformations** 363
Gleb Koshevoy

PART II OTHER CONTRIBUTED ARTICLES

**Hypergeometric Motives and Calabi–Yau Differential
Equations**

Fields of Definition of Finite Hypergeometric Functions 391
Frits Beukers

L-Series and Feynman Integrals 401
David Broadhurst and David P. Roberts

**Arithmetic Properties of Hypergeometric Mirror Maps
and Dwork’s Congruences** 405
Éric Delaygue

**Appell–Lauricella Hypergeometric Functions over Finite Fields,
and a New Cubic Transformation Formula** 417
Sharon Frechette, Holly Swisher, and Fang-Ting Tu

Sequences, Modular Forms and Cellular Integrals 423
Dermot McCarthy, Robert Osburn, and Armin Straub

Some Supercongruences for Truncated Hypergeometric Series 425
Ling Long and Ravi Ramakrishna

The Explicit Formula and a Motivic Splitting 429
David P. Roberts

Hypergeometric Supercongruences 435
 David P. Roberts and Fernando Rodriguez Villegas

Alternate Mirror Families and Hypergeometric Motives..... 441
 Charles F. Doran, Tyler L. Kelly, Adriana Salerno, Steven Sperber,
 John Voight, and Ursula Whitcher

Schwarzian Equations and Equivariant Functions 449
 Abdellah Sebbar

Hypergeometric Functions over Finite Fields..... 461
 Jenny Fuselier, Ling Long, Ravi Ramakrishna, Holly Swisher,
 and Fang-Ting Tu

**Supercongruences Occurred to Rigid Hypergeometric Type
 Calabi–Yau Threefolds** 467
 Ling Long, Fang-Ting Tu, Noriko Yui, and Wadim Zudilin

***p*-Adic Hypergeometrics** 471
 Fernando Rodriguez Villegas

On *p*-Adic Unit-Root Formulas 475
 Masha Vlasenko

Triangular Modular Curves 481
 John Voight

Jacobi Sums and Hecke Grössencharacters..... 485
 Mark Watkins

**Special Values of Hypergeometric Functions and Periods of CM
 Elliptic Curves**..... 487
 Yifan Yang

CY-Operators and L-Functions 491
 Duco van Straten

Computational Inverse Problems

A Matrix Theoretic Derivation of the Kalman Filter 505
 Johnathan M. Bardsley

**Approximate Bayesian Computational Methods for the Inference
 of Unknown Parameters**..... 515
 Yuqin Ke and Tianhai Tian

**Combinatorics, Statistical Mechanics, and Conformal
 Field Theory**

**The Loop-Weight Changing Operator in the Completely Packed
 Loop Model** 531
 Bernard Nienhuis and Kayed Al Qasimi

Mathematics of Risk

A Note on Optimal Double Spending Attacks 545
 Juri Hinz and Peter Taylor

Stochastic Maximum Principle on a Continuous-Time Behavioral Portfolio Model 553
 Qizhu Liang and Jie Xiong

Number of Claims and Ruin Time for a Refracted Risk Process 559
 Yanhong Li, Zbigniew Palmowski, Chunming Zhao, and Chunsheng Zhang

Numerical Approximations to Distributions of Weighted Kolmogorov-Smirnov Statistics via Integral Equations 579
 Dan Wu, Lin Yee Hin, Nino Kordzakhia, and Alexander Novikov

Introduction to Extreme Value Theory: Applications to Risk Analysis and Management 591
 Marie Kratz

Monotone Sharpe Ratios and Related Measures of Investment Performance 637
 Mikhail Zhitlukhin

On Chernoff’s Test for a Fractional Brownian Motion 667
 Alexey Muravlev and Mikhail Zhitlukhin

Geometric R-Matrices: from Geometry to Probability

How to Sheafify an Elliptic Quantum Group 675
 Yaping Yang and Gufang Zhao

Part I
Refereed Articles

A Metropolis-Hastings-Within-Gibbs Sampler for Nonlinear Hierarchical-Bayesian Inverse Problems



Johnathan M. Bardsley and Tiangang Cui

Abstract We investigate the use of the randomize-then-optimize (RTO) method as a proposal distribution for sampling posterior distributions arising in nonlinear, hierarchical Bayesian inverse problems. Specifically, we extend the hierarchical Gibbs sampler for linear inverse problems to nonlinear inverse problems by embedding RTO-MH within the hierarchical Gibbs sampler. We test the method on a nonlinear inverse problem arising in differential equations.

1 Introduction

In this paper, we focus on inverse problems of the form

$$\mathbf{y} = \mathbf{F}(\mathbf{u}) + \mathbf{e}, \quad \mathbf{e} \sim \mathcal{N}(\mathbf{0}, \lambda^{-1}\mathbf{I}), \quad (1)$$

where $\mathbf{y} \in \mathbb{R}^m$ is the observed data, $\mathbf{u} \in \mathbb{R}^n$ is the unknown parameter, $\mathbf{F} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is the forward operator, and λ is known as the measurement precision parameter. The likelihood function then has the form

$$p(\mathbf{y}|\mathbf{u}, \lambda) = (2\pi)^{-\frac{m}{2}} \lambda^{m/2} \exp\left(-\frac{\lambda}{2} \|\mathbf{F}(\mathbf{u}) - \mathbf{y}\|^2\right). \quad (2)$$

J. M. Bardsley
Department of Mathematical Sciences, University of Montana, Missoula, MT, USA
e-mail: bardsleyj@mso.umt.edu

T. Cui (✉)
School of Mathematical Sciences, Monash University, Clayton, VIC, Australia
e-mail: tiangang.cui@monash.edu

Next, we assume that the prior is a zero-mean Gaussian random vector, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, (\delta \mathbf{L})^{-1})$, which has distribution

$$p(\mathbf{u}|\delta) = (2\pi)^{-\frac{n}{2}} \delta^{\bar{n}/2} \exp\left(-\frac{\delta}{2} \mathbf{u}^T \mathbf{L} \mathbf{u}\right), \quad (3)$$

where \mathbf{L} is defined via a Gaussian Markov random field (GMRF) [1], and \bar{n} is the rank of \mathbf{L} . In the one-dimensional numerical example considered at the end of the paper, we choose \mathbf{L} to be a discretization of the negative-Laplacian operator. The hyper-parameter δ , which is known as the prior precision parameter, provides the relative weight given to the prior as compared to the likelihood function.

In keeping with the Bayesian paradigm, we assume hyper-priors $p(\lambda)$ and $p(\delta)$ on λ and δ , respectively. A standard choice in the linear Gaussian case is to choose Gamma hyper-priors:

$$p(\lambda) \propto \lambda^{\alpha_\lambda - 1} \exp(-\beta_\lambda \lambda), \quad (4)$$

$$p(\delta) \propto \delta^{\alpha_\delta - 1} \exp(-\beta_\delta \delta). \quad (5)$$

This is due to the fact that the conditional densities for λ and δ are then also Gamma-distributed (a property known as conjugacy), and hence are easy to sample from. We choose hyper-parameters $\alpha_\lambda = \alpha_\delta = 1$ and $\beta_\lambda = \beta_\delta = 10^{-4}$, making the hyper-priors exponentially distributed with small decay parameters β_λ and β_δ . In the test cases we have considered, these hyper-priors work well, though they should be chosen carefully in a particular situation. Specifically, it is important that they are chosen to be relatively flat over the regions of high probability for λ and δ defined by the posterior density function, so that they are not overly informative.

Taking into account the likelihood, the prior, and the hyper-priors, the posterior probability density function over all of the unknown parameters is given, by Bayes' law, as

$$\begin{aligned} & p(\mathbf{u}, \lambda, \delta | \mathbf{y}) \\ &= p(\mathbf{y} | \mathbf{u}, \lambda) p(\mathbf{u} | \delta) p(\lambda) p(\delta) / p(\mathbf{y}) \\ &\propto \lambda^{m/2 + \alpha_\lambda - 1} \delta^{\bar{n}/2 + \alpha_\delta - 1} \exp\left(-\frac{\lambda}{2} \|\mathbf{F}(\mathbf{u}) - \mathbf{y}\|^2 - \frac{\delta}{2} \mathbf{u}^T \mathbf{L} \mathbf{u} - \beta_\lambda \lambda - \beta_\delta \delta\right), \quad (6) \end{aligned}$$

where $p(\mathbf{y})$ is the normalizing constant for the posterior. Our focus in this paper is to develop a Gibbs sampler for sampling from the full posterior (6). For this, we

need the full conditionals, which are given by

$$p(\lambda|\mathbf{b}, \mathbf{u}, \delta) \propto \lambda^{m/2+\alpha_\lambda-1} \exp\left(\left[-\frac{1}{2}\|\mathbf{F}(\mathbf{u}) - \mathbf{y}\|^2 - \beta_\lambda\right]\lambda\right), \quad (7)$$

$$p(\delta|\mathbf{y}, \mathbf{u}, \lambda) \propto \delta^{\bar{n}/2+\alpha_\delta-1} \exp\left(\left[-\frac{1}{2}\mathbf{u}^T \mathbf{L} \mathbf{u} - \beta_\delta\right]\delta\right), \quad (8)$$

$$p(\mathbf{u}|\mathbf{y}, \lambda, \delta) \propto \exp\left(-\frac{\lambda}{2}\|\mathbf{F}(\mathbf{u}) - \mathbf{y}\|^2 - \frac{\delta}{2}\mathbf{u}^T \mathbf{L} \mathbf{u}\right). \quad (9)$$

The Gamma-hyper priors are conjugate, and hence the conditional densities for λ and δ are also Gamma-distributed:

$$\lambda|\mathbf{u}, \delta, \mathbf{b} \sim \Gamma\left(m/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{F}(\mathbf{u}) - \mathbf{b}\|^2 + \beta_\lambda\right), \quad (10)$$

$$\delta|\mathbf{u}, \lambda, \mathbf{b} \sim \Gamma\left(\bar{n}/2 + \alpha_\delta, \frac{1}{2}\mathbf{u}^T \mathbf{L} \mathbf{u} + \beta_\delta\right). \quad (11)$$

The distributions (10) and (11) are independent so that $p(\lambda, \delta|\mathbf{b}, \mathbf{x}) = p(\lambda|\mathbf{b}, \mathbf{x})p(\delta|\mathbf{b}, \mathbf{x})$. Hence, computing independent samples from (10) and (11) yields a sample from $p(\lambda, \delta|\mathbf{b}, \mathbf{x})$. Moreover, in the linear case, \mathbf{F} is a matrix and the Gaussian prior is also conjugate, leading to a Gaussian conditional (9), which can be equivalently expressed

$$\mathbf{u} \sim \mathcal{N}\left((\lambda\mathbf{F}^T \mathbf{F} + \delta\mathbf{L})^{-1}\lambda\mathbf{F}^T \mathbf{y}, (\lambda\mathbf{F}^T \mathbf{F} + \delta\mathbf{L})^{-1}\right).$$

Taking these observations all together leads to the two-stage Gibbs sampler given next, which is also presented in [1, 2].

The Hierarchical Gibbs Sampler, Linear Case

0. Initialize (λ_0, δ_0) , $\mathbf{u}^0 = (\lambda_0\mathbf{F}^T \mathbf{F} + \delta_0\mathbf{L})^{-1}\lambda_0\mathbf{F}^T \mathbf{y}$, set $k = 1$, define k_{total} .

1. Compute $(\lambda_k, \delta_k) \sim p(\lambda, \delta|\mathbf{y}, \mathbf{u}^{k-1})$ as follows.

a. Compute $\lambda_k \sim \Gamma\left(m/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{F}\mathbf{u}^{k-1} - \mathbf{y}\|^2 + \beta_\lambda\right)$.

b. Compute $\delta_k \sim \Gamma\left(\bar{n}/2 + \alpha_\delta, \frac{1}{2}(\mathbf{u}^{k-1})^T \mathbf{L} \mathbf{u}^{k-1} + \beta_\delta\right)$.

2. Compute $\mathbf{u}^k \sim \mathcal{N}\left((\lambda_k\mathbf{F}^T \mathbf{F} + \delta_k\mathbf{L})^{-1}\lambda_k\mathbf{F}^T \mathbf{y}, (\lambda_k\mathbf{F}^T \mathbf{F} + \delta_k\mathbf{L})^{-1}\right)$.

3. If $k = k_{\text{total}}$ stop, otherwise, set $k = k + 1$ and return to Step 1.

When \mathbf{F} is nonlinear, the conditional density $p(\mathbf{u}|\mathbf{y}, \lambda, \delta)$, defined in (9), is no longer Gaussian and cannot be sampled from it directly. To overcome this, we embed a Metropolis-Hastings (MH) step within step 2 of hierarchical Gibbs, as advocated in [4, Algorithm A.43]. For the MH proposal, we use the randomize-then-optimize (RTO) [3], and thus we begin in Sect. 2 by describing the RTO proposal. In Sect. 3, we describe RTO-MH and its embedding within hierarchical Gibbs for sampling

from the full posterior (6). Finally, we use RTO-MH-within-hierarchical Gibbs to sample from (6) in a specific nonlinear inverse problem arising in differential equations. Concluding remarks are provided in Sect. 5.

2 The Randomize-Then-Optimize Proposal Density

We first define the augmented forward model and observation taking the form

$$\mathbf{F}_{\lambda,\delta}(\mathbf{u}) \stackrel{\text{def}}{=} \begin{bmatrix} \lambda^{1/2}\mathbf{F}(\mathbf{u}) \\ \delta^{1/2}\mathbf{L}^{1/2}\mathbf{x} \end{bmatrix} \quad \text{and} \quad \mathbf{y}_{\lambda,\delta} \stackrel{\text{def}}{=} \begin{bmatrix} \lambda^{1/2}\mathbf{y} \\ \mathbf{0} \end{bmatrix}.$$

For motivation, note that in the linear case, $p(\mathbf{u}|\mathbf{y}, \lambda, \delta)$ is Gaussian and can be sampled by solving the stochastic least squares problem

$$\mathbf{u}|\mathbf{y}, \lambda, \delta = \arg \min_{\boldsymbol{\psi}} \|\mathbf{F}_{\lambda,\delta}\boldsymbol{\psi} - (\mathbf{y}_{\lambda,\delta} + \boldsymbol{\epsilon})\|^2, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (12)$$

This follows from the fact that if $\mathbf{F}_{\lambda,\delta} = \mathbf{Q}_{\lambda,\delta}\mathbf{R}_{\lambda,\delta}$ is the thin (or condensed) QR-factorization of $\mathbf{F}_{\lambda,\delta}$, and $\mathbf{F}_{\lambda,\delta}$ has full column rank, then $\mathbf{Q}_{\lambda,\delta} \in \mathbb{R}^{(M+N) \times N}$ has orthonormal columns spanning the column space of $\mathbf{F}_{\lambda,\delta}$; $\mathbf{R}_{\lambda,\delta} \in \mathbb{R}^{N \times N}$ is upper-triangular and invertible; and the solution of (12) is unique and can be expressed

$$\mathbf{Q}_{\lambda,\delta}^T \mathbf{F}_{\lambda,\delta}(\mathbf{u}|\mathbf{b}, \lambda, \delta) = \mathbf{Q}_{\lambda,\delta}^T (\mathbf{y}_{\lambda,\delta} + \boldsymbol{\epsilon}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (13)$$

Note that in the linear case $\mathbf{Q}_{\lambda,\delta}^T \mathbf{F}_{\lambda,\delta} = \mathbf{R}_{\lambda,\delta}$, and it follows that (13) yields samples from $p(\mathbf{u}|\mathbf{y}, \lambda, \delta)$.

In the nonlinear case, Eq. (13) can still be used, but the resulting samples do not have distribution $p(\mathbf{u}|\mathbf{y}, \lambda, \delta)$. To derive the form of the distribution, we first define

$$\mathbf{r}_{\lambda,\delta}(\mathbf{u}) \stackrel{\text{def}}{=} \mathbf{F}_{\lambda,\delta}(\mathbf{u}) - \mathbf{y}_{\lambda,\delta}$$

and denote the Jacobian of $\mathbf{F}_{\lambda,\delta}$, evaluated at \mathbf{u} , by $\mathbf{J}_{\lambda,\delta}(\mathbf{u})$. Then, provided $\mathbf{Q}_{\lambda,\delta}^T \mathbf{F}_{\lambda,\delta}$ is a one-to-one function with continuous first partial derivatives, and its Jacobian, $\mathbf{Q}_{\lambda,\delta}^T \mathbf{J}_{\lambda,\delta}$, is invertible, the probability density function for $\mathbf{u}|\mathbf{b}, \lambda, \delta$ defined by (13) is

$$\begin{aligned} p_{\text{RTO}}(\mathbf{u}|\mathbf{b}, \lambda, \delta) &\propto \left| \det(\mathbf{Q}_{\lambda,\delta}^T \mathbf{J}_{\lambda,\delta}(\mathbf{u})) \right| \exp\left(-\frac{1}{2} \|\mathbf{Q}_{\lambda,\delta}^T \mathbf{r}_{\lambda,\delta}(\mathbf{u})\|^2\right) \\ &= c_{\lambda,\delta}(\mathbf{x}) p(\mathbf{u}|\mathbf{b}, \lambda, \delta), \end{aligned} \quad (14)$$

where

$$c_{\lambda,\delta}(\mathbf{u}) = \left| \det(\mathbf{Q}_{\lambda,\delta}^T \mathbf{J}_{\lambda,\delta}(\mathbf{u})) \right| \exp \left(\frac{1}{2} \|\mathbf{r}_{\lambda,\delta}(\mathbf{u})\|^2 - \frac{1}{2} \|\mathbf{Q}_{\lambda,\delta}^T \mathbf{r}_{\lambda,\delta}(\mathbf{u})\|^2 \right). \quad (15)$$

There is flexibility in how to choose $\mathbf{Q}_{\lambda,\delta} \in \mathbb{R}^{(M+N) \times N}$, though $\mathbf{Q}_{\lambda,\delta}^T \mathbf{F}_{\lambda,\delta}$ must satisfy the conditions mentioned in the previous sentence. In our implementations of RTO, we have used $\mathbf{Q}_{\lambda,\delta}$ from the thin **QR**-factorization $\mathbf{J}_{\lambda,\delta}(\mathbf{u}_{\lambda,\delta}) = \mathbf{Q}_{\lambda,\delta} \mathbf{R}_{\lambda,\delta}$, where $\mathbf{u}_{\lambda,\delta}$ is the MAP estimator, i.e., $\mathbf{u}_{\lambda,\delta} = \arg \min_{\mathbf{u}} \|\mathbf{F}_{\lambda,\delta}(\mathbf{u}) - \mathbf{y}_{\lambda,\delta}\|^2$.

In practice, we compute samples from (13) by solving the stochastic optimization problem

$$\mathbf{u}^* = \arg \min_{\boldsymbol{\psi}} \frac{1}{2} \|\mathbf{Q}_{\lambda,\delta}^T (\mathbf{F}_{\lambda,\delta}(\boldsymbol{\psi}) - (\mathbf{y}_{\lambda,\delta} + \boldsymbol{\epsilon}^*))\|^2, \quad \boldsymbol{\epsilon}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (16)$$

The name randomize-then-optimize stems from (16), where $\mathbf{y}_{\lambda,\delta}$ is first ‘randomized’, by adding $\boldsymbol{\epsilon}^*$, and then ‘optimized’, by solving (16). Finally, we note that if the cost function minimum in (16) is greater than zero, (13) has no solution, and we must discard the corresponding sample. In practice, we discard solutions \mathbf{x}^* of (16) with a cost function minimum greater than $\eta = 10^{-8}$, though we have found this to occur very rarely in practice.

3 RTO-Metropolis-Hastings and Its Embedding Within Hierarchical Gibbs

Although RTO does not yield samples from $p(\mathbf{u}|\mathbf{y}, \lambda, \delta)$ for nonlinear problems, it can be used as a proposal for MH. At step k of the MH algorithm, given the current sample \mathbf{u}^{k-1} , one can use (16) to compute $\mathbf{u}^* \sim p_{\text{RTO}}(\mathbf{u}|\mathbf{b}, \lambda, \delta)$ and then set $\mathbf{u}^k = \mathbf{u}^{k-1}$ with probability

$$\begin{aligned} r_{\lambda,\delta} &= \min \left(1, \frac{p(\mathbf{u}^*|\mathbf{y}, \lambda, \delta) p_{\text{RTO}}(\mathbf{u}^{k-1}|\mathbf{y}, \lambda, \delta)}{p(\mathbf{u}^{k-1}|\mathbf{y}, \lambda, \delta) p_{\text{RTO}}(\mathbf{u}^*|\mathbf{y}, \lambda, \delta)} \right) \\ &= \min \left(1, \frac{p(\mathbf{u}^*|\mathbf{y}, \lambda, \delta) c_{\lambda,\delta}(\mathbf{u}^{k-1}) p(\mathbf{u}^{k-1}|\mathbf{y}, \lambda, \delta)}{p(\mathbf{u}^{k-1}|\mathbf{y}, \lambda, \delta) c_{\lambda,\delta}(\mathbf{u}^*) p(\mathbf{u}^*|\mathbf{y}, \lambda, \delta)} \right) \\ &= \min \left(1, \frac{c_{\lambda,\delta}(\mathbf{u}^{k-1})}{c_{\lambda,\delta}(\mathbf{u}^*)} \right). \end{aligned} \quad (17)$$

Note that it is often advantageous, for numerical reasons, to replace the ratio in (17) by the equivalent expression

$$c_{\lambda,\delta}(\mathbf{u}^{k-1})/c_{\lambda,\delta}(\mathbf{u}^*) = \exp \left(\ln c_{\lambda,\delta}(\mathbf{u}^{k-1}) - \ln c_{\lambda,\delta}(\mathbf{u}^*) \right),$$

where

$$\ln c_{\lambda,\delta}(\mathbf{u}) \simeq \ln \left| \mathbf{Q}_{\lambda,\delta}^T \mathbf{J}_{\lambda,\delta}(\mathbf{u}) \right| + \frac{1}{2} \|\mathbf{r}_{\lambda,\delta}(\mathbf{x})\|^2 - \frac{1}{2} \|\mathbf{Q}_{\lambda,\delta}^T \mathbf{r}_{\lambda,\delta}(\mathbf{x})\|^2,$$

and ‘ \simeq ’ denotes ‘equal up to an additive, unimportant constant.’

The RTO-MH Algorithm

1. Choose initial vector \mathbf{u}^0 , parameter $0 < \eta \ll 0$, and samples N . Set $k = 1$.
2. Compute $\mathbf{u}^* \sim p_{\text{RTO}}(\mathbf{u}|\mathbf{y}, \lambda, \delta)$ by solving (16) for a fixed realization $\boldsymbol{\epsilon}^* \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. If

$$\|\mathbf{Q}_{\lambda,\delta}^T (\mathbf{F}_{\lambda,\delta}(\mathbf{u}^*) - (\mathbf{y}_{\lambda,\delta} + \boldsymbol{\epsilon}^*))\|^2 > \eta,$$

then repeat step 2.

3. Set $\mathbf{u}^k = \mathbf{u}^*$ with probability $r_{\lambda,\delta}$ defined by (17). Else, set $\mathbf{u}^k = \mathbf{u}^{k-1}$.
4. If $k < N$, set $k = k + 1$ and return to Step 2, otherwise stop.

The proposed sample \mathbf{u}^* is independent of \mathbf{u}^{k-1} , making RTO-MH an independence MH method. Thus, we can apply [4, Theorem 7.8] to obtain the result that RTO-MH will produce a uniformly ergodic chain that converges in distribution to $p(\mathbf{u}|\mathbf{y}, \lambda, \delta)$ provided there exists $M > 0$ such that $p(\mathbf{u}|\mathbf{y}, \lambda, \delta) \leq M \cdot p_{\text{RTO}}(\mathbf{u}|\mathbf{y}, \lambda, \delta)$, for all $\mathbf{u} \in \mathbb{R}^N$. Given (14), this inequality holds if and only if $c_{\lambda,\delta}(\mathbf{u})$, defined by (15), is bounded away from zero for all \mathbf{u} .

3.1 RTO-MH-Within-Hierarchical Gibbs

In the hierarchical setting, we embed a single RTO-MH step within the hierarchical Gibbs sampler, to obtain the following MCMC method.

RTO-MH-Within-Hierarchical Gibbs

0. Initialize (λ_0, δ_0) , set $k = 1$, define k_{total} , and set

$$\mathbf{u}^0 = \arg \min_{\mathbf{u}} \|\mathbf{F}_{\lambda_0, \delta_0}(\mathbf{u}) - \mathbf{y}_{\lambda_0, \delta_0}\|^2.$$

1. Simulate $(\lambda_k, \delta_k) \sim p(\lambda, \delta|\mathbf{y}, \mathbf{u}^{k-1})$ as follows.
 - a. Compute $\lambda_k \sim \Gamma\left(m/2 + \alpha_\lambda, \frac{1}{2}\|\mathbf{F}(\mathbf{x}^{k-1}) - \mathbf{y}\|^2 + \beta_\lambda\right)$.
 - b. Compute $\delta_k \sim \Gamma\left(\bar{n}/2 + \alpha_\delta, \frac{1}{2}(\mathbf{x}^{k-1})^T \mathbf{L} \mathbf{x}^{k-1} + \beta_\delta\right)$.
2. Simulate u^k using RTO as follows.
 - a. Compute $\mathbf{u}^* \sim p_{\text{RTO}}(\mathbf{u}|\mathbf{y}, \lambda_k, \delta_k)$ by solving (16) with $(\lambda, \delta) = (\lambda_k, \delta_k)$.
 - b. Set $\mathbf{u}^k = \mathbf{u}^*$ with probability r_{λ_k, δ_k} defined by (17), else set $\mathbf{u}^k = \mathbf{u}^{k-1}$.
3. If $k = k_{\text{total}}$ stop, otherwise, set $k = k + 1$ and return to Step 1.

In step 2a, note that two optimization problems must be solved. First, the MAP estimator $\mathbf{u}_{\lambda_k, \delta_k}$ is computed; then the **QR**-factorization $\mathbf{J}(\mathbf{u}_{\lambda_k, \delta_k}) = \mathbf{Q}_{\lambda_k, \delta_k} \mathbf{R}_{\lambda_k, \delta_k}$ is computed; and finally, the stochastic optimization problem (16) is solved, with $(\lambda, \delta) = (\lambda_k, \delta_k)$, to obtain the RTO sample \mathbf{u}^* . One could take multiple RTO-MH steps in Step 2, within each outer loop, to improve the chances of updating \mathbf{u}^{k-1} , but we do not implement that here.

4 Numerical Experiment

To test RTO-MH-within-hierarchical Gibbs, we consider a nonlinear inverse problem from [1, Chapter 6]. The inverse problem is to estimate the diffusion coefficient $u(s)$ from measurements of the solution $x(s)$ of the Poisson equation

$$-\frac{d}{ds} \left(u(s) \frac{dx}{ds} \right) = f(s), \quad 0 < s < 1, \quad (18)$$

with zeros boundary conditions $x(0) = x(1) = 0$. Assuming a uniform mesh on $[0, 1]$, after numerical discretization, (18) takes the form

$$\mathbf{B}(\mathbf{u})\mathbf{x} = \mathbf{f}, \quad \mathbf{B}(\mathbf{u}) \stackrel{\text{def}}{=} \mathbf{D}^T \text{diag}(\mathbf{u})\mathbf{D}, \quad (19)$$

where $\mathbf{u} \in \mathbb{R}^n$ and $\mathbf{D} \in \mathbb{R}^{n \times n-1}$ is a discrete derivative matrix. To generate data, we compute numerical solutions corresponding to two discrete Dirac delta forcing functions, $f_1(s)$ and $f_2(s)$, centered at $s = 1/3$ and $s = 2/3$, respectively. After discretization, f_1 and f_2 become $(n-1) \times 1$ Kronecker delta vectors \mathbf{f}_1 and \mathbf{f}_2 , and the measurement model takes the form of (1) with

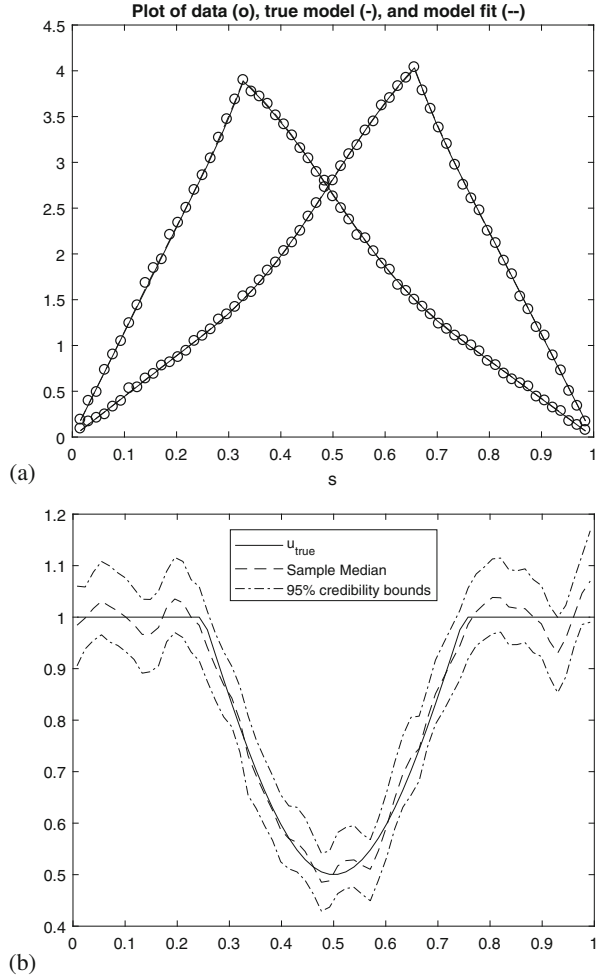
$$\mathbf{y} \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}_{2n-2} \quad \text{and} \quad \mathbf{F}(\mathbf{u}) \stackrel{\text{def}}{=} \begin{bmatrix} \mathbf{B}(\mathbf{u})^{-1} \mathbf{f}_1 \\ \mathbf{B}(\mathbf{u})^{-1} \mathbf{f}_2 \end{bmatrix}_{2n-2},$$

so that $m = 2n - 2$. We generate data using (1) with $n = 50$ and \mathbf{u}_{true} obtained by discretizing

$$u(s) = \min \{1, 1 - 0.5 \sin(2\pi(s - 0.25))\},$$

and λ^{-1} chosen so that signal-to-noise ratio, $\|\mathbf{F}(\mathbf{u}_{\text{true}})\|/\sqrt{m\lambda^{-1}}$, is 100. The data vectors \mathbf{y}_1 and \mathbf{y}_2 are plotted in Fig. 1a together with the noise-free data $\mathbf{B}(\mathbf{x})^{-1}\mathbf{f}_1$ and $\mathbf{B}(\mathbf{x})^{-1}\mathbf{f}_2$.

Fig. 1 (a) Plots of the measured data \mathbf{b}_1 and \mathbf{b}_2 , the true state \mathbf{u} , and the model fits. (b) Plots of the true diffusion coefficient \mathbf{x} together with the RTO-MH sample median and the element-wise 95% credibility bounds



With the measurements in hand, we implement RTO-MH-within-hierarchical Gibbs for sampling from (6). The results are plotted in Figs. 1 and 2. In Fig. 1b, we see the sample median together with 95% credibility intervals computed from the \mathbf{u} -chain generated by the MCMC method. In Fig. 2a, we plot the individual chains for λ , δ , and a randomly chosen element of the \mathbf{u} -chain. And finally, in Fig. 2b, we plot the auto correlation functions and associated integrated autocorrelation times (τ_{int}) for these three parameters [1].

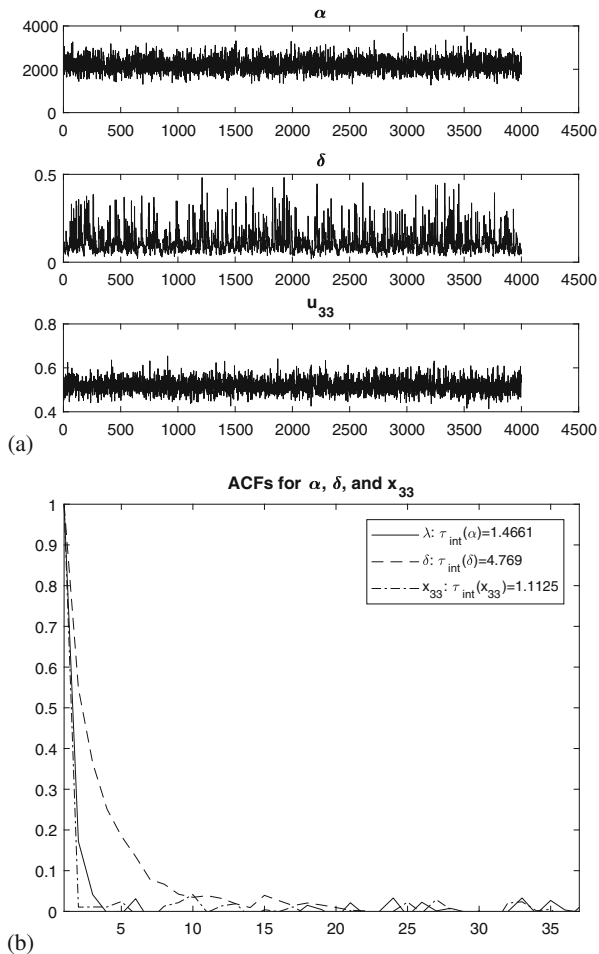


Fig. 2 (a) Plots of the chains for three randomly selected elements of \mathbf{x} . (b) Autocorrelation times associated with these chains

5 Conclusions

In this paper, we have tackled the problem of sampling from the full posterior (6) when \mathbf{F} is a nonlinear function. To do this, we followed the same approach as the hierarchical Gibbs algorithm of [2], however in that algorithm, \mathbf{F} is linear, the conditional density $p(\mathbf{u}|\mathbf{y}, \lambda, \delta)$ is Gaussian, and hence samples from $p(\mathbf{u}|\mathbf{y}, \lambda, \delta)$ can be computed by solving a linear system of equations. In the nonlinear case, $p(\mathbf{u}|\mathbf{y}, \lambda, \delta)$ is non-Gaussian, but we can use RTO-MH to obtain samples, as described in [3]. We obtain a MH-within-Gibbs method for sampling from (6) by embedding a single RTO-MH step with hierarchical Gibbs. We then tested the

method on a nonlinear inverse problem arising in differential equations and found that it worked well.

References

1. Bardsley, J.M.: Computational Uncertainty Quantification for Inverse Problems, vol. 19. SIAM (2018)
2. Bardsley, J.M.: MCMC-based image reconstruction with uncertainty quantification. *SIAM J. Sci. Comput.* **34**(3), A1316–A1332 (2012)
3. Bardsley, J.M., Solonen, A., Haario, H., Laine, M.: Randomize-then-optimize: a method for sampling from posterior distributions in nonlinear inverse problems. *SIAM J. Sci. Comput.* **36**(4), A1359–C399 (2014)
4. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, New York (2004)

Sequential Bayesian Inference for Dynamical Systems Using the Finite Volume Method



Colin Fox, Richard A. Norton, Malcolm E. K. Morrison,
and Timothy C. A. Molteno

Abstract Optimal Bayesian sequential inference, or filtering, for the state of a deterministic dynamical system requires simulation of the Frobenius-Perron operator, that can be formulated as the solution of an initial value problem in the continuity equation on filtering distributions. For low-dimensional, smooth systems the finite-volume method is an effective solver that conserves probability and gives estimates that converge to the optimal continuous-time values. A Courant–Friedrichs–Lewy condition assures that intermediate discretized solutions remain positive density functions. We demonstrate this finite-volume filter (FVF) in a simulated example of filtering for the state of a pendulum, including a case where rank-deficient observations lead to multi-modal probability distributions.

1 Introduction

In 2011, one of us (TCAM) offered to improve the speed and accuracy of the Tru-Test scales for ‘walk over weighing’ (WOW) of cattle, and wagered a beer on the outcome [8]. Tru-Test is a company based in Auckland, New Zealand, that manufactures measurement and productivity tools for the farming industry, particularly for dairy. Tru-Test’s XR3000 WOW system was already in the market, though they could see room for improvement in terms of on-farm usage, as well as speed and accuracy. Indeed, advertising material for the XR3000 stated that WOW

requires that the animals pass the platform regularly and smoothly

which hinted at the existing processing requiring somewhat constrained movement by the cows for it to deliver a weight.

C. Fox (✉) · R. A. Norton · M. E. K. Morrison · T. C. A. Molteno
University of Otago, Dunedin, New Zealand
e-mail: fox@physics.otago.ac.nz; richard.norton@otago.ac.nz; tim@physics.otago.ac.nz

Fig. 1 A dairy cow walking over a weigh bridge placed near the milking shed (photocredit: Wayne Johnson/Pizzini Productions)

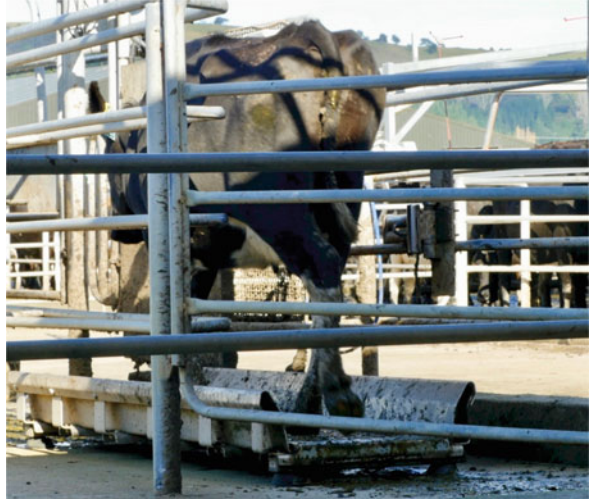


Figure 1 shows a cow walking over the weigh-bridge in the WOW system located on the ground in the path from a milking shed. The weigh bridge consists of a low platform with strain gauges beneath the platform, at each end, that are used to measure a time series of downward force from which weight (more correctly, mass) of the cow is derived.

The plan, for improving estimates of cow mass from strain-gauge time series, was to apply Bayesian modeling and computational inference. Bayesian inference allows uncertain measurements to be modeled in terms of probability distributions, and interpreted in terms of physical models that describe how the data is produced. This leads to estimates of parameters in the model, such as the mass of a cow, and meaningful uncertainty quantification on those estimates. At the outset we developed dynamical-systems models for the moving cow, with some models looking like one or more pogo sticks. Operation in real-time would require developing new algorithms for performing the inference sequentially—as the data arrives—and new hardware with sufficient computing speed to implement those algorithms. Figure 2 (right) shows hardware developed for this application, that includes strain-gauge signal conditioning, digitization, and an embedded ARM processor, alongside the XR3000 electronics and display (left).

This paper describes an algorithm for optimal sequential Bayesian inference that we developed in response to this application in cow weighing. We first give a stylized model of WOW, then a method for optimal filtering for tracking the state of a nonlinear dynamical system, then present numerical examples for the stylized model.

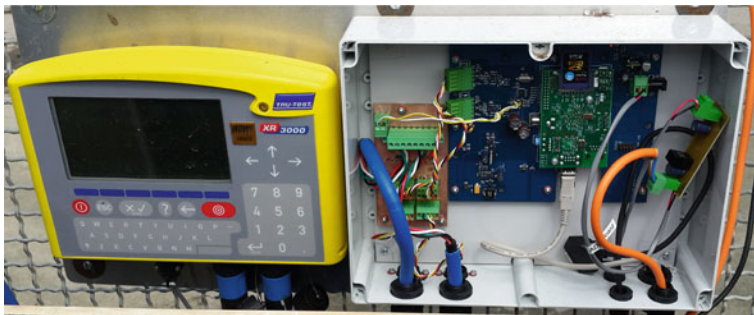


Fig. 2 WOW hardware in 2016: existing commercial unit (left), and prototype with embedded processing (right) (photocredit and hardware design: Phill Brown)

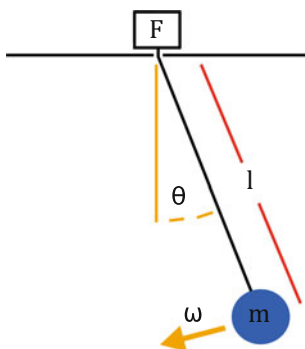


Fig. 3 A pendulum of length l with mass m , undergoing motion with angular displacement θ and angular velocity ω . The force F in the string is measured

1.1 A Stylized Problem

A (very) stylized model of WOW is the problem of tracking a simple pendulum of length l and mass m when only the force F in the string is measured, as depicted in Fig. 3. For this system the kinematic variables are the angular displacement (from the vertical downwards) θ and the angular velocity ω . The kinematic state (θ, ω) evolves according to

$$\frac{d}{dt}(\theta, \omega) = \left(\omega, -\frac{g}{l} \sin \theta \right)$$

where g is the acceleration due to gravity, l the length of the pendulum.

The force F_k is measured at times $t_k, k = 1, 2, 3, \dots$, with the (noise free) value related to the state variables by

$$F_k = ml\omega^2(t_k) + mg \cos \theta(t_k).$$

Estimation of the parameter m may be performed by considering the augmented system

$$\frac{d}{dt}(\theta, \omega, m) = \left(\omega, -\frac{g}{l} \sin \theta, 0 \right).$$

See [7] for a computed example of parameter estimation in this system.

2 Sequential Bayesian Inference for Dynamical Systems

Consider now a general dynamical system that evolves according to the (autonomous) differential equation

$$\frac{d}{dt}\mathbf{x} = \mathbf{f}(\mathbf{x}), \tag{1}$$

where \mathbf{f} is a known velocity field and $\mathbf{x}(t)$ is the state vector of the system at time t . Given an initial state $\mathbf{x}(0) = \mathbf{x}_0$ at $t = 0$, Eq. (1) may be solved to determine the future state $\mathbf{x}(t), t > 0$, that we also write $\mathbf{x}(t; \mathbf{x}_0)$ to denote this deterministic solution.

At increasing discrete times $t_k, k = 1, 2, 3, \dots$, the system is observed, returning measurement z_k that provides noisy and incomplete information about $\mathbf{x}_k = \mathbf{x}(t_k)$. We assume that we know the conditional distribution over observed value z_k , given the state \mathbf{x}_k ,

$$\rho(z_k | \mathbf{x}_k).$$

Let $Z_t = \{z_k : t_k \leq t\}$ denote the set of observations up to time t , and let (the random variable) $\mathbf{x}_t = \mathbf{x}(t)$ denote the unknown state at time t . The formal Bayesian solution corresponds to determining the time-varying sequence of filtering distributions

$$\rho(\mathbf{x}_t | Z_t) \tag{2}$$

over the state at time t conditioned on all available measurements to time t .

Discrete-Time Formulation A standard approach [1] is to discretize the system equation (1) and treat the discrete-time system [3]. When uncertainty in \mathbf{f} is included via ‘process noise’ \mathbf{v}_k , observation errors via ‘observation noise’ \mathbf{n}_k , the discrete-time problem is written as

$$\begin{aligned}\mathbf{x}_k &= f_k(\mathbf{x}_{k-1}, \mathbf{v}_k) \\ \mathbf{z}_k &= h_k(\mathbf{x}_k, \mathbf{n}_k)\end{aligned}$$

with functions f_k and h_k assumed known.

When the random processes \mathbf{v}_k and \mathbf{n}_k are independently distributed from the current and previous states, the system equation defines a Markov process, as does Eq. (1), while the observation equation defines the conditional probability $\rho(\mathbf{z}_k|\mathbf{x}_k)$.

We will treat the continuous-time problem directly, defining a family of numerical approximations that converge in distribution to the desired continuous-time distributions.

Continuous-Time Bayesian Filtering Sequential Bayesian inference iterates two steps to generate the filtering distributions in Eq. (2) [5].

Prediction Between measurements times t_k and t_{k+1} , Z_t is constant and the continuous-time evolution of the filtering distribution may be derived from the (forward) Chapman-Kolmogorov equation

$$\begin{aligned}\rho(\mathbf{x}_{t+\Delta t}|Z_{t+\Delta t}) &= \rho(\mathbf{x}_{t+\Delta t}|Z_t) = \int \rho(\mathbf{x}_{t+\Delta t}|\mathbf{x}_t, Z_t) \rho(\mathbf{x}_t|Z_t) d\mathbf{x}_t \\ &= \int \delta(\mathbf{x}_{t+\Delta t} - \mathbf{x}(\Delta t; \mathbf{x}_t)) \rho(\mathbf{x}_t|Z_t) d\mathbf{x}_t,\end{aligned}$$

which defines a linear operator on the space of probability distributions,

$$S_{\Delta t} : \rho(\mathbf{x}_t|Z_t) \mapsto \rho(\mathbf{x}_{t+\Delta t}|Z_t). \quad (3)$$

$S_{\Delta t}$ is the Frobenius-Perron (or Foias) operator for time increment Δt .

Update At measurement times t_k , Z_t changes, from Z_{k-1} to Z_k , and the filtering distribution changes, typically discontinuously, as

$$\rho(\mathbf{x}_k|Z_k) = \frac{\rho(\mathbf{z}_k|\mathbf{x}_k) \rho(\mathbf{x}_k|Z_{k-1})}{\rho(\mathbf{z}_k|Z_{k-1})}, \quad (4)$$

which is simply Bayes’ rule written at observation time t_k . We have written $\mathbf{x}_k = \mathbf{x}_{t_k}$ and $Z_k = Z_{t_k}$, and used conditional independence of \mathbf{z}_k and Z_{k-1} given \mathbf{x}_k .

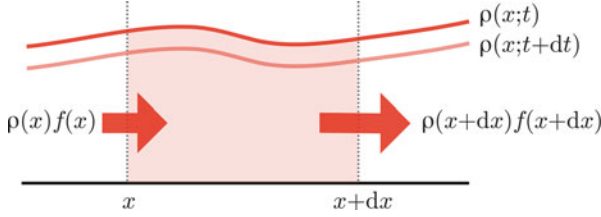


Fig. 4 A schematic of probability flux in region $(x, x + dx)$, and for time $(t, t + dt)$. The schematic shows greater flux exiting the region than entering, correspondingly the pdf at $t + dt$ is decreased with respect to the pdf at t

2.1 The Frobenius-Perron Operator is a PDE

The Frobenius-Perron operator in Eq. (3), that evolves the filtering density forward in time, may be written as the solution of an initial value problem (IVP) in a partial differential equation (PDE) for the probability density function (pdf).

For pdf $\rho(x; t)$ over state x and depending on time t , the velocity field $f(x)$ implies a flux of probability equal to $\rho(x; t) f(x)$. Figure 4 shows a schematic of the pdf and probability flux in region $(x, x + dx)$, and for the time interval $(t, t + dt)$. Equating the rate of change in the pdf with the rate at which probability mass enters the region, and taking $dx, dt \rightarrow 0$, gives the continuity equation

$$\frac{\partial}{\partial t} \rho = -\nabla \cdot (\rho f). \quad (5)$$

The Frobenius-Perron operator $S_{\Delta t}$, for time interval Δt , may be simulated by solving the PDE (5) with initial condition $\rho(x; 0) = \rho(x_t | Z_t)$ to evaluate $\rho(x; \Delta t) = \rho(x_{t+\Delta t} | Z_t)$.

Equation (5) is a linear advection equation. When the state equation has additive stochastic forcing, as is often used to represent model error, evolution of the filtering pdf is governed by a linear advection-diffusion (Fokker-Planck) equation.

3 Finite Volume Solver

The finite volume method (FVM) discretizes the continuity equation in its integral form, for each ‘cell’ K in a mesh,

$$\frac{\partial}{\partial t} \int_K \rho \, dx + \oint_{\partial K} \rho (f \cdot \hat{n}) \, dS = 0.$$

Write $L \sim K$ if cells L and K share a common interface, denoted E_{KL} , and denote by $\hat{\mathbf{n}}_{KL}$ the unit normal on E_{KL} directed from K to L . Define the initial vector of cell values by $\mathbf{P}_K^0 = \frac{1}{|K|} \int_K \rho(\mathbf{x}; 0) \, d\mathbf{x}$ then for $m = 0, 1, \dots, r$ compute \mathbf{P}^{m+1} as

$$\frac{\mathbf{P}_K^{m+1} - \mathbf{P}_K^m}{\Delta t} + \frac{1}{|K|} \sum_{L \sim K} f_{KL} \mathbf{P}_{KL}^m = 0,$$

where

$$f_{KL} = \int_{E_{KL}} \mathbf{f} \cdot \hat{\mathbf{n}}_{KL} \, dS \quad \text{and} \quad \mathbf{P}_{KL}^m = \begin{cases} \mathbf{P}_K^m & \text{if } f_{KL} \geq 0 \\ \mathbf{P}_L^m & \text{if } f_{KL} < 0 \end{cases}$$

is the normal velocity on E_{KL} and first-order *upwinding* scheme, respectively.

In matrix form, the FVM step for time increment Δt is

$$\mathbf{P}^{m+1} = (\mathbf{I} - \Delta t \mathbf{A}) \mathbf{P}^m,$$

where \mathbf{I} is the identity matrix and \mathbf{A} is a sparse matrix defined above. This formula is essentially Euler's famous formula for the (matrix) exponential.

Since $f_{KL} = -f_{LK}$, the FVM conserves probability at each step, i.e., $\sum_K |K| \mathbf{P}_K^{m+1} = \sum_K |K| \mathbf{P}_K^m$. The FVM also preserves positivity of the pdf when the time step Δt is small enough that the matrix $\mathbf{I} - \Delta t \mathbf{A}$ has all non-negative entries. It is straightforward to show that positive entries of the matrix \mathbf{A} can occur on the diagonal, only. Hence, the Courant–Friedrichs–Lewy (CFL) type condition, that assures that the FVM iteration is positivity preserving, may be written

$$\Delta t \leq \frac{1}{\max_i A_{ii}}. \quad (6)$$

With this condition, the FVM both conserves probability and is positivity preserving, hence is a (discrete) *Markov operator*. In contrast, the numerical method for the matrix exponential in MATLAB, for example, does not preserve positivity for the class of matrices considered here.

4 Continuous-Time Frobenius-Perron Operator and Convergence of the FVM Approximation

In this section we summarize results from [7], to develop some analytic properties of the continuous-time solution, and establish properties and distributional convergence of the numerical approximations produced by the FVM solver.

Let $X(\cdot, t) : \mathbb{R}^d \rightarrow \mathbb{R}^d$ denote the map from initial condition to the solution of Eq. (1) at time $t \geq 0$, and $Y(\cdot, t) = X(\cdot, t)^{-1}$. If $X(\cdot, t)$ is *non-singular* ($|Y(E, t)| = 0$ if $|E| = 0 \forall$ Borel subsets $E \subset \mathbb{R}^d$), then $\forall t \geq 0$, the associated Frobenius-Perron operator [6] $S_t : L^1(\mathbb{R}^d) \rightarrow L^1(\mathbb{R}^d)$ is defined by

$$\int_E S_t \rho \, dx = \int_{Y(E, t)} \rho \, dx \quad \forall \text{ Borel subsets } E \subset \mathbb{R}^d.$$

Given an initial pdf p_0 , the pdf $p(\cdot; t)$ at some future time, $t > 0$, may be computed by solving (see, e.g., [6, Def. 3.2.3 and §7.6])

$$\begin{aligned} \frac{\partial}{\partial t} p + \operatorname{div}(fp) &= 0 & \forall x \in \mathbb{R}^d, t > 0 \\ p(x; 0) &= p_0(x) & \forall x \in \mathbb{R}^d \end{aligned} \quad (7)$$

Then, $\forall t \geq 0$, the Frobenius-Perron operator $S_t : L^1(\mathbb{R}^d) \rightarrow L^1(\mathbb{R}^d)$ is defined such that for any $\rho \in L^1(\mathbb{R}^d)$,

$$S_t \rho := p(\cdot; t),$$

where p is a solution to the IVP (7) with $p_0 = \rho$. Existence of a Frobenius-Perron operator and (weak) solutions to the IVP depends on the regularity of f .

Definition 1 (Definition 3.1.1. in [6]) A linear operator $S : L^1(\mathbb{R}^d) \rightarrow L^1(\mathbb{R}^d)$ is a *Markov operator* (or satisfies the *Markov property*) if for any $f \in L^1(\mathbb{R}^d)$ such that $f \geq 0$,

$$Sf \geq 0 \quad \text{and} \quad \|Sf\|_{L^1(\mathbb{R}^d)} = \|f\|_{L^1(\mathbb{R}^d)}.$$

If f has continuous first order derivatives and solutions to Eq. (1) exist for all initial points $x_0 \in \mathbb{R}^d$ and all $t \geq 0$ then the Frobenius-Perron operator is well-defined, satisfies the *Markov property*, and $\{S_t : t \geq 0\}$ defines a continuous semigroup of Frobenius-Perron operators.

FVM Approximation For computational purposes it is necessary to numerically approximate the Frobenius-Perron operators. We use piece-wise constant function approximations on a mesh and the FVM.

Define a mesh \mathcal{T} on \mathbb{R}^d as a family of bounded, open, connected, polygonal, disjoint subsets of \mathbb{R}^d such that $\mathbb{R}^d = \cup_{K \in \mathcal{T}} \overline{K}$. We assume that the common interface between two cells is a subset of a hyperplane of \mathbb{R}^d , and the mesh is *admissible*, i.e.,

$$\exists \alpha > 0 : \begin{cases} \alpha h^d \leq |K| \\ |\partial K| \leq \frac{1}{\alpha} h^{d-1} \end{cases} \quad \forall K \in \mathcal{T}$$

where $h = \sup\{\text{diam}(K) : K \in \mathcal{T}\}$, $|K|$ is the d -dimensional Lebesgue measure of K , and $|\partial K|$ is the $(d - 1)$ -dimensional Lebesgue measure of ∂K .

We will use superscript h to denote numerical approximations (though, strictly, we should use \mathcal{T} as h does not uniquely define the mesh).

The following gives the CFL condition for the (unstructured) mesh \mathcal{T} . Suppose that for some $\xi \in [0, 1)$ and $c_0 \geq 0$, we say that Δt satisfies the CFL condition if

$$\Delta t \sum_{L \sim K} \max\{0, f_{KL}\} \leq (1 - \xi)|K| \quad \forall K \in \mathcal{T} \text{ and } \Delta t \leq c_0 h. \quad (8)$$

Lemma 1 *If Δt satisfies the CFL condition in Eq. (8) and $p_0 \geq 0$ then*

$$p^h(x; t) \geq 0 \quad \forall x \in \mathbb{R}^d, t > 0,$$

and S_t^h is a Markov operator.

The following theorems establish convergence of solutions of the FVM, and convergence of expectations with respect to the filtering distributions.

Theorem 1 *Suppose $\text{div } f = 0$, $\rho \in BV(\mathbb{R}^d)$, and Δt satisfies the CFL condition for some $\xi \in (0, 1)$. Then $\forall t \geq 0$,*

$$\|S_t \rho - S_t^h \rho\|_{L^1(\mathbb{R}^d)} \leq C \xi^{-1} \|\rho\|_{TV} (t^{1/2} h^{1/2} + \xi^{1/2} t h).$$

Convergence of expectations is a consequence of convergence of our FVM.

Theorem 2 *Suppose $H, T < \infty$. Under the same assumptions as previous Theorem, if:*

1. $g \in L^\infty(\mathbb{R}^d)$, or
2. $g \in L_{\text{loc}}^\infty(\mathbb{R}^d)$ and ρ has compact support,

then there exists a constant C independent of h and t such that

$$\left| \mathbb{E}_{S_t^h \rho}[g] - \mathbb{E}_{S_t \rho}[g] \right| \leq C h^{1/2} \quad \forall t \in [0, T], h \in (0, H).$$

This guarantees convergence in distribution of the discrete approximation to the continuous-time filtering pdfs in the limit $h \rightarrow 0$.

In numerical tests [7] we found convergence to be $\mathcal{O}(h)$, which is twice the order predicted by Theorem 2. Since the CFL condition requires the time step is also $\mathcal{O}(h)$, the method is $\mathcal{O}(\Delta t)$ accurate. Thus the FVM method we use achieves the highest order permitted by the meta theorem of Bolley and Crouzeix [2], that positivity-preserving Runge-Kutta methods can be first order accurate, at most.

5 Computed Examples

We now present a computed example of the finite volume filter (FVF) that uses the FVM for implementing the Frobenius-Perron operator during the prediction phase, and the update rule, Eq. (4) evaluated using mid-point quadrature.

Further details of these numerical experiments can be found in [4], including comparison to filtering produced by the unscented Kalman filter (UKF).

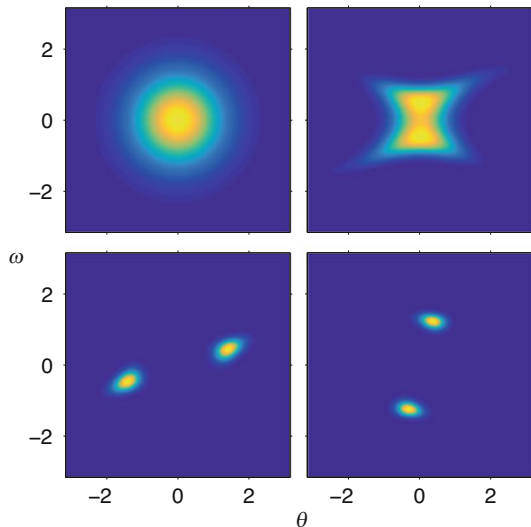
5.1 FVF Tracking of a Pendulum from Measured Force

Figure 5 shows four snapshots of the filtering pdf for the stylized model of Sect. 1.1. The ‘true’ pendulum was simulated with initial condition $(\theta_0, \omega_0) = (0.2\pi, 0)$. Eight measured values of the force were recorded, per 2π time, over time period of 3π , with added Gaussian noise having $\sigma = 0.2$. The FVF was initialized with $N(0, 0.8^2 I)$.

Since the initial and filtering pdfs are symmetric about the origin, the means of angular displacement and velocity are always identically zero. Hence, filtering methods that such as the UKF, or any extension of the Kalman filter that assumes Gaussian pdfs, or that focus on the mean as a ‘best’ estimate, will estimate the state as identically zero, for all time. Clearly, this is uninformative.

In contrast, the FVF has localized the true state after 3π time (about 1.5 periods), albeit with ambiguity in sign. Properties of the system that do not depend on the sign of the state, such as the period, the length of the pendulum, or the mass of the pendulum, can then be accurately estimated. The computed example in [7]

Fig. 5 Initial ($t = 0$) and filtered pdfs in phase-space after measurements at times $t = \pi/4, \pi,$ and 3π (left to right, top to bottom)



shows that the length of the pendulum is correctly determined after just 1.5, and that the accuracy of the estimate improves with further measurements (and time). The same feature holds for estimating mass. Hence, the FVM is successful in accurately estimating the parameters in the stylized model, even when the measurements leave ambiguity in the kinematic state.

6 Conclusions

Bayes-optimal filtering for a dynamical system requires solving a PDE. It is interesting to view filters in terms of the properties of the implicit, or explicit, PDE solver, such as the density function representation and technology used in the PDE solver. This paper develops a FVM solver using the simplest-possible discretization to implement a Bayes-optimal filter, that turned out to be computationally feasible for low-dimensional smooth systems.

The reader may be interested to know that TCAM won his wager, and beer, with new sequential inference algorithms now producing useful results on the farm. In the interests of full disclosure we should also report that the original notion of utilizing a simple dynamical-systems model for a walking cow did not perform well, as the model ‘cow’ would eventually walk upside down, just as the pendulum prefers to hang downwards. In response, we developed models for cow locomotion based on energy conservation, that are beyond the scope of this paper. However, the FVM has found immediate application in other dynamic estimation problems where a dynamical model that evolves a state vector works well, such as estimating the equilibrium temperature of milk during steaming as a tool for training coffee baristas.

References

1. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Proces.* **50**(2), 174–188 (2002)
2. Bolley, C., Crouzeix, M.: Conservation de la positivité lors de la discrétisation des problèmes d’évolution paraboliques. *R.A.I.R.O. Analyse Numérique* **12**, 237–245 (1978)
3. Cappé, O., Moulines, E., Ryden, T.: *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, New York (2005). <https://doi.org/10.1007/0-387-28982-8>
4. Fox, C., Morrison, M.E.K., Norton, R.A., Molteni, T.C.A.: Optimal nonlinear filtering using the finite-volume method. *Phys. Rev. E* **97**(1), 010201 (2018)
5. Jazwinski, A.H.: *Stochastic Processes and Filtering Theory*. Academic, New York (1970)
6. Lasota, A., Mackey, M.C.: *Chaos, fractals, and noise*. In: *Applied Mathematical Sciences*, vol. 97, 2nd edn. Springer, New York (1994). <https://doi.org/10.1007/978-1-4612-4286-4>
7. Norton, R.A., Fox, C., Morrison, M.E.K.: Numerical approximation of the Frobenius-Perron operator using the finite volume method. *SIAM J. Numer. Anal.* **56**(1), 570–589 (2018)
8. Treblicock, K.: Building better scales – for a beer. *N. Z. Dairy Exporter* **2011**, 101 (2011)

Correlation Integral Likelihood for Stochastic Differential Equations



Heikki Haario, Janne Hakkarainen, Ramona Maraia, and Sebastian Springer

Abstract A new approach was recently introduced for the task of estimation of parameters of chaotic dynamical systems. Here we apply the method for stochastic differential equation (SDE) systems. It turns out that the basic version of the approach does not identify such systems. However, a modification is presented that enables efficient parameter estimation of SDE models. We test the approach with basic SDE examples, compare the results to those obtained by usual state-space filtering methods, and apply it to more complex cases where the more traditional methods are no more available.

1 Introduction

The difficulty of estimating parameters of chaotic dynamical models is related to the fact that a fixed model parameter does not correspond to a unique model integration, but to a set of quite different solutions as obtained by setting slightly different initial values, selecting numerical solvers used to integrate the system, or tolerances specified for a given solver. But while all such trajectories are different, they approximate the same underlying attractor and should be considered in this sense equivalent. In [4] we introduced a distance concept for chaotic systems based on this insight. Modifying one of the fractal dimension definitions, the correlation dimension, we calculate samples from the phase space of the system and map these points onto a stochastic vector. The vector turns out to be Gaussian, providing a

H. Haario · R. Maraia (✉) · S. Springer

School of Engineering Science, Lappeenranta University of Technology, Lappeenranta, Finland
e-mail: heikki.haario@lut.fi; ramona.maraia@lut.fi; sebastian.springer@lut.fi

J. Hakkarainen

Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland

Earth Observation, Finnish Meteorological Institute, Earth Observation, Finnish Meteorological Institute, Helsinki, Finland

e-mail: janne.hakkarainen@fmi.fi

© Springer Nature Switzerland AG 2019

D. R. Wood et al. (eds.), *2017 MATRIX Annals*, MATRIX Book Series 2,

https://doi.org/10.1007/978-3-030-04161-8_3

natural likelihood concept that quantifies the chaotic variability of points of a chaotic system within a given setting of observations.

Stochastic differential equation (SDE) systems behave partly in a similar way: each integration of a given system with fixed model parameters produces a different realization. This calls for methods that can quantify the variability of the realizations. On the other hand, the stochastic nature of a SDE system is clearly different from the chaotic variability of a deterministic chaotic system. Consequently, the phase space behavior of each type of systems is different as well. The aim of this work is to study to which extent the parameter estimation approach originally developed for chaotic systems can be applied to SDE models.

The rest of the paper is organized as follows. In the Background section we recall the correlation integral likelihood concept and outline the results obtained for chaotic systems. In Numerical experiments we exhibit the performance of the method for the Ornstein-Uhlenbeck model and extensions of it, together with comparisons to more standard, Kalman filter based methods.

2 Background

The standard way of estimating parameters of dynamical systems is based on the residuals between the data and the model responses, both given at the time points of the measurements. Supposing the statistics of the measurement error is known, a well defined likelihood function can be written. The maximum likelihood point is typically considered as the best point estimator, and it coincides with the usual least squares fit in the case of Gaussian noise. The full posterior distribution of parameters can be sampled by Markov chain Monte Carlo (MCMC) methods. The approach has become routine for the parameter estimation of deterministic models in Bayesian inference.

The estimation of the parameters of stochastic models is not so straightforward. A given model parameter does not correspond to a fixed solution, but a whole range of possible realizations. Several methods have been proposed to overcome this difficulty. State-based approaches estimate the joint distribution of the state vector and the parameters. The likelihood for the parameter is obtained as a marginal distribution, effectively by ‘integrating out’ the state space. This approach is routine in the context of linear time series modeling, and implemented by the likelihood obtained by application of the Kalman filter formulas, see [2, 7, 11].

Here we study a different way of characterizing the stochastic variability of the state space. Supposing that a sufficient amount of data is available, we create a mapping from it onto a feature vector. The mapping is based on averaging, and the vector turns out to be asymptotically Gaussian. From real data, the mean and covariance of this Gaussian distribution can be empirically estimated. Thus we have a likelihood available, both for maximum likelihood parameter estimation and for MCMC sampling of the parameter posterior. The idea is the same as that earlier used for estimating parameters of chaotic models in [4], but certain modifications

are needed for SDE systems. We discuss the basic setting of the approach below, as well as the reasons behind the modifications needed.

2.1 Likelihood via Filtering

A standard way of estimating the parameters with stochastic models is to use filtering methods for constructing the likelihood (see, e.g., [2, 7, 11] for basic references and implementation, or [8] for recent variant). By using the Kalman filter, the idea is to build the marginal filter likelihood from the prediction residual \mathbf{r}_k and its error covariance matrix \mathbf{C}_k^r at each filtering time step k .

The basic linear Kalman filter is written as a pair

$$\mathbf{x}_k = \mathbf{M}_k \mathbf{x}_{k-1} + \boldsymbol{\xi}_k, \quad (1)$$

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \boldsymbol{\varepsilon}_k, \quad (2)$$

where \mathbf{x}_k is the state and \mathbf{y}_k is the measurement vector. Matrix \mathbf{M}_k is the linear state-space model, and matrix \mathbf{H}_k is the observation operator that maps from the state space to the observation space. The error terms $\boldsymbol{\xi}_k$ and $\boldsymbol{\varepsilon}_k$ are typically assumed zero mean and Gaussian: $\boldsymbol{\xi}_k \sim N(\mathbf{0}, \mathbf{Q}_k)$ and $\boldsymbol{\varepsilon}_k \sim N(\mathbf{0}, \mathbf{R}_k)$. This dynamical system is solved using Kalman filter formulas (see, e.g., [11]).

Given a set of observation $\mathbf{y}_{1:K}$ and the parameter vector $\boldsymbol{\theta}$, the marginal filter likelihood is written as

$$p(\mathbf{y}_{1:K} | \boldsymbol{\theta}) = \exp\left(-\frac{1}{2} \sum_{k=1}^K \left[\mathbf{r}_k^T (\mathbf{C}_k^r)^{-1} \mathbf{r}_k + \log |\mathbf{C}_k^r| \right]\right), \quad (3)$$

where $|\cdot|$ denotes the matrix determinant. Here the prediction residual and its error covariance matrix are calculated by the formulas

$$\mathbf{r}_k = \mathbf{y}_k - \mathbf{H}_k \mathbf{x}_k^{\text{prior}}, \quad (4)$$

$$\mathbf{C}_k^r = \mathbf{H}_k \mathbf{C}_k^{\text{prior}} \mathbf{H}_k^T + \mathbf{R}_k, \quad (5)$$

where $\mathbf{x}_k^{\text{prior}}$ is the prior estimate computed from the previous state $\mathbf{x}_k^{\text{prior}} = \mathbf{M}_k \mathbf{x}_{k-1}^{\text{est}}$, and $\mathbf{C}_k^{\text{prior}} = \mathbf{M}_k \mathbf{C}_{k-1}^{\text{est}} \mathbf{M}_k^T + \mathbf{Q}_k$ is the respective error covariance matrix. Note that the normalizing ‘‘constant’’ $|\mathbf{C}_k^r|$ has to be included, since it depends on the parameters via the prediction model.

This approach is well established in the framework of linear time series or linear SDE systems, where the additive model noise is known or may be estimated, as one of the unknowns in the vector $\boldsymbol{\theta}$. In case the drift part of the system (1) is nonlinear, one still may use the approach in the extended Kalman filter (EKF)

form, based on the approximation by linearization. Often the EKF approach is also applied to filtering of deterministic systems. In that setting the model error term is rather postulated and interpreted as a measure of bias. The covariances \mathbf{Q} and \mathbf{R} represent then our trust on the model and data, respectively, previous work [5], motivated by closure parameter estimation in climate research, is an example of this approach. A related option is to employ ensemble filtering. In [12] this approach was employed in order to tune the ensemble prediction system parameters. It was observed, however, that the method resulted in a highly stochastic cost function that prevented a successful application of parameter optimization algorithms. Moreover, the tuning parameters of the filter itself may bias the model parameter estimation, see [6]. Recently, some additional criticism toward using the filtering for estimating the parameters in real-world applications (other than finance) has been presented see [10].

Next, we present the method developed in [4] for deterministic chaotic systems. While computationally more demanding, it is free of the pitfalls listed above, and can be applied to stochastic systems more general than the class of additive noise given by (1).

2.2 Correlation Integral Likelihood

In this section we briefly summarize the correlation integral likelihood method used for creating a likelihood for complex patterns [4].

Let us use the notation $\mathbf{s} = \mathbf{s}(\boldsymbol{\theta}, \mathbf{x})$ for a state vector \mathbf{s} that depends on parameters $\boldsymbol{\theta}$ and other inputs \mathbf{x} such as, e.g., the initial values of a dynamical system. We consider two different trajectories, $\mathbf{s} = \mathbf{s}(\boldsymbol{\theta}, \mathbf{x})$ and $\tilde{\mathbf{s}} = \mathbf{s}(\tilde{\boldsymbol{\theta}}, \tilde{\mathbf{x}})$, evaluated at $N \in \mathbb{N}$ time points t_i , $i = 1 : N$, with explicit dependency on the respective initial and parameter values. For $R \in \mathbb{R}$, the *modified correlation sum* is defined as

$$C(R, N, \boldsymbol{\theta}, \mathbf{x}, \tilde{\boldsymbol{\theta}}, \tilde{\mathbf{x}}) = \frac{1}{N^2} \sum_{i,j} \# (\|\mathbf{s}_i - \tilde{\mathbf{s}}_j\| < R). \quad (6)$$

In the case $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}$ and $\tilde{\mathbf{x}} = \mathbf{x}$ the formula reduces to the well known definition of *correlation sum*, the *Correlation Integral* is then defined as the limit $C(R) = \lim_{N \rightarrow \infty} C(R, N)$, and the *Correlation Dimension* ν as the limit

$$\nu = \lim_{R \rightarrow 0} \log C(R) / \log(R).$$

In numerical practice, the limit $R \rightarrow 0$ is approximated by the small scale values of the ratio above, by the log-log plot obtained by computing $\log C(R)$ at various values of $\log R$.

However, we do not focus on the small-scale limit as in the above definition, but rather use the expression (6) at all relevant scales R to characterize the distance between two trajectories. For this purpose, a finite set of decreasing radii $R = (R_k)$, $k = 1, \dots, M$, is chosen. The radii values R_k are selected so as to involve both small and large scale properties of the trajectory samples. Typically, the radii are chosen as $R_k = b^{-k} R_0$, with $R_0 = \max_{i,j} \|s_i - \tilde{s}_j\|$ or somewhat larger to ensure that all the values are inside the largest radius. The values of M and b should be chosen in a way that R_M is small enough. For more details see [4].

Consider now the case with given data s_{i_2} which corresponds to the case of a fixed but unknown model parameter vector, $\theta = \theta = \theta_0$. We select two subsets \mathbf{s} and $\tilde{\mathbf{s}}$ of size N from the data (see more details below). If we fix the radii values $R = (R_k)$, $k = 1, \dots, M$ the expression (6) defines a M dimensional vector with components $y_k = C(R_k, \theta_0, \mathbf{x})$. A training set of these vectors is created by repeatedly selecting the subsets \mathbf{s} and $\tilde{\mathbf{s}}$. The statistics of this vector can then be estimated in a straightforward way.

Indeed, the expression (6) is an average of distances, so by the Central Limit Theorem it might be expected to get Gaussian. More exactly, each expression $\mathbf{y} = (y_k)$ gives the empirical cumulative distribution function of the respective set of distances. The basic form of the Donsker's theorem tells that empirical distribution functions asymptotically tend to a Brownian bridge. In a more general setting, close to what we employ here, the Gaussianity was established by Borovkova et al. [1].

At a pseudo code level the procedure can be summarized as follow:

- Using the measured data, create a training set of the vectors \mathbf{y} for fixed radii values (R_k) by sampling data at measurement times (t_i).
- Create the empirical statistical distribution of the training set \mathbf{y} as a Gaussian likelihood, by computing the mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$ of the training set vectors.
- Find the maximum likelihood model parameter θ_0 of the distribution

$$P_{\theta_0}(\theta, x) \sim \exp -\frac{1}{2}(\boldsymbol{\mu} - y(\theta, x))^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - y(\theta, x))$$

- Sample the likelihood to find those model parameters θ for which the vector $\mathbf{y} = C(\theta_0; \mathbf{x}; \theta; \tilde{\mathbf{x}})$ belongs to the distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The first step will be discussed more in detail in the examples below. Note that in [4] we assumed a parameter value θ_0 given and created the training data by model simulations, while here we start with given data, create the training set from subsets of data, and proceed to estimate a maximum likelihood parameter value θ_0 .

Remark In all the cases the prior distribution is assumed to be flat uniform.

3 Numerical Experiments

The main objective of this section is to modify the Correlation integral likelihood (CIL) method for identifying SDE system parameters. The new version of the method is compared with the filter likelihood results. After this validation the approach is applied to a more complex case.

3.1 Ornstein-Uhlenbeck with Modification for Dynamics

We start with a basic SDE example, the Ornstein-Uhlenbeck (OU) process model. We use it as a benchmark to verify that the CIL method is able to produce results comparable to standard filter likelihood methods in a setting where these classical methods perform perfectly well. The OU process equation is given by

$$dX_t = -\theta X_t dt + \sigma dW_t. \quad (7)$$

In the numerical simulations, we use $\theta = 10$ and $\sigma = 1.5$ as the ‘true’ values. For simplicity, the mean value of the process is set to zero (but all the results and conclusions are valid for a non-zero mean as well). We create a data signal of 3000 points on the time interval $[0, 30]$, with initial value $X = 0$.

Figure 1 exhibits the signal used as data, obtained by integration of (7) using the Euler-Maryama method, with a time step $dt = 0.01$ and using a fixed Gaussian $N(0, \sigma^2)$ as the diffusion part. The figure presents three different realizations. Note that essentially the same results as those given below were obtained by any realizations used.

Let us first apply the CIL method in the basic form. To create the sample sets \mathbf{s}_i we randomly select 1500 of the data points of the signal in Fig. 1 and use the rest of the points as \mathbf{s}_j to get the set of distances needed in (6). This process is repeated around 2000 times to get a representative set of the feature vectors \mathbf{y} . The likelihood is then obtained by computing the mean and covariance of the training vectors \mathbf{y} , and the Normality of the vectors can be verified by the usual χ^2 test.

Next, we find the distribution of the model parameters θ, σ that follows this distribution by creating a MCMC chain of length 20,000 using adaptive Metropolis [3]. The result in Fig. 2 shows, however, that the model parameters are not identified by this likelihood. This situation is different from those reported in [4], and several unpublished cases, for chaotic systems, where the same likelihood construction is able to identify the model parameters.

We conclude that too much information is lost in the mapping from data to the feature vectors \mathbf{y} . Indeed, this is not surprising in view of the fact that only the distances between randomized data points is considered, while the order or differences between consecutive points is lost. A trivial example is given by any

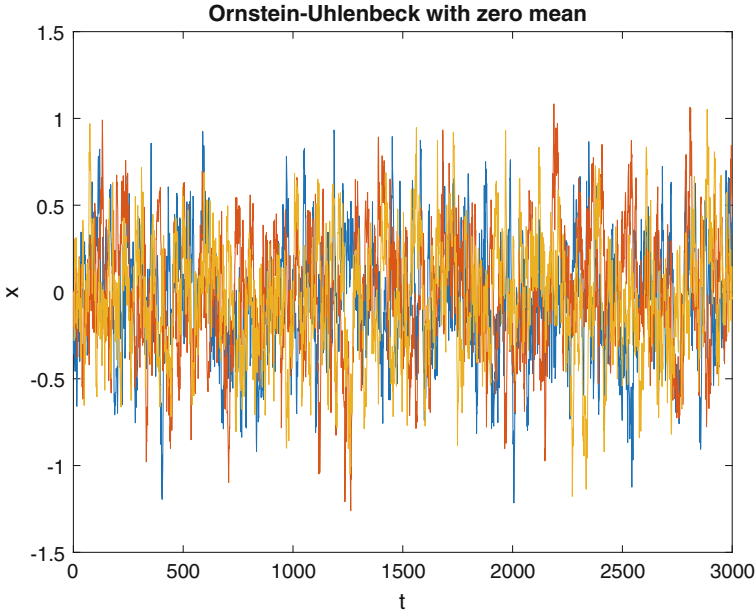


Fig. 1 Ornstein-Uhlenbeck signal used for the experiments

vector or random points: sorting it in increasing order gives a definitely different signal, but with just the same set of points and distances between them.

Intuitively, the mean reverting dynamics is lost here, so some additional modification of the method is needed. The large posterior in Fig. 2 exhibits only what it is programmed to do: signals whose distance distributions remain close, which in this case does not characterize the signals. The feature vector can be modified in various ways. Here we present the impact of extending it in the obvious way: we include the differences between consecutive points. We create the feature vectors separately for the signal and for the differences. The final feature vector is created by concatenating the curves, and the Gaussianity of the combined vector can be tested by the χ^2 test. Figure 2 illustrates the posterior obtained using three different levels of information: only the data signal, only difference between consecutive points, and both together. We see how the first two are not enough, while the posterior of the extended case, practically the intersection of the two other posteriors, significantly improves the identification.

Next, we compare the Correlation Integral Likelihood results with that obtained by filter likelihood estimation based on Kalman filtering. We use the same data signal as above, using all the points X_k , $k = 1, \dots, 3000$ as exact measurements (no noise added) of the state vectors, and create MCMC samples of the likelihood given by the expression (3). The comparison presented in Fig. 3. As expected, the filtering method is more accurate with this amount of data (we use every Euler-Maryama integration step as data for filtering), but the results by CIL are comparable.

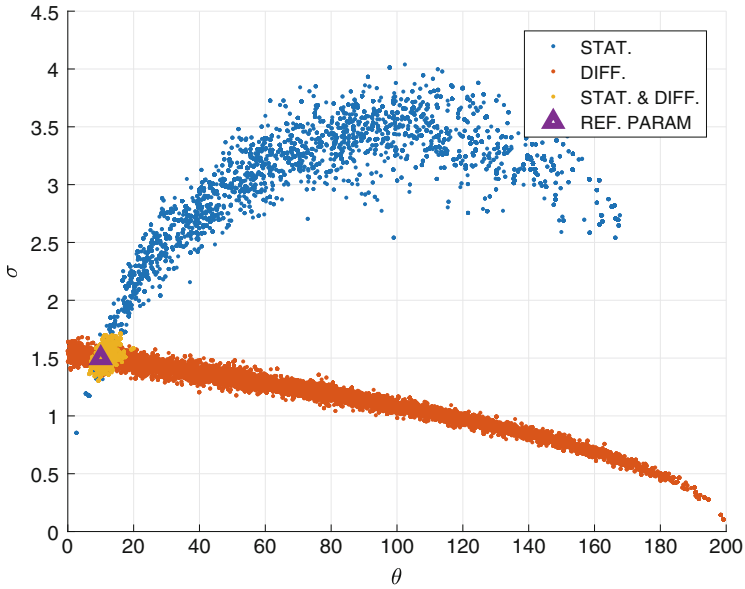


Fig. 2 The use of both state and difference information leads to a posterior (yellow) that is located around the intersection of the posterior generated by the state information only (blue) and the one generated using the difference only (orange)

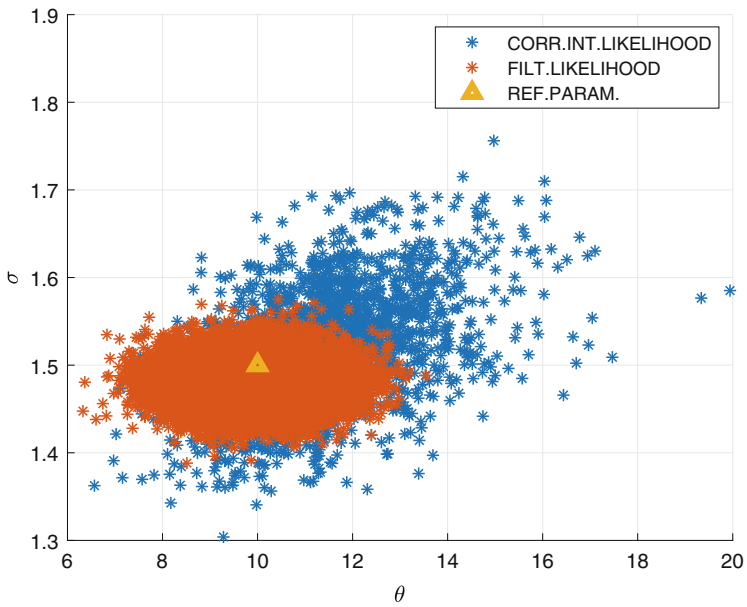


Fig. 3 Illustration of the results obtained by comparing CIL with the Filter likelihood method in parameter estimation for a zero mean Ornstein-Uhlenbeck

Remarks In the above examples we have used the known value of θ_0 as the starting point for the MCMC sampling. However, as the likelihood is created by the data signal, we can equally well use it as the cost function to estimate θ_0 first. We omit here the details of this step.

Note that there is a difference in computational times of the two methods, in this particular case they are approximately 20 min for CIL and around 6 min for KF. The difference is basically due to the additional computation of the distances needed for CIL.

Note that using a larger time step between data points would decrease the accuracy of the KF estimate. However, it does not impact the CIL estimate, as it is based on independent samples X_i in random order, not on predicting X_{i+1} by X_i .

Finally, we note that the use of the present modification, including the system ‘dynamics’ by signal differences, is not limited to the OU example. Rather, it can be used generally to improve the model parameter identification of both SDE and deterministic chaotic systems. However, a more detailed discussion is outside the scope of this work.

3.2 Stochastic Chaos

Here we study the CIL approach for chaotic dynamics, extended with stochastic perturbations. Now the stochasticity is no more of the additive form (1) but is contained in the model equations in a nonlinear way. The specific forms of the perturbations discussed here come from meteorology. In the so called Ensemble Prediction Systems (EPS) an ensemble of weather predictions, with carefully perturbed initial values, is launched together with the main prediction. The motive is to create probabilistic estimates for the uncertainty of the prediction. However, it is difficult to create a spread of the ensemble predictions that would match the observed uncertainty; the spread of the model simulations tends to be too narrow. To increase the spread the so called stochastic physics is employed: the right hand side of the model differential equation is multiplied by a random factor (close to one) at every integration step. More recently, so called stochastic parametrization is used in addition: certain model parameters are randomized likewise at every integration step of the system. For more details of these methods see [9].

As a case study for the parameter estimation with stochastic physics and stochastic parametrization a classical chaotic attractor, the Rossler system, is chosen. We give the Rossler system in the form where the stochastic physics is introduced by the multiplicative factors $1 + c_k \epsilon$, and the model parameters α, β, γ are likewise replaced by perturbed terms $\alpha + c_k \epsilon$, etc., $k = 1 : 6$, $\epsilon \sim N(0, 1)$. The system reads as

$$\begin{cases} \dot{X} = (1 + c_1 \epsilon_1) (-Y - Z) \\ \dot{Y} = (1 + c_2 \epsilon_2) (X + (\alpha + c_3 \epsilon_3) Y) \\ \dot{Z} = (1 + c_4 \epsilon_4) ((\beta + c_5 \epsilon_5) + Z (X - (\gamma + c_6 \epsilon_6))) \end{cases} \quad (8)$$

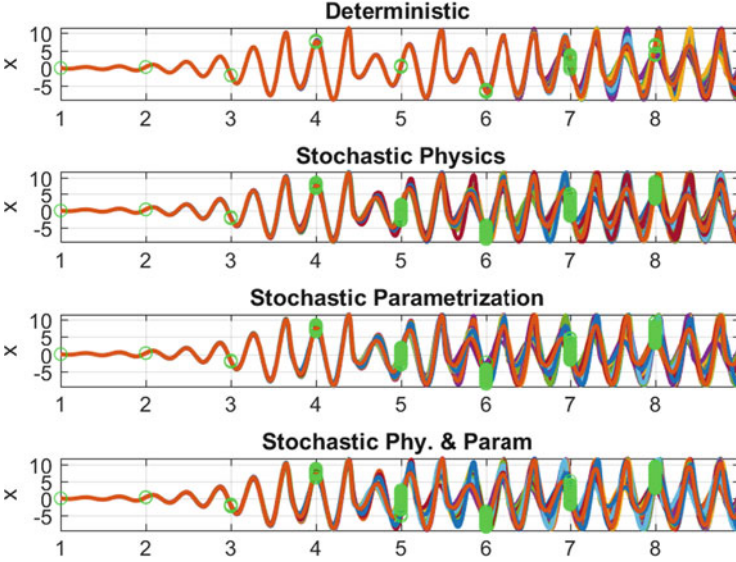


Fig. 4 The X component of the Rossler model with four different options for stochasticity

with ‘true’ parameters $\alpha = \beta = 0.2$ and $\gamma = 5.7$. The magnitudes c_k were chosen so that the maximum relative error would not exceed 40% in any of the cases.

Figure 4 shows the time evolutions of one of the components, the values of X for different combinations of added stochasticity. Each plot consists of 80 runs with slightly perturbed initial values. We see that the interval of predictable behavior shrinks to almost one half of that of deterministic chaos when both types of perturbations are added.

The task of parameter estimation is now to try to find the distribution of the mean value of each of the perturbed parameters. The construction of the likelihood is performed via the standard procedure: from a long enough data signal (here, produced by simulating (8)) we sample subsets to calculate the distances, and repeat this for a number of times to be able to empirically determine the statistics of the feature vectors. Again, the Gaussianity of the statistics can be verified. Both a maximum likelihood parameter estimate, and the subsequent MCMC sampling for the posterior can then be performed.

For the examples we create the data by simulating (8) over a total time interval $[0, 120,000]$ and select data points at frequency shown in Fig. 4 with the green circles. To get one feature vector y we select two disjoint sets of 2000 consecutive data points. To create the statistics for y we repeat this procedure for around 1800 times. The number of radius values used was 10.

The results of the runs for different setting of the perturbations are given in Fig. 5. We can conclude that the approach performs as expected: the more stochasticity in the model, the wider are the parameter posteriors. However, in all cases we get bounded posteriors, and the algorithm performs without any technical issues.

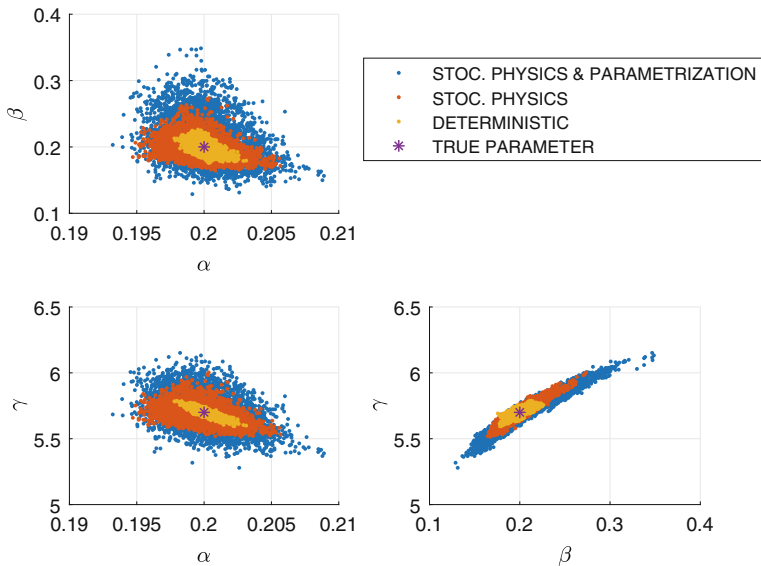


Fig. 5 Parameter posteriors for three different options of stochasticity for the Rossler model

4 Conclusions

In this work we have applied the recently developed Correlation Integral Likelihood method to estimate parameters of stochastic differential equation systems. Certain modifications are needed to get satisfactory results, comparable to those achieved by standard filter likelihood methods for basic SDE systems. But the main focus is on situations where the standard methods are not available, such as the stochastic physics and parametrizations employed in meteorology for uncertainty quantification. Several extensions of the approach are left for future work.

Acknowledgements This work was supported by the Centre of Excellence of Inverse Problems, Academy of Finland.

References

1. Borovkova, S., Burton, R., Dehling, H.: Limit theorems for functionals of mixing processes with applications to U -statistics and dimension estimation. *Trans. Am. Math. Soc.* **353**(11), 4261–4318 (2001). <https://doi.org/10.1090/S0002-9947-01-02819-7>
2. Durbin, J., Koopman, S.J.: *Time Series Analysis by State Space Methods*. Oxford University Press, Oxford (2012)
3. Haario, H., Laine, M., Mira, A., Saksman, E.: DRAM: Efficient adaptive MCMC. *Stat. Comput.* **16**(4), 339–354 (2006). <https://doi.org/10.1007/s11222-006-9438-0>

4. Haario, H., Kalachev, L., Hakkarainen, J.: Generalized correlation integral vectors: a distance concept for chaotic dynamical systems. *Chaos: Interdiscipl. J. Nonlinear Sci.* **25**(6), 063102 (2015). <http://dx.doi.org/10.1063/1.4921939>
5. Hakkarainen, J., Ilin, A., Solonen, A., Laine, M., Haario, H., Tamminen, J., Oja, E., Järvinen, H.: On closure parameter estimation in chaotic systems. *Nonlinear Process. Geophys.* **19**(1), 127–143 (2012). <http://dx.doi.org/10.5194/npg-19-127-2012>
6. Hakkarainen, J., Solonen, A., Ilin, A., Susiluoto, J., Laine, M., Haario, H., Järvinen, H.: A dilemma of the uniqueness of weather and climate model closure parameters. *Tellus A Dyn. Meteorol. Oceanogr.* **65**(1), 20147 (2013). <http://dx.doi.org/10.3402/tellusa.v65i0.20147>
7. Laine, M., Latva-Pukkila, N., Kyrölä, E.: Analysing time-varying trends in stratospheric ozone time series using the state space approach. *Atmos. Chem. Phys.* **14**(18), 9707–9725 (2014). <https://doi.org/10.5194/acp-14-9707-2014>. <https://www.atmos-chem-phys.net/14/9707/2014/>
8. Mbalawata, I.S., Särkkä, S., Haario, H.: Parameter estimation in stochastic differential equations with Markov chain Monte Carlo and non-linear Kalman filtering. *Comput. Stat.* **28**(3), 1195–1223 (2013). <https://doi.org/10.1007/s00180-012-0352-y>
9. Ollinaho, P., Lock, S.J., Leutbecher, M., Bechtold, P., Beljaars, A., Bozzo, A., Forbes, R.M., Haiden, T., Hogan, R.J., Sandu, I.: Towards process-level representation of model uncertainties: stochastically perturbed parametrizations in the ECMWF ensemble. *Quart. J. R. Meteorol. Soc.* **143**(702), 408–422 (2017). <http://dx.doi.org/10.1002/qj.2931>
10. Rougier, J.: ‘Intractable and unsolved’: some thoughts on statistical data assimilation with uncertain static parameters. *Philos. Trans. R. Soc. Lond. A Math. Phys. Eng. Sci.* **371**(1991) (2013). <https://doi.org/10.1098/rsta.2012.0297>. <http://rsta.royalsocietypublishing.org/content/371/1991/20120297>
11. Särkkä, S.: *Bayesian Filtering and Smoothing*. Cambridge University Press, Cambridge (2013)
12. Solonen, A., Järvinen, H.: An approach for tuning ensemble prediction systems. *Tellus A Dyn Meteorol Oceanogr* **65**(1), 20594 (2013). <http://dx.doi.org/10.3402/tellusa.v65i0.20594>

A Set Optimization Technique for Domain Reconstruction from Single-Measurement Electrical Impedance Tomography Data



Bastian Harrach and Janosch Rieger

Abstract We propose and test a numerical method for the computation of the convex source support from single-measurement electrical impedance tomography data. Our technique is based on the observation that the convex source support is the unique minimum of an optimization problem in the space of all convex and compact subsets of the imaged body.

1 Introduction

Electrical impedance tomography is a modern non-invasive imaging technology with the potential to complement computerized tomography in treatments like pulmonary function diagnostics and breast cancer screening. From a mathematical perspective, the reconstruction of the exact conductivity within the imaged body from electrical impedance tomography data amounts to solving a strongly ill-posed inverse problem.

The difficulty of this problem can partly be avoided by noting that one is usually not interested in the conductivity as such, but only in the domain where it differs from the conductivity of healthy tissue. A technique introduced in [5] for scattering problems and adapted to electrical impedance tomography later in [3] takes this approach one step further by considering a convex set, called the convex source support, which contains information on the desired domain, but can be computed from a single measurement.

B. Harrach

Goethe University Frankfurt, Institute for Mathematics, Frankfurt am Main, Germany
e-mail: harrach@math.uni-frankfurt.de

J. Rieger (✉)

Monash University, School of Mathematical Sciences, Clayton, VIC, Australia
e-mail: janosch.rieger@monash.edu

We propose a numerical method for the computation of the convex source support from electrical impedance tomography data. It is based on the observation that this particular set is the unique minimum of an optimization problem in the space of all convex and compact subsets of the imaged body. In Sect. 2, we recall the notion of the convex source support, and in Sect. 3, we formulate the above-mentioned optimization problem and manipulate its constraint into a convenient form. In Sect. 4, we introduce Galerkin approximations, which are spaces of polytopes with fixed outer normals, to the space of all convex and compact subsets of a given Euclidean vector space. These spaces possess a sparse and unique representation in terms of coordinates. In Sect. 5, we discuss how the derivatives of the objective function and the constraint of our optimization problem can be computed efficiently. In Sect. 6, we gather all the above ingredients and solve the optimization problem numerically using a standard interior point method on the Galerkin approximation, which yields a numerical approximation of the convex source support.

This paper is a report on work in progress, which aims to present ideas rather than a complete solution of the problem. In particular, we assume that we can measure the potential on the entire boundary of the imaged body, which is not possible in real-world applications, and we neither include an error analysis nor stability results for the proposed algorithm.

2 The Convex Source Support in Electrical Impedance Tomography

We consider the following idealistic model of the EIT problem. Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a smoothly bounded domain describing the imaged body, let $\sigma \in L^\infty(\Omega)$ be the conductivity within Ω , and let $g \in L_\diamond^2(\partial\Omega)$ be the electric current applied to $\partial\Omega$, where $L_\diamond^2(\partial\Omega)$ denotes the subspace of $L^2(\partial\Omega)$ with vanishing integral mean on $\partial\Omega$. Then the electrical potential $u \in H_\diamond^1(\Omega)$ solves

$$\begin{aligned} \nabla \cdot (\sigma(x)\nabla u(x)) &= 0, & x \in \Omega, \\ \sigma \partial_\nu u|_{\partial\Omega}(x) &= g(x), & x \in \partial\Omega, \end{aligned} \tag{1}$$

where ν is the outer normal on $\partial\Omega$, and $H_\diamond^1(\partial\Omega)$ is the subspace of $H^1(\Omega)$ -functions with vanishing integral mean on $\partial\Omega$.

Our aim is to find inclusions or anomalies in Ω where the conductivity σ differs from a reference conductivity value σ_0 (e.g. that of healthy tissue) from measuring the electric potential $u|_{\partial\Omega}$ on $\partial\Omega$. To simplify our exposition we assume $\sigma_0 \equiv 1$ throughout this work. More precisely, we aim to find information on $\text{supp}(\sigma - \sigma_0)$ from the data $(u - u_0)|_{\partial\Omega}$, where u_0 solves (1) with the same Neumann boundary data g , and σ replaced by σ_0 . This is usually referred to as the problem of single measurement EIT since only one current g is applied to the patient.

Now we introduce the convex source support, following [3] and [5]. First note that since u and u_0 are solutions of (1) with conductivities σ and σ_0 and identical Neumann data $g \in L^2_\diamond(\partial\Omega)$, their difference $w := u - u_0$ solves the equation

$$\Delta w = \operatorname{div}((1 - \sigma)\nabla u) \text{ in } \Omega, \quad \partial_\nu w = 0 \text{ on } \partial\Omega, \quad (2)$$

with a source term satisfying

$$\operatorname{supp}((1 - \sigma)\nabla u) \subset \operatorname{supp}(1 - \sigma). \quad (3)$$

This motivates the following construction of the convex source support. Let us define the virtual measurement operator

$$L : L^2(\Omega)^d \rightarrow L^2_\diamond(\partial\Omega), \quad F \mapsto w|_{\partial\Omega},$$

where $w \in H^1_\diamond(\Omega)$ solves

$$\Delta w = \operatorname{div} F \text{ in } \Omega, \quad \partial_\nu w = 0 \text{ on } \partial\Omega.$$

Given a measurement $f = (u - u_0)|_{\partial\Omega} \in L^2_\diamond(\partial\Omega)$, the convex source support of problem (1) is defined by

$$\mathcal{C}f := \bigcap_{LF=f} \operatorname{co}(\operatorname{supp}(F)),$$

which is the intersection of the convex hulls of all supports of sources that could possibly generate the measurement f . By Eqs. (2) and (3),

$$\mathcal{C}(u - u_0)|_{\partial\Omega} \subset \operatorname{co}(\operatorname{supp}((1 - \sigma)\nabla u)) \subset \operatorname{co}(\operatorname{supp}(\sigma - \sigma_0)),$$

which means that the convex source support provides coarse, but reliable information about the position of the set $\operatorname{supp}(\sigma - \sigma_0)$. In fact, much more is known. The following theorem, e.g., can be found in [3].

Theorem 1 *We have $\mathcal{C}f = \emptyset$ if and only if $f = 0$, and for every $\epsilon > 0$, there exists $F_\epsilon \in L^2(\Omega)^d$ such that $LF_\epsilon = f$ and $\operatorname{dist}(\operatorname{co}(\operatorname{supp}(F_\epsilon)), \mathcal{C}f) < \epsilon$.*

3 An Optimization Problem in $\mathcal{H}_c(\mathbb{R}^d)$

For given data $f \in L^2_\diamond(\partial\Omega)$, we recast the computation of the convex source support as a minimization problem

$$\operatorname{vol}(D) = \min! \quad \text{subject to} \quad D \in \mathcal{H}_c(\mathbb{R}^d), \quad \mathcal{C}f \subset D \subset \Omega \quad (4)$$

in the space $\mathcal{K}_c(\mathbb{R}^d)$ of all nonempty convex and compact subsets of \mathbb{R}^d , which obviously has the unique solution $D^* = \mathcal{C}f$. To solve the problem (4), we mainly need a handy criterion to check whether $\mathcal{C}f \subset D$.

By Theorem 1, we have $\mathcal{C}f \subset \text{int } D$ if and only if there exists $F \in L^2(\Omega)^d$ with $\text{supp}(F) \subset D$ and $LF = f$. In other words, we have to check whether $f \in \mathcal{R}(L_D)$, i.e. whether f is in the range of the operator L_D , where

$$L_D : L^2(D)^d \rightarrow L^2_\diamond(\partial\Omega), \quad F \mapsto w|_{\partial\Omega},$$

and $w \in H^1_\diamond(\Omega)$ solves

$$\Delta w = \text{div } F \text{ in } \Omega, \quad \partial_\nu w = 0 \text{ on } \partial\Omega.$$

Proposition 1 *If the interior of $D \subseteq \Omega$ is not empty, then L_D is a compact linear operator with dense range, and*

$$f \notin \mathcal{R}(L_D) \quad \text{if and only if} \quad \lim_{\alpha \rightarrow 0} \|R_\alpha^D f\| = \infty.$$

where $R_\alpha^D f := (L_D^* L_D + \alpha I)^{-1} L_D^* f$.

Proof L_D is the concatenation of the linear bounded solution operator and the linear compact trace operator from $H^1_\diamond(\Omega)$ to $L^2_\diamond(\partial\Omega)$ and thus linear and compact. The adjoint of L_D is given by (see [4, Lemma 2])

$$L_D^* : L^2_\diamond(\partial\Omega) \rightarrow L^2(D)^n, \quad \varphi \mapsto \nabla v_0|_D,$$

where $v_0 \in H^1_\diamond(D)$ solves

$$\Delta v_0 = 0 \text{ in } \Omega, \quad \partial_\nu v_0|_{\partial\Omega} = \varphi \text{ on } \partial\Omega.$$

By unique continuation, L_D^* is injective and thus L_D has dense range. This also implies that the domain of definition of the Moore-Penrose inverse L_D^+ (cf. [2, Def. 2.2]) is given by

$$\mathcal{D}(L_D^+) = \mathcal{R}(L_D) + \mathcal{R}(L_D)^\perp = \mathcal{R}(L_D).$$

Since R_α^D is a linear regularization (the Tikhonov regularization, cf. [2, Section 5]), and a simple computation shows that

$$\sup_{\alpha > 0} \|L_D R_\alpha^D\| \leq 1,$$

it follows from standard regularization theory (cf., e.g., [2, Prop. 3.6]) that

$$\lim_{\alpha \rightarrow 0} R_\alpha^D f = L_D^+ f \quad \text{if } f \in \mathcal{D}(L_D^+) = \mathcal{R}(L_D),$$

and that

$$\lim_{\alpha \rightarrow 0} \|R_\alpha^D f\| = \infty \quad \text{if } f \notin \mathcal{D}(L_D^+) = \mathcal{R}(L_D).$$

This proves the assertion.

To implement Proposition 1, we therefore have to check whether the quantity

$$\begin{aligned} \|R_\alpha^D f\|^2 &= \|(L_D^* L_D + \alpha I)^{-1} L_D^* f\|^2 = \|L_D^* (L_D L_D^* + \alpha I)^{-1} f\|^2 \\ &= ((L_D L_D^* + \alpha I)^{-1} L_D L_D^* (L_D L_D^* + \alpha I)^{-1} f, f) \end{aligned}$$

remains bounded as $\alpha \rightarrow 0$. Writing $M_D := L_D L_D^* : L_\diamond^2(\partial\Omega) \rightarrow L_\diamond^2(\partial\Omega)$, we obtain the convenient representation

$$\|R_\alpha^D f\|^2 = ((M_D + \alpha I)^{-1} M_D (M_D + \alpha I)^{-1} f, f). \quad (5)$$

Fix an orthonormal basis $(\varphi_j)_j$ of $L_\diamond^2(\partial\Omega)$. The characterization of L_D^* in [4, Lemma 2] shows that

$$(M_D \varphi_j, \varphi_k) = (L_D^* \varphi_j, L_D^* \varphi_k) = \int_D \nabla u_0^j \cdot \nabla u_0^k dx, \quad (6)$$

where u_0^j solves

$$\Delta u_0^j = 0 \text{ in } \Omega, \quad \partial_\nu u_0^j = \varphi_j \text{ on } \partial\Omega. \quad (7)$$

Note that the integrands $\nabla u_0^j \cdot \nabla u_0^k$ do not depend on D and hence can be precomputed. Since

$$\int_D \nabla u_0^j \cdot \nabla u_0^k dx = \int_{\partial D} \partial_\nu u_0^j \cdot u_0^k ds$$

by the Gauß-Green theorem, even more computational effort can be shifted to the offline phase, provided the sets under consideration possess essentially finitely many different normals, which is the situation we consider in Sect. 4 and what follows.

Proposition 1 gives a mathematically rigorous criterion to check whether a set D contains the convex source support. In the following we describe a heuristic numerical implementation of this criterion and test it on a simple test example. Let us stress that we do not have any theoretical results on convergence or stability of the proposed numerical implementation, and that it is completely unclear whether such an implementation exists. Checking whether a function lies in the dense range of an infinite-dimensional operator seems intrinsically unstable to discretization errors and errors in the function or the operator. Likewise, it is unclear how to numerically check whether the sequence in Proposition 1 diverges or not.

In other words, the following heuristic numerical algorithm is motivated by a rigorous theoretical result but it is completely heuristic and we do not have any theoretical justification for this algorithm. Since, to the knowledge of the authors, no convergent numerical methods are known for the considered problem, we believe that this algorithm might still be of interest and serve as a first step towards mathematically rigorously justified algorithms.

To heuristically check, whether $\|R_\alpha^D f\| \rightarrow \infty$, we fix suitable constants $\alpha, C > 0$ and $N \in \mathbb{N}$. Consider the finite-dimensional subspace $V_N := \text{span}(\varphi_1, \dots, \varphi_N)$ of $L_\diamond^2(\partial\Omega)$ and the corresponding L^2 orthogonal projector $P_N : L_\diamond^2(\partial\Omega) \rightarrow V_N$. Instead of M_D , we consider the truncated operator $M_D^N := P_N M|_{V_N} : V_N \rightarrow V_N$, which satisfies

$$(M_D^N \varphi_j, \varphi_k) = (P_N M_D \varphi_j, \varphi_k) = (M_D \varphi_j, \varphi_k) \quad \text{for } 1 \leq j, k \leq N,$$

so that formula (6) holds for M_D^N as well. We define

$$\|R_{\alpha,N}^D v\|^2 := ((M_D^N + \alpha I)^{-1} M_D^N (M_D^N + \alpha I)^{-1} P_N v, P_N v) \quad \text{for all } v \in L_\diamond^2(\partial\Omega)$$

and note that

$$R_{\alpha,N}^D f \rightarrow R_\alpha^D f \quad \text{as } N \rightarrow \infty$$

follows from a discussion, which involves a variant of the Banach Lemma. Therefore, the use of the criterion

$$\|R_{\alpha,N}^D f\|^2 \leq C$$

instead of $\|R_\alpha^D f\| \rightarrow \infty$ is well-motivated, and we solve

$$\text{vol}(D) = \min! \quad \text{subject to } D \in \mathcal{K}_c(\mathbb{R}^d), D \subset \Omega, \|R_{\alpha,N}^D f\|^2 \leq C. \quad (8)$$

with, e.g., $\alpha = 10^{-4}$ and $C = 10^6$ instead of (4).

4 Galerkin Approximations to $\mathcal{K}_c(\mathbb{R}^2)$

We outline a setting for a first-discretize-then-optimize approach to numerical optimization in the space $\mathcal{K}_c(\mathbb{R}^2)$, which we use to solve problem (8). To this end, we define Galerkin subspaces of $\mathcal{K}_c(\mathbb{R}^2)$ in terms of polytopes with prescribed sets of outer normals. These spaces have good global approximation properties (see Proposition 2), they possess a unique representation in terms of few coordinates, and their sets of admissible coordinates are characterized by sparse linear inequalities. A theory of these spaces in arbitrary dimension is work in progress.

Fix a matrix $A \in \mathbb{R}^{m \times 2}$ with rows a_i^T , $i = 1, \dots, m$, where $a_i \in \mathbb{R}^2$, $\|a_i\|_2 = 1$ for all i and $a_i \neq a_j$ for all i, j with $i \neq j$. For every $b \in \mathbb{R}^m$, we consider the convex polyhedron

$$Q_{A,b} := \{x \in \mathbb{R}^2 : Ax \leq b\},$$

and we define a space $\mathcal{G}_A \subset \mathcal{K}_c(\mathbb{R}^2)$ of convex polyhedra by setting

$$\mathcal{G}_A := \{Q_{A,b} : b \in \mathbb{R}^m\} \setminus \{\emptyset\}.$$

The choice of these spaces is motivated by an approximation result from [8]. Recall the definition of the one-sided Hausdorff distance

$$\text{dist} : \mathcal{K}_c(\mathbb{R}^2) \times \mathcal{K}_c(\mathbb{R}^2) \rightarrow \mathbb{R}_+, \quad \text{dist}(D, D') := \sup_{x \in D} \inf_{x' \in D'} \|x - x'\|_2.$$

Proposition 2 *Assume that the matrix $A \in \mathbb{R}^{m \times 2}$ satisfies*

$$\delta := \max_{x \in \mathbb{R}^2, \|x\|_2=1} \text{dist}(\{x\}, \{a_1^T, \dots, a_m^T\}) < 1. \quad (9)$$

Then the associated space \mathcal{G}_A consists of convex polytopes, and for all $D \in \mathcal{K}_c(\mathbb{R}^2)$, there exists $Q_{A,b} \in \mathcal{G}_A$ such that $D \subset Q_{A,b}$ and

$$\text{dist}(Q_{A,b}, D) \leq \frac{2\delta - \delta^2}{1 - \delta} \text{dist}(D, \{0\}).$$

Hence, if the matrix A is augmented in such a way that $\delta \rightarrow 0$ as $m \rightarrow \infty$, then \mathcal{G}_A converges to $\mathcal{K}_c(\mathbb{R}^2)$ uniformly on every bounded subset of $\mathcal{K}_c(\mathbb{R}^2)$.

It is, at present, not entirely clear how to represent the spaces \mathcal{G}_A in terms of coordinates. There are $b \in \mathbb{R}^m$ with $Q_{A,b} = \emptyset$, and two different vectors $b, b' \in \mathbb{R}^m$ may encode the same polytope $Q_{A,b} = Q_{A,b'}$. In our concrete optimization problem, the constraint $\mathcal{C}f \subset D$ will enforce $Q_{A,b} \neq \emptyset$. For the time being, we treat the second issue by forcing all hyperplanes $\{x \in \mathbb{R}^2 : a_k^T x = b_k\}$, $k = 1, \dots, m$, to possess at least one common point with $Q_{A,b}$. This approach will be made rigorous in the future.

Definition 1 We call the set \mathcal{C}_A of all $b \in \mathbb{R}^m$ satisfying

$$p_1 \begin{pmatrix} b_m \\ b_2 \end{pmatrix} \geq b_1, \quad p_k \begin{pmatrix} b_{k-1} \\ b_{k+1} \end{pmatrix} \geq b_k, \quad k = 2, \dots, m-1, \quad \text{and} \quad p_m \begin{pmatrix} b_{m-1} \\ b_1 \end{pmatrix} \geq b_m$$

with

$$p_1 := a_1^T \begin{pmatrix} a_m^T \\ a_2^T \end{pmatrix}^{-1}, \quad p_m := a_m^T \begin{pmatrix} a_{m-1}^T \\ a_1^T \end{pmatrix}^{-1},$$

$$p_k := a_k^T \begin{pmatrix} a_{k-1}^T \\ a_{k+1}^T \end{pmatrix}^{-1}, \quad k = 2, \dots, m-1$$

the set of admissible coordinates.

Note that the inverse matrices above exist when $\delta < 1$ as required in Proposition 2. Hence it is easy to assemble the sparse matrix

$$H_A = \begin{pmatrix} 1 & -p_{1,2} & & & -p_{1,1} \\ -p_{2,1} & 1 & -p_{2,2} & & \\ & \ddots & \ddots & \ddots & \\ & & -p_{m-1,1} & 1 & -p_{m-1,2} \\ -p_{m,2} & & & -p_{m,1} & 1 \end{pmatrix},$$

which gives rise to the following characterization of the set $\mathcal{C}_A \subset \mathbb{R}^m$.

Lemma 1 *The set of admissible coordinates can be written as*

$$\mathcal{C}_A = \{b \in \mathbb{R}^m : H_A b \leq 0\}.$$

All in all, we replaced the relatively inaccessible space $\mathcal{K}_c(\mathbb{R}^2)$ with a Galerkin subspace \mathcal{G}_A that is parametrized over a set $\mathcal{C}_A \subset \mathbb{R}^m$ of coordinates, which, in turn, is described by a sparse linear inequality. For the practical computations in this paper, we fix the matrix $A = (a_1, \dots, a_m)^T$ given by

$$a_k = (\cos(2k\pi/m), \sin(2k\pi/m))^T, \quad k = 1, \dots, m,$$

which is probably the best choice in the absence of detailed information on the set to be approximated. As we will have $\Omega = B_1(0)$ in our computational example, we will replace problem (4) with the fully discrete optimization problem

$$\left. \begin{array}{l} \text{vol}(Q_{A,b}) = \min! \\ \text{subject to } b \in \mathbb{R}^m, H_A b \leq 0, b \leq 1, \|R_{\alpha,N}^{Q_{A,b}} f\|^2 \leq C. \end{array} \right\} \quad (10)$$

5 Gradients of Functions on \mathcal{G}_A

The objective function $D \mapsto \text{vol}(D)$ and the constraint $D \mapsto \|R_{\alpha, N}^D f\|^2$ are both given in terms of integrals of a real-valued function over the set D . The evaluation of these integrals is straight-forward and efficient. The efficient evaluations of the gradients of both integrals with respect to coordinates requires some preparation. We follow [6, Lemma 2.2] and [7, Theorem 1].

Proposition 3 *Let $b \in \mathcal{C}_A$, let $k \in \{1, \dots, m\}$, and let $Q_{A,b}^k$ be the facet*

$$Q_{A,b}^k := Q_{A,b} \cap \{x \in \mathbb{R}^2 : a_k^T x = b_k\}.$$

If we assume that $\text{vol}_2(Q_{A,b}) > 0$ and $\text{vol}_1(Q_{A,b}^k) > 0$, then for any continuous function $h : \mathbb{R}^2 \rightarrow \mathbb{R}$ we have

$$\frac{d}{db_k} \int_{Q_{A,b}} h(x) dx = \int_{Q_{A,b}^k} h(\xi) d\xi.$$

The above proposition shows that whenever $Q_{A,b}$ is not degenerate, we have

$$\nabla_b \text{vol}_2(Q_{A,b}) = (\text{vol}_1(Q_{A,b}^1), \dots, \text{vol}_1(Q_{A,b}^m))^T.$$

To compute $\nabla_b \|R_{\alpha, N}^{Q_{A,b}} f\|^2$, we need the following lemma. The construction of the matrices P , S and U reduces the costs for the computation of the desired derivative.

Lemma 2 *Let $\varepsilon > 0$, let $M : (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}^{N \times N}$, $\gamma \mapsto M(\gamma)$, be differentiable with $M(\gamma)$ symmetric and $M(\gamma) + \alpha I$ invertible for all $\gamma \in (-\varepsilon, \varepsilon)$. Using the abbreviations*

$$X := M + \alpha I, \quad Y := X^{-1} M' X^{-1} \quad \text{and} \quad Z := M X^{-1},$$

we find that

$$(X^{-1} M X^{-1})' = -Y Z + Y - (Y Z)^T.$$

The proof is elementary and therefore omitted. An application of Lemma 2 to the matrix representation of $M_{Q_{A,b}}^N$ yields

$$\begin{aligned} \frac{d}{db_k} \|R_{\alpha, N}^{Q_{A,b}} f\|^2 &= \frac{d}{db_k} ((M_{Q_{A,b}}^N + \alpha I)^{-1} M_{Q_{A,b}}^N (M_{Q_{A,b}}^N + \alpha I)^{-1} f, f) \\ &= ((-Y_k Z + Y_k - (Y_k Z)^T) f, f) \end{aligned}$$

with the abbreviations

$$X := M_{Q_{A,b}}^N + \alpha I, \quad Y_k := X^{-1} \left(\frac{d}{db_k} M_{Q_{A,b}}^N \right) X^{-1} \quad \text{and} \quad Z := M_{Q_{A,b}}^N X^{-1},$$

where

$$\left[\frac{d}{db_k} M_{Q_{A,b}}^N \right]_{ij} = \int_{Q_{A,b}^k} \nabla u_0^i \cdot \nabla u_0^j \, d\xi$$

by Proposition 3. Thus we obtain a formula for $\nabla_b \|R_{\alpha,N}^{Q_{A,b}} f\|^2$ that is not only more precise than a numerical approximation by finite differences, but also much cheaper to compute, because the area of integration is just a lower-dimensional surface.

6 A First Numerical Simulation

We test our numerical algorithm on a simple 2d example, where all quantities are known explicitly and the algorithm can be observed under controlled conditions. Let $\Omega = B_1(0)$ be the unit circle and let $\sigma_0 \equiv 1$. We consider a point inhomogeneity which leads to a difference potential (cf. [1])

$$w(x) = \frac{1}{\pi} \frac{\langle z^* - x, \eta \rangle}{\|z^* - x\|_2^2}, \quad x \in B_1(0),$$

that solves the partial differential equation

$$\Delta w = \eta \cdot \nabla \delta_{z^*} \text{ in } \Omega, \quad \sigma_0 \partial_\nu w|_{\partial\Omega} = 0,$$

where z^* is the location of the point inhomogeneity, and $\eta \in \mathbb{R}^2$, $\|\eta\|_2 = 1$ is a dipole orientation vector depending on the applied current pattern. Using a standard smoothing argument, it is easily checked (see, e.g., [4]) that for each open set O containing z^* there exists $F \in L^2(O)^2$ so that

$$\Delta w = \operatorname{div} F.$$

Hence the convex source support of the difference measurement $w|_{\partial\Omega}$ is the inhomogeneity location z^* . In our example we used $z^* = (\frac{3}{10}, \frac{3}{10})^T$ and $\eta = (1, 0)^T$.

In the following computations, it is convenient to switch between standard coordinates (x_1, x_2) and polar coordinates (r, ξ) . Consider the basis

$$\varphi_{2j}(\xi) = \frac{1}{\sqrt{\pi}} \cos(j\xi), \quad \varphi_{2j+1}(\xi) = \frac{1}{\sqrt{\pi}} \sin(j\xi), \quad j \in \mathbb{N}_1,$$

of $L_{\diamond}^2(\partial\Omega)$. Since the Laplace operator satisfies

$$u_{x_1x_1} + u_{x_2x_2} = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\xi\xi},$$

it is easy to see that the corresponding solutions of problem (7) are

$$u_0^{2j}(r, \xi) = \frac{1}{j\sqrt{\pi}} \cos(j\xi)r^j, \quad u_0^{2j+1}(r, \xi) = \frac{1}{j\sqrt{\pi}} \sin(j\xi)r^j, \quad j \in \mathbb{N}_1.$$

Since the gradient satisfies

$$u_{x_1} = \cos \xi \cdot u_r + \frac{1}{r} \sin \xi \cdot u_{\xi}, \quad u_{x_2} = \sin \xi \cdot u_r + \frac{1}{r} \cos \xi \cdot u_{\xi},$$

we have explicit representations

$$\begin{aligned} \frac{d}{dx_1}u_0^{2j} &= \frac{r^{j-1}}{\sqrt{\pi}} \cos((j+1)\xi), & \frac{d}{dx_2}u_0^{2j} &= -\frac{r^{j-1}}{\sqrt{\pi}} \sin((j-1)\xi), \\ \frac{d}{dx_1}u_0^{2j+1} &= \frac{r^{j-1}}{\sqrt{\pi}} \sin((j+1)\xi), & \frac{d}{dx_2}u_0^{2j+1} &= \frac{r^{j-1}}{\sqrt{\pi}} \cos((j-1)\xi). \end{aligned}$$

Now we fix the matrix $A = (a_1, \dots, a_8)^T$ given by

$$a_k = (\cos(k\pi/4), \sin(k\pi/4))^T, \quad k = 1, \dots, 8,$$

and solve optimization problem (4) approximately by applying Matlab's interior point method to problem (10) with initial value $b_0 = (\frac{4}{5}, \dots, \frac{4}{5})^T$ and $N = 6$, computing values and gradients of the objective $b \mapsto \text{vol}_2(Q_{A,b})$ and the constraint $b \mapsto \|R_{\alpha,N}^{Q_{A,b}} w|_{\partial\Omega}\|^2$ as in Sect. 5. The results and the computation times on an ordinary desktop computer are displayed in Fig. 1.

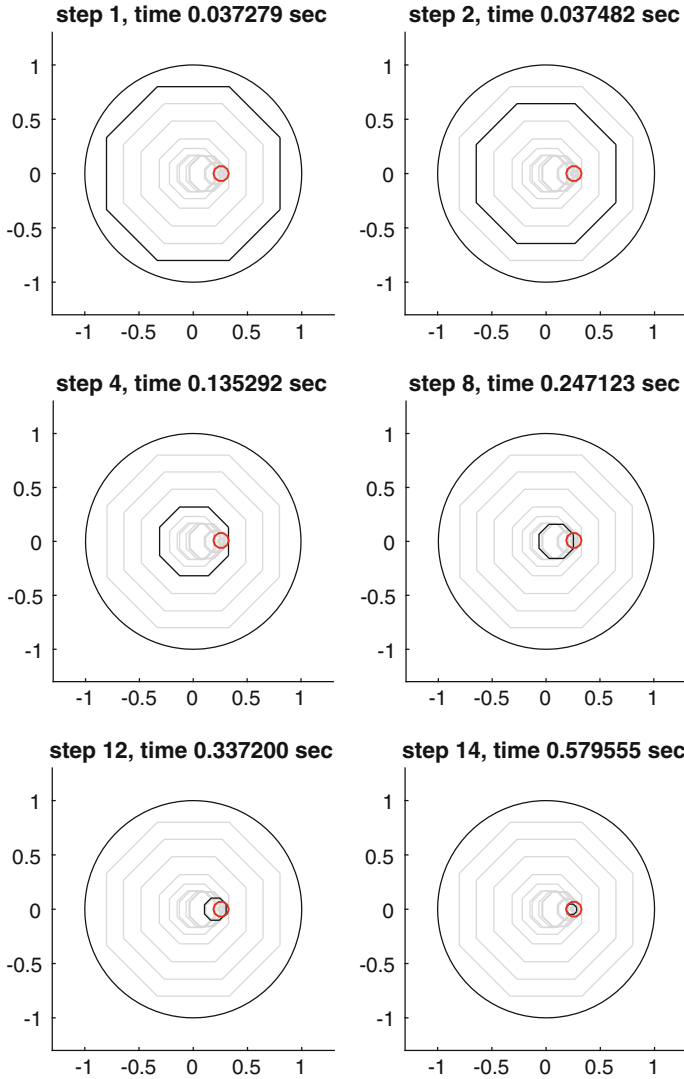


Fig. 1 Selected iterates of Matlab's interior point optimization tool applied to (10) with data specified in Sect. 6. The current iterate is highlighted in black. The position of the dipole is the center of the red circle

References

1. Ammari, H., Griesmaier, R., Hanke, M.: Identification of small inhomogeneities: asymptotic factorization. *Math. Comp.* **76**(259), 1425–1448 (2007)
2. Engl, H.W., Hanke, M., Neubauer, A.: *Regularization of Inverse Problems*. Springer Science & Business Media, Dordrecht (1996)

3. Hanke, M, Hyvönen, N, Reusswig, S.: Convex source support and its applications to electric impedance tomography. *SIAM J. Imaging Sci.* **1**(4), 364–378 (2008)
4. Harrach, B.: Recent progress on the factorization method for electrical impedance tomography. *Comput. Math. Methods Med.* **8** (2013), Art. ID 425184
5. Kusiak, S, Sylvester, J.: The scattering support. *Commun. Pure Appl. Math.* **56**(11), 1525–1548 (2003)
6. Lasserre, J.B.: Integration on a convex polytope. *Proc. Am. Math. Soc.* **126**(8), 2433–2441 (1998)
7. Müller, C, Feuer, A, Goodwin, G.C.: Derivative of an integral over a convex polytope. *Appl. Math. Lett.* **24**(7), 1120–1123 (2011)
8. Rieger, J.: Discretizations of linear elliptic partial differential inclusions. *Numer. Funct. Anal. Optim.* **32**(8), 904–925 (2011)

Local Volatility Calibration by Optimal Transport



Ivan Guo, Grégoire Loeper, and Shiyi Wang

Abstract The calibration of volatility models from observable option prices is a fundamental problem in quantitative finance. The most common approach among industry practitioners is based on the celebrated Dupire’s formula, which requires the knowledge of vanilla option prices for a continuum of strikes and maturities that can only be obtained via some form of price interpolation. In this paper, we propose a new local volatility calibration technique using the theory of optimal transport. We formulate a time continuous martingale optimal transport problem, which seeks a martingale diffusion process that matches the known densities of an asset price at two different dates, while minimizing a chosen cost function. Inspired by the seminal work of Benamou and Brenier, we formulate the problem as a convex optimization problem, derive its dual formulation, and solve it numerically via an augmented Lagrangian method and the alternative direction method of multipliers (ADMM) algorithm. The solution effectively reconstructs the dynamic of the asset price between the two dates by recovering the optimal local volatility function, without requiring any time interpolation of the option prices.

1 Introduction

A fundamental assumption of the classical Black-Scholes option pricing framework is that the underlying risky asset has a constant volatility. However, this assumption can be easily dispelled by the option prices observed in the market, where the implied volatility surfaces are known to exhibit “skews” or “smiles”. Over the

I. Guo · G. Loeper

School of Mathematical Sciences, Monash University, Clayton, VIC, Australia

Centre for Quantitative Finance and Investment Strategies, Monash University, Clayton, VIC, Australia

S. Wang (✉)

School of Mathematical Sciences, Monash University, Clayton, VIC, Australia

e-mail: Leo.Wang@monash.edu

years, many sophisticated volatility models have been introduced to explain this phenomenon. One popular class of model is the local volatility models. In a local volatility model, the volatility function $\sigma(t, S_t)$ is a function of time t as well as the asset price S_t . The calibration of the local volatility function involves determining σ from available option prices.

One of the most prominent approaches for calibrating local volatility is introduced by the path-breaking work of Dupire [6], which provides a method to recover the local volatility function $\sigma(t, s)$ if the prices of European call options $C(T, K)$ are known for a continuum of maturities T and strikes K . In particular, the famous Dupire's formula is given by

$$\sigma^2(T, K) = \frac{\frac{\partial C(T, K)}{\partial T} + \mu_t K \frac{\partial C(T, K)}{\partial K}}{\frac{K^2}{2} \frac{\partial^2 C(T, K)}{\partial K^2}}, \quad (1)$$

where μ_t is a deterministic function. However, in practice, option prices are only available at discrete strikes and maturities, hence interpolation is required in both variables to utilize this formula, leading to many inaccuracies. Furthermore, the numerical evaluation of the second derivative in the denominator can potentially cause instabilities in the volatility surface as well as singularities. Despite these drawbacks, Dupire's formula and its variants are still used prevalently in the financial industry today.

In this paper, we introduce a new technique for the calibration of local volatility functions that adopts a variational approach inspired by optimal transport. The optimal transport problem was first proposed by Monge [10] in 1781 in the context of civil engineering. The basic problem is to transfer material from one site to another while minimizing transportation cost. In the 1940s, Kantorovich [8] provided a modern treatment of the problem based on linear programming techniques, leading to the so-called Monge-Kantorovich problem. Since then, the theory of optimal transport has attracted considerable attention with applications in many areas such as fluid dynamics, meteorology and econometrics (see, e.g., [7] and [14]). Recently, there have been a few studies extending optimal transport to stochastic settings with applications in financial mathematics. For instance, Tan and Touzi [13] studied an extension of the Monge-Kantorovich problem for semimartingales, while Dolinsky and Soner [5] applied martingale optimal transport to the problem of robust hedging.

In our approach, we begin by recovering the probability density of the underlying asset at times t_0 and t_1 from the prices of European options expiring at t_0 and t_1 . Then, instead of interpolating between different maturities, we seek a martingale diffusion process which transports the density from t_0 to t_1 , while minimizing a particular cost function. This is similar to the classical optimal transport problem, with the additional constraint that the diffusion process must be a martingale driven by a local volatility function. In the case where the cost function is convex, we find that the problem can be reformulated as a convex optimization problem under linear constraints. Theoretically, the stochastic control problem can be reformulated as an

optimization problem which involves solving a non-linear PDE at each step, and the PDE is closely connected with the ones studied in Bouchard et al. [2, 3] and Loeper [9] in the context of option pricing with market impact. For this paper, we approach the problem via the augmented Lagrangian method and the alternative direction method of multipliers (ADMM) algorithm, which was also used in Benamou and Brenier [1] for classical optimal transport problems.

The paper is organized as follows. In Sect. 2, we introduce the classical optimal transport problem as formulated by Benamou and Brenier [1]. In Sect. 3, we introduce the martingale optimal transport problem and its augmented Lagrangian. The numerical method is detailed in Sect. 4 and numerical results are given in Sect. 5.

2 Optimal Transport

In this section, we briefly outline the optimal transport problem as formulated by Benamou and Brenier [1]. Given density functions $\rho_0, \rho_1 : \mathbb{R}^d \rightarrow [0, \infty)$ with equal total mass $\int_{\mathbb{R}^d} \rho_0(x) dx = \int_{\mathbb{R}^d} \rho_1(x) dx$. We say that a map $s : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is an admissible transport plan if it satisfies

$$\int_{x \in A} \rho_1(x) dx = \int_{s(x) \in A} \rho_0(x) dx, \quad (2)$$

for all bounded subset $A \subset \mathbb{R}^d$. Let \mathcal{T} denote the collection of all admissible maps. Given a cost function $c(x, y)$, which represents the transportation cost of moving one unit of mass from x to y , the optimal transport problem is to find an optimal map $s^* \in \mathcal{T}$ that minimizes the total cost

$$\inf_{s \in \mathcal{T}} \int_{\mathbb{R}^d} c(x, s(x)) \rho_0(x) dx. \quad (3)$$

In particular, when $c(x, y) = |y - x|^2$ where $|\cdot|$ denotes the Euclidean norm, this problem is known as the L^2 Monge-Kantorovich problem (MKP).

The L^2 MKP is reformulated in [1] in a fluid mechanics framework. In the time interval $t \in [0, 1]$, consider all possible smooth, time-dependent, densities $\rho(t, x) \geq 0$ and velocity fields $v(t, x) \in \mathbb{R}^d$, that satisfy the continuity equation

$$\partial_t \rho(t, x) + \nabla \cdot (\rho(t, x) v(t, x)) = 0, \quad \forall t \in [0, 1], \quad \forall x \in \mathbb{R}^d, \quad (4)$$

and the initial and final conditions

$$\rho(0, x) = \rho_0, \quad \rho(1, x) = \rho_1. \quad (5)$$

In [1], it is proven that the L^2 MKP is equivalent to finding an optimal pair (ρ^*, v^*) that minimizes

$$\inf_{\rho, v} \int_{\mathbb{R}^d} \int_0^1 \rho(t, x) |v(t, x)|^2 dt dx, \quad (6)$$

subject to the constraints (4) and (5). This problem is then solved numerically in [1] via an augmented Lagrangian approach. The specific numerical algorithm used is known as the alternative direction method of multipliers (ADMM), which has applications in statistical learning and distributed optimization.

3 Formulation

3.1 The Martingale Problem

Let $(\Omega, \mathbb{F}, \mathbb{Q})$ be a probability space, where \mathbb{Q} is the risk-neutral measure. Suppose the dynamic of an asset price X_t on $t \in [0, 1]$ is given by the local volatility model

$$dX_t = \sigma(t, X_t) dW_t, \quad t \in [0, 1], \quad (7)$$

where $\sigma(t, x)$ is a local volatility function and W_t is a one-dimensional Brownian motion. For the sake of simplicity, suppose the interest and dividend rates are zero. Denote by $\rho(t, x)$ the density function of X_t and $\gamma(t, x) = \sigma(t, x)^2/2$ the diffusion coefficient. It is well known that $\rho(t, x)$ follows the Fokker-Planck equation

$$\partial_t \rho(t, x) - \partial_{xx}(\rho(t, x)\gamma(t, x)) = 0. \quad (8)$$

Suppose that the initial and the final densities are given by $\rho_0(x)$ and $\rho_1(x)$, which are recovered from European option prices via the Breeden-Litzenberger [4] formula,

$$\rho_T(K) = \frac{\partial^2 C(T, K)}{\partial K^2}.$$

Let $F : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a convex cost function. We are interested in minimizing the quantity

$$\mathbb{E} \left(\int_0^1 F(\gamma(t, X_t)) dt \right) = \int_D \int_0^1 \rho(t, x) F(\gamma(t, X_t)) dt dx,$$

where $F(x) = +\infty$ if $x < 0$, and $D \subseteq \mathbb{R}$ is the support of $\{X_t, t \in [0, 1]\}$. Unlike the classical optimal transport problem, the existence of a solution here requires an additional condition: there exists a martingale transport plan if and only if ρ_0 and ρ_1

satisfy:

$$\int_{\mathbb{R}} \varphi(x) \rho_0(x) dx \leq \int_{\mathbb{R}} \varphi(x) \rho_1(x) dx,$$

for all convex function $\varphi(x) : \mathbb{R} \rightarrow \mathbb{R}$. This is known as Strassen's Theorem [12]. This condition is naturally satisfied by financial models in which the asset price follows a martingale diffusion process.

Remark 1 The formulation here is actually quite general and it can be easily adapted to a large family of models. For example, the case of a geometric Brownian motion with local volatility can be recovered by substituting $\tilde{\sigma}(t, X_t)X_t = \sigma(t, X_t)$ everywhere, including in the Fokker-Planck equation. The cost function F would then also be dependent on x . The later arguments involving convex conjugates still hold since F remains a convex function of $\tilde{\sigma}$.

Since $\rho F(\gamma)$ is not convex in (ρ, γ) (which is crucial for our method), the substitution $m(t, x) := \rho(t, x)\gamma(t, x)$ is applied. So we obtain the following *martingale optimal transport problem*:

$$\inf_{\rho, m} \int_D \int_0^1 \rho(t, x) F\left(\frac{m(t, x)}{\rho(t, x)}\right) dt dx, \quad (9)$$

subject to the constraints:

$$\rho(0, x) = \rho_0(x), \quad \rho(1, x) = \rho_1(x), \quad (10)$$

$$\partial_t \rho(t, x) - \partial_{xx} m(t, x) = 0. \quad (11)$$

Using the convexity of F , the term $\rho F(m/\rho)$ can be easily verified to be convex in (ρ, m) . Also note that we have the natural restrictions of $\rho > 0$ and $m \geq 0$. Note that $m \geq 0$ is enforced by penalizing the cost function F , and $\rho > 0$ will be encoded in the convex conjugate formulation. (see Proposition 1)

Next, introduce a time-space dependent Lagrange multiplier $\phi(t, x)$ for the constraints (10) and (11). Hence the associated Lagrangian is

$$L(\phi, \rho, m) = \int_{\mathbb{R}} \int_0^1 \rho(t, x) F\left(\frac{m(t, x)}{\rho(t, x)}\right) + \phi(t, x) (\partial_t \rho(x) - \partial_{xx}(m(t, x))) dt dx. \quad (12)$$

Integrating (12) by parts and letting $m = \rho\gamma$ vanish on the boundaries of D , the martingale optimal transport problem can be reformulated as the following saddle

point problem:

$$\begin{aligned} \inf_{\rho, m} \sup_{\phi} L(\phi, \rho, m) &= \inf_{\rho, m} \sup_{\phi} \int_D \int_0^1 \left(\rho F\left(\frac{m}{\rho}\right) - \rho \partial_t \phi - m \partial_{xx} \phi \right) dt dx \\ &\quad - \int_D (\phi(0, x) \rho_0 - \phi(1, x) \rho_1) dx. \end{aligned} \quad (13)$$

As shown by Theorem 3.6 in [13], (13) has an equivalent dual formulation which leads to the following representation:

$$\begin{aligned} \sup_{\phi} \inf_{\rho, m} L(\phi, \rho, m) &= \sup_{\phi} \inf_{\rho} \int_D \int_0^1 -\rho (\partial_t \phi + F^*(\partial_{xx} \phi)) dt dx \\ &\quad - \int_D (\phi(0, x) \rho_0 - \phi(1, x) \rho_1) dx. \end{aligned} \quad (14)$$

In particular, the optimal ϕ must satisfy the condition

$$\partial_t \phi + F^*(\partial_{xx} \phi) = 0, \quad (15)$$

where F^* is the convex conjugate of F (see (16) and Proposition 1). We will later use (15) to check the optimality of our algorithm.

3.2 Augmented Lagrangian Approach

Similar to [1], we solve the martingale optimal transport problem using the augmented Lagrangian approach. Let us begin by briefly recalling the well-known definition and properties of the convex conjugate. For more details, the readers are referred to Section 12 of Rockafellar [11].

Fix $D \subseteq \mathbb{R}^d$, let $f : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ be a proper convex and lower semi-continuous function. Then the *convex conjugate* of f is the function $f^* : \mathbb{R}^d \rightarrow \mathbb{R} \cup \{+\infty\}$ defined by

$$f^*(y) := \sup_{x \in \mathbb{R}^d} (x \cdot y - f(x)). \quad (16)$$

The convex conjugate is also often known as the *Legendre-Fenchel transform*.

Proposition 1 *We have the following properties:*

- (i) f^* is a proper convex and lower semi-continuous function with $f^{**} \equiv f$;
- (ii) if f is differentiable, then $f(x) + f^*(f'(x)) = x f'(x)$.

Returning to the problem at hand, recall that $G(x, y) := xF(y/x)$, $x > 0$ is convex in (x, y) . By adopting the convention of $G(x, y) = \infty$ whenever $x \leq 0$, it can be expressed in terms of the convex conjugate, as shown in the following proposition.

Proposition 2 *Denote by F^* the convex conjugate of F .*

(i) *Let $G(x, y) = xF(y/x)$, the convex conjugate of G is given by:*

$$G^*(a, b) = \begin{cases} 0, & \text{if } a + F^*(b) \leq 0, \\ \infty, & \text{otherwise.} \end{cases} \quad (17)$$

(ii) *For $x > 0$, We have the following equality,*

$$xF\left(\frac{y}{x}\right) = \sup_{(a,b) \in \mathbb{R}^2} \{ax + by : a + F^*(b) \leq 0\}. \quad (18)$$

Proof

(i) By definition, the convex conjugate of G is given by

$$G^*(a, b) = \sup_{(x,y) \in \mathbb{R}^2} \left\{ ax + by - xF\left(\frac{y}{x}\right) : x > 0 \right\} \quad (19)$$

$$= \sup_{(x,y) \in \mathbb{R}^2} \left\{ ax + x \left(b \frac{y}{x} - F\left(\frac{y}{x}\right) \right) : x > 0 \right\} \quad (20)$$

$$= \sup_{x > 0} \{ x(a + F^*(b)) \}, \quad (21)$$

If $a + F^*(b) \leq 0$, the supremum is achieved by limit $x \rightarrow 0$, otherwise, G^* becomes unbounded as x increases. This establishes part (i).

(ii) The required equality follows immediately from part (i) and the fact that

$$xF\left(\frac{y}{x}\right) = \sup_{(a,b) \in \mathbb{R}^2} \{ ax + by - G^*(a, b) : a + F^*(b) \leq 0 \}.$$

□

Now we are in a position to present the augmented Lagrangian. First, let us introduce the following notations:

$$K = \left\{ (a, b) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R} \times \mathbb{R} \mid a + F^*(b) \leq 0 \right\}, \quad (22)$$

$$\mu = (\rho, m) = (\rho, \rho\gamma), \quad q = (a, b), \quad \langle \mu, q \rangle = \int_D \int_0^1 \mu \cdot q, \quad (23)$$

$$H(q) = G^*(a, b) = \begin{cases} 0, & \text{if } q \in K, \\ \infty, & \text{otherwise,} \end{cases} \quad (24)$$

$$J(\phi) = \int_D [\phi(0, x)\rho_0 - \phi(1, x)\rho_1], \quad (25)$$

$$\nabla_{t,xx} = (\partial_t, \partial_{xx}). \quad (26)$$

By using the above notations, we can express the equality from Proposition 2 (ii) in the following way,

$$\rho F\left(\frac{m}{\rho}\right) = \sup_{\{a,b\} \in K} \{a\rho + bm\} = \sup_{q \in K} \{\mu \cdot q\}. \quad (27)$$

Since the restriction $q \in K$ is checked point-wise for every (t, x) , we can exchange the supremum with the integrals in the following equality

$$\int_D \int_0^1 \sup_{q \in K} \{\mu \cdot q\} = \sup_q \left\{ -H(q) + \int_D \int_0^1 \mu \cdot q \right\} = \sup_q \left\{ -H(q) + \langle \mu, q \rangle \right\}. \quad (28)$$

Therefore, the saddle point problem specified by (13) can be rewritten as

$$\sup_{\mu} \inf_{\phi, q} \left\{ H(q) + J(\phi) + \langle \mu, \nabla_{t,xx}\phi - q \rangle \right\}. \quad (29)$$

Note that in the new saddle point problem (29), μ is the Lagrange multiplier of the new constraint $\nabla_{t,xx}\phi = q$. In order to turn this into a convex problem, we define the augmented Lagrangian as follows:

$$L_r(\phi, q, \mu) = H(q) + J(\phi) + \langle \mu, \nabla_{t,xx}\phi - q \rangle + \frac{r}{2} \langle \nabla_{t,xx}\phi - q, \nabla_{t,xx}\phi - q \rangle, \quad (30)$$

where $r > 0$ is a penalization parameter. The saddle point problem then becomes

$$\sup_{\mu} \inf_{\phi, q} L_r(\phi, q, \mu), \quad (31)$$

which has the same solution as (13).

4 Numerical Method

In this section, we describe in detail the alternative direction method of multipliers (ADMM) algorithm for solving the saddle point problem given by (30) and (31). In each iteration, using $(\phi^{n-1}, q^{n-1}, \mu^{n-1})$ as a starting point, the ADMM algorithm

performs the following three steps:

$$\text{Step A: } \phi^n = \arg \min_{\phi} L_r(\phi, q^{n-1}, \mu^{n-1}), \quad (32)$$

$$\text{Step B: } q^n = \arg \min_q L_r(\phi^n, q, \mu^{n-1}), \quad (33)$$

$$\text{Step C: } \mu^n = \arg \max_{\mu} L_r(\phi^n, q^n, \mu). \quad (34)$$

Step A: $\phi^n = \arg \min_{\phi} L_r(\phi, q^{n-1}, \mu^{n-1})$

To find the function ϕ^n that minimizes $L_r(\phi, q^{n-1}, \mu^{n-1})$, we set the functional derivative of L_r with respect to ϕ to zero:

$$J(\phi) + \langle \mu^{n-1}, \nabla_{t,xx} \phi \rangle + r \langle \nabla_{t,xx} \phi^n - q^{n-1}, \nabla_{t,xx} \phi \rangle = 0. \quad (35)$$

By integrating by parts, we arrive at the following variational equation

$$-r(\partial_{tt} \phi^n - \partial_{xxxx} \phi^n) = \partial_t(\rho^{n-1} - r a^{n-1}) - \partial_{xx}(m^{n-1} - r b^{n-1}), \quad (36)$$

with Neumann boundary conditions in time $\forall x \in D$:

$$r \partial_t \phi^n(0, x) = \rho_0 - \rho^{n-1}(0, x) + r a^{n-1}(0, x), \quad (37)$$

$$r \partial_t \phi^n(1, x) = \rho_1 - \rho^{n-1}(1, x) + r a^{n-1}(1, x). \quad (38)$$

For the boundary conditions in space, let $D = [\underline{D}, \overline{D}]$. We give the following boundary condition to the diffusion coefficient:

$$\gamma(t, \underline{D}) = \gamma(t, \overline{D}) = \overline{\gamma} := \arg \min_{\gamma \in \mathbb{R}} F(\gamma).$$

From (13) and (15), we know $\partial_{xx} \phi$ is the dual variable of γ . Since $\overline{\gamma}$ minimizes F , the corresponding $\partial_{xx} \phi$ must be zero. Therefore, we have the following boundary conditions:

$$\partial_{xx} \phi(t, \underline{D}) = \partial_{xx} \phi(t, \overline{D}) = 0, \quad \forall t \in [0, 1]. \quad (39)$$

In [1], periodic boundary conditions were used in the spatial dimension and a perturbed equation was used to yield a unique solution. Since periodic boundary conditions are inappropriate for martingale diffusion and we are dealing with a bi-Laplacian term in space, we impose the following additional boundary conditions in order to enforce a unique solution:

$$\phi(t, \underline{D}) = \phi(t, \overline{D}) = 0, \quad \forall t \in [0, 1]. \quad (40)$$

Now, the 4th order linear PDE (36) can be numerically solved by the finite difference method or the finite element method.

Step B: $q^n = \arg \min_q L_r(\phi^n, q, \mu^{n-1})$

Since $H(q)$ is not differentiable, we cannot differentiate L_r with respect to q . Nevertheless, we can simply obtain q^n by solving the minimization problem

$$\inf_q L_r(\phi^n, q, \mu^{n-1}). \quad (41)$$

This is equivalent to solving

$$\inf_{q \in K} \left\langle \nabla_{t,xx} \phi^n + \frac{\mu^{n-1}}{r} - q, \nabla_{t,xx} \phi^n + \frac{\mu^{n-1}}{r} - q \right\rangle. \quad (42)$$

Now, let us define

$$p^n(t, x) = \{\alpha^n(t, x), \beta^n(t, x)\} = \nabla_{t,xx} \phi^n(t, x) + \frac{\mu^{n-1}(t, x)}{r}, \quad (43)$$

then we can find $q^n(t, x) = \{a^n(t, x), b^n(t, x)\}$ by solving

$$\inf_{\{a,b\} \in \mathbb{R} \times \mathbb{R}} \left\{ (a(t, x) - \alpha^n(t, x))^2 + (b(t, x) - \beta^n(t, x))^2 : a + F^*(b) \leq 0 \right\} \quad (44)$$

point-wise in space and time. This is a simple one-dimensional projection problem. If $\{\alpha^n, \beta^n\}$ satisfies the constraint $\alpha^n + F^*(\beta^n) \leq 0$, then it is also the minimum. Otherwise, the minimum must occur on the boundary $a + F^*(b) = 0$. In this case we substitute the condition into (44) to obtain

$$\inf_{b \in \mathbb{R}} \left\{ (F^*(b(t, x)) + \alpha(t, x))^2 + (b(t, x) - \beta(t, x))^2 \right\}, \quad (45)$$

which must be solved point-wise. The minimum of (45) can be found using standard root finding methods such as Newton's method. In some simple cases it is even possible to compute the solution analytically.

Step C: $\mu^n = \arg \max_\mu L_r(\phi^n, q^n, \mu)$

Begin by computing the gradient by differentiating the augmented Lagrangian L_r respect to μ . Then, simply update μ by moving it point-wise along the gradient as follows,

$$\mu^n(t, x) = \mu^{n-1}(t, x) + r(\nabla_{t,xx} \phi^n(t, x) - q^n(t, x)). \quad (46)$$

Stopping criteria:

Recall the HJB equation (15):

$$\partial_t \phi + F^*(\partial_{xx} \phi) = 0. \quad (47)$$

We use (47) to check for optimality. Define the residual:

$$res^n = \max_{t \in [0, 1], x \in D} \rho \left| \partial_t \phi + F^*(\partial_{xx} \phi) \right|. \quad (48)$$

This quantity converges to 0 when it approaches the optimal solution of the problem. The residual is weighted by the density ρ to alleviate any potential issues caused by small values of ρ .

5 Numerical Results

The algorithm was implemented and tested on the following simple example. Consider the computational domain $x \in [0, 1]$ and the time interval $t \in [0, 1]$. We set the initial and final distributions to be $X_0 \sim N(0.5, 0.05^2)$ and $X_1 \sim N(0.5, 0.1^2)$ respectively, where $N(\mu, \sigma^2)$ denotes the normal distribution. The following cost function was chosen:

$$F(\gamma) = \begin{cases} (\gamma - \bar{\gamma})^2, & \gamma \geq 0, \\ +\infty, & \text{otherwise,} \end{cases} \quad (49)$$

where $\bar{\gamma}$ was set to 0.00375 so that the optimal value of variance is constant $\sigma^2 = 0.1^2 - 0.05^2 = 0.0075$. Then we discretized the space-time domain as a 128×128 lattice. The penalization parameter is set to $r = 64$. The results after 3000 iterations are shown in Figs. 1 and 2, and the convergence of the residuals is shown in Fig. 3. The convergence speed decays quickly, but we reach a good approximation after about 500 iterations. The noisy tails in Fig. 2 correspond to regions where the density ρ is close to zero. The diffusion process has a very low probability of reaching these regions, so the value of σ^2 has little impact. In areas where ρ is not close to zero, σ^2 remains constant which matches the analytical solution.

6 Summary

This paper focuses on a new approach for the calibration of local volatility models. Given the distributions of the asset price at two fixed dates, the technique of optimal transport is applied to interpolate the distributions and recover the local volatility function, while maintaining the martingale property of the underlying

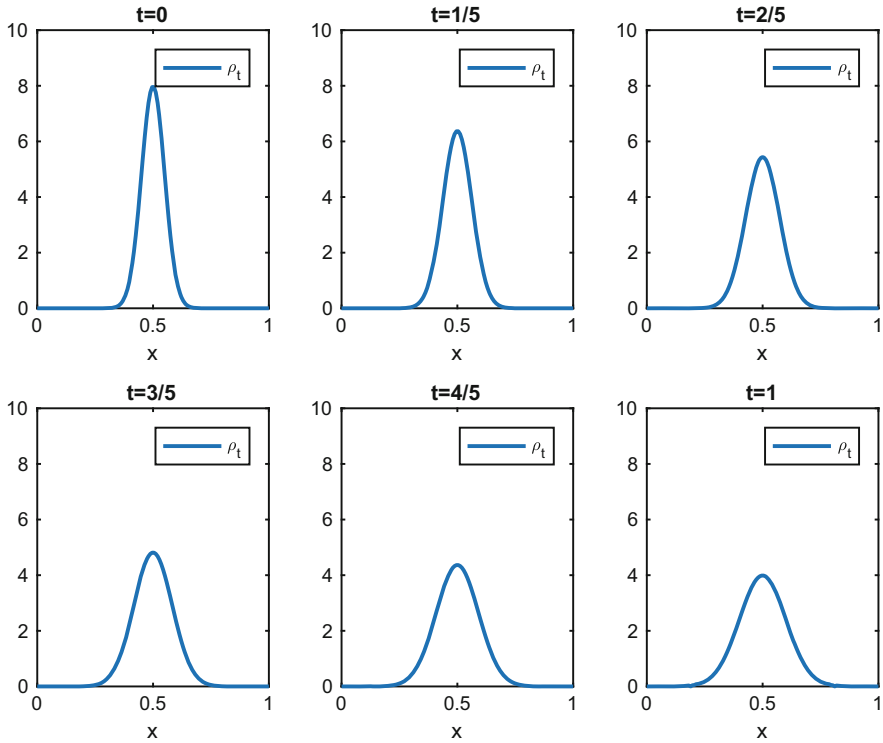


Fig. 1 The density function $\rho(t, x)$

process. Inspired by Benamou and Brenier [1], the problem is first converted into a saddle point problem, and then solved numerically by an augmented Lagrangian approach and the alternative direction method of multipliers (ADMM) algorithm. The algorithm performs well on a simple case in which the numerical solution matches its analytical counterpart. The main drawback of this method is due to the slow convergence rate of the ADMM algorithm. We observed that a higher penalization parameter may lead to faster convergence. Further research is required to conduct more numerical experiment, improve the efficiency of the algorithm and apply it to more complex cases.

Acknowledgements The Centre for Quantitative Finance and Investment Strategies has been supported by BNP Paribas.

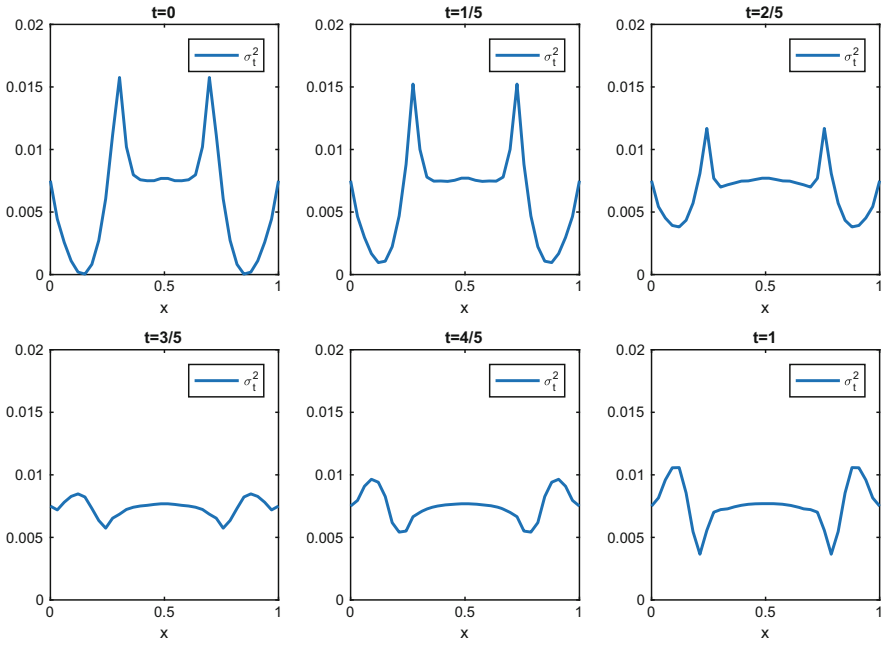


Fig. 2 The variance $\sigma^2(t, x)$

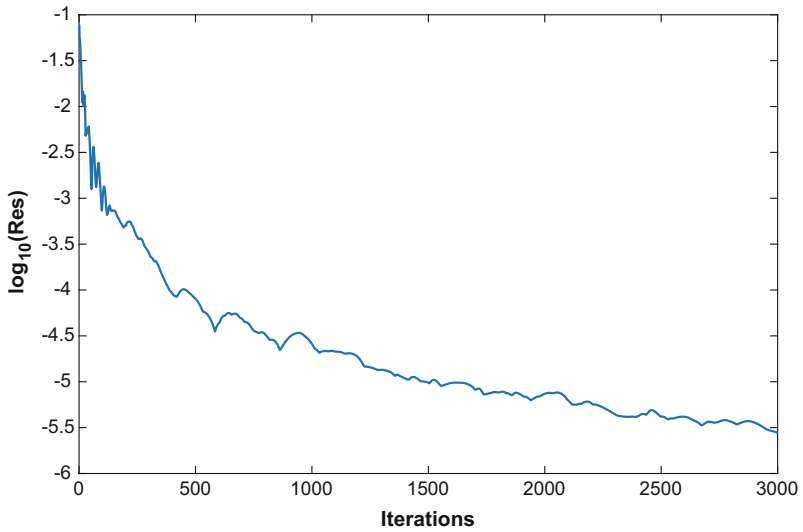


Fig. 3 The residual res^n

References

1. Benamou, J.D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.* **84**(3), 375–393 (2000)
2. Bouchard, B., Loeper, G., Zou, Y.: Hedging of covered options with linear market impact and gamma constraint (2015). Preprint. arXiv:1512.07087
3. Bouchard, B., Loeper, G., Zou, Y.: Almost-sure hedging with permanent price impact. *Finance Stochast.* **20**(3), 741–771 (2016)
4. Breeden, D.T., Litzenberger, R.H.: Prices of state-contingent claims implicit in option prices. *J. Bus.* **51**, 621–651 (1978)
5. Dolinsky, Y., Soner, H.M.: Martingale optimal transport and robust hedging in continuous time. *Probab. Theory Relat. Fields* **160**(1–2), 391–427 (2014)
6. Dupire, B.: Pricing with a smile. *Risk Mag.* **7**, 18–20 (1994)
7. Evans, L.C.: Partial differential equations and Monge-Kantorovich mass transfer. *Curr. Dev. Math.* **1997**(1), 65–126 (1997)
8. Kantorovich, L.V.: On a problem of Monge. *J. Math. Sci.* **133**(4), 1383–1383 (2006)
9. Loeper, G.: Option pricing with linear market impact and non-linear Black-Scholes equations (2013). Preprint. arXiv:1301.6252
10. Monge, G.: Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris* (1781)
11. Rockafellar, R.T.: *Convex Analysis*. Princeton University Press, Princeton (2015)
12. Strassen, V.: The existence of probability measures with given marginals. *Ann. Math. Stat.* **36**, 423–439 (1965)
13. Tan, X., Touzi, N., et al.: Optimal transportation under controlled stochastic dynamics. *Ann. Probab.* **41**(5), 3201–3240 (2013)
14. Villani, C.: *Optimal Transport: Old and New*, vol. 338. Springer Science & Business Media, Berlin (2008)

Likelihood Informed Dimension Reduction for Remote Sensing of Atmospheric Constituent Profiles



Otto Lamminpää, Marko Laine, Simo Tukiainen, and Johanna Tamminen

Abstract We use likelihood informed dimension reduction (LIS) (Cui et al. *Inverse Prob* 30(11):114015, 28, 2014) for inverting vertical profile information of atmospheric methane from ground based Fourier transform infrared (FTIR) measurements at Sodankylä, Northern Finland. The measurements belong to the word wide TCCON network for greenhouse gas measurements and, in addition to providing accurate greenhouse gas measurements, they are important for validating satellite observations.

LIS allows construction of an efficient Markov chain Monte Carlo sampling algorithm that explores only a reduced dimensional space but still produces a good approximation of the original full dimensional Bayesian posterior distribution. This in effect makes the statistical estimation problem independent of the discretization of the inverse problem. In addition, we compare LIS to a dimension reduction method based on prior covariance matrix truncation used earlier (Tukiainen et al., *J Geophys Res Atmos* 121:10312–10327, 2016).

1 Introduction

Atmospheric composition measurements have an increasingly crucial role in monitoring the green house gas concentrations in order to understand and predict changes in climate. The warming effect of greenhouse gases, such as carbon dioxide (CO₂) and methane (CH₄), is based on the absorption of electromagnetic radiation originating from the sun by these trace gases. This mechanism has a strong theoretical base and has been confirmed by recent observations [4].

Remote sensing measurements of atmospheric composition, and greenhouse gases in particular, are carried out by ground-based Fourier transform infrared (FTIR) spectrometers, and more recently by a growing number of satellites (for

O. Lamminpää (✉) · M. Laine · S. Tukiainen · J. Tamminen
Finnish Meteorological Institute, Helsinki, Finland
e-mail: otto.lamminpaa@fmi.fi

example SCIAMACHY, ACE-FTS, GOSAT, OCO-2). The advantage of satellite measurements is that they provide global coverage. They are used for anthropogenic emission monitoring, detecting trends in atmospheric composition and studying the effects of biosphere, to name but a few examples. Accurate ground-based measurements are crucial to satellite measurement validation, and the global Total Carbon Column Observing Network (TCCON [17]) of FTIR spectrometers, consisting of around 20 measurement sites around the world, is widely used as a reference [3]. The FTIR instrument looks directly at sun, returning an absorption spectrum as measured data.

Determining atmospheric gas density profiles, or *retrieval*, from the absorption spectra is an ill-defined *inverse problem* as the measurement contains only a limited amount of information about the state of the atmosphere. Based on prior knowledge and using the Bayesian approach to regularize the problem, the profile retrieval is possible, provided that our prior accurately describes the possible states that may occur in the atmosphere. When retrieving a vertical atmospheric profile, the dimension of the estimation problem depends on the discretization. For accurate retrievals a high number of layers are needed, leading to a computationally costly algorithms. However, fast methods are required for the operational algorithm. For this purpose, different ways of reducing the dimension of the problem have been developed. The official operational TCCON GGG algorithm [17] solves the inverse problem by scaling the prior profile based on the measured data. This method is robust and computationally efficient, but only retrieves one piece of information and thus can give largely inaccurate results about the density profiles.

An improved dimension reduction method for the FTIR retrieval based on reducing the rank of the prior covariance matrix was used by Tukiainen et al. [16] using computational methods developed by Solonen et al. [13]. This method confines the solution to a subspace spanned by the non-negligible eigenvectors of the prior covariance matrix. This approach allows a retrieval using more basis functions than the operational method and thus gives more accurate solutions. However, the prior has to be hand tuned to have a number of non-zero singular values that correspond to the number of degrees of freedom for the signal in the measurement. Moreover, whatever information lies in the complement of this subspace remains unused.

In this work, we introduce an analysis method for determining the number of components the measurement can provide information from [12], as well as the *likelihood informed subspace* dimension reduction method for non-linear statistical inverse problems [2, 14]. We show that these two formulations are in fact equal. We then proceed to implement a dimension reduction scheme for the FTIR inverse problem using adaptive MCMC sampling [5, 6] to fully characterize the non-linear posterior distribution, and show that this method gives an optimal result with respect to Hellinger distance to the non-approximated full dimensional posterior distribution. In contrast with the previously implemented prior reduction method, the likelihood informed subspace method is also shown to give the user freedom to use a prior derived directly from an ensemble of previously conducted atmospheric composition measurements.

2 Methodology

We consider the atmospheric composition measurement carried out at the FMI Arctic Research Centre, Sodankylä, Finland [9]. The on-site Fourier transform infrared spectrometer (FTIR) measures solar light arriving to the device directly from the sun, or more precisely, the absorption of solar light at different wavelengths within the atmosphere. From the absorption spectra of different trace gases (CO_2 , CH_4 , . . .) we can compute the corresponding vertical density profiles, i.e. the fraction of the trace gas in question as a function of height.

Let us consider the absorption spectrum with m separate wavelengths. The solar light passing through the atmosphere and hitting the detector can be modeled using the *Beer-Lambert law*, which gives, for wavelengths λ_j , $j \in [1, \dots, m]$, the intensity of detected light as

$$I(\lambda_j) = I_0(\lambda_j) \exp\left(-\sum_{k=1}^K \int_0^\infty \mathcal{C}_k(\lambda_j, z) \rho_k(z) dz\right) (a\lambda_j^2 + b\lambda_j + c) + d, \quad (1)$$

where I_0 is the intensity of solar light when it enters the atmosphere, the atmosphere has K absorbing trace gases, $\mathcal{C}_k(\lambda_j, z)$ is the absorption coefficient of gas k , which depends on height z and on the wavelength λ_j , and $\rho_k(z)$ is the density of gas k at height z . The second degree polynomial and the constant d in (1) are used to describe instrument related features and the continuity properties of the spectrum. In reality, solar light is scattered on the way by atmospheric particles. This phenomenon is relatively weak in the wavelength band we are considering in this work, so it will be ignored for simplicity.

The absorption in continuous atmosphere is modeled by discretizing the integral in Eq. (1) into a sum over atmospheric layers and assuming a constant absorption for each separate layer. This way, a discrete computational *forward model* can be constructed, giving an absorption spectrum as data produced by applying the forward model to a state vector x describing the discretized atmospheric density profile for a certain trace gas. In this work, we limit ourselves to consider the retrieval of atmospheric methane (CH_4).

2.1 Bayesian Formulation of the Inverse Problem

Consider an inverse problem of estimating unknown parameter vector $x \in \mathbb{R}^n$ from observation $y \in \mathbb{R}^m$,

$$y = F(x) + \varepsilon, \quad (2)$$

where our physical model is describe by the *forward model* $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and the random variable $\varepsilon \in \mathbb{R}^m$ represents the observation error arising from instrument

noise and forward model approximations. In the Bayesian approach to inverse problems [7] our uncertainty about x is described by statistical distributions. The solution to the problem is obtained as posterior distribution of x conditioned on a realization of the data y and depending on our prior knowledge. By the Bayes' formula, we have

$$\pi(x|y) \propto \pi(y|x)\pi_{pr}(x), \quad (3)$$

where $\pi(x|y)$ is the posterior distribution, $\pi(y|x)$ the likelihood and $\pi_{pr}(x)$ the prior distribution. The proportionality \propto comes from a constant that does not depend on the unknown x . In this work, we assume the prior to be Gaussian, $\mathcal{N}(x_0, \Sigma_{pr})$, e.g.

$$\pi_{pr}(x) \propto \exp\left(-\frac{1}{2}(x - x_0)^T \Sigma_{pr}^{-1}(x - x_0)\right). \quad (4)$$

Also, the additive noise is assumed to be zero-mean Gaussian with known covariance matrix, $\varepsilon \sim \mathcal{N}(0, \Sigma_{obs})$, so the likelihood will have form

$$\pi(y|x) \propto \exp\left(-\frac{1}{2}(y - F(x))^T \Sigma_{obs}^{-1}(y - F(x))\right). \quad (5)$$

When the forward model is non-linear, the posterior distribution can be explored by Markov chain Monte Carlo (MCMC) sampling. When the dimension of the unknown is high, for example by discretization of the inverse problem, MCMC is known to be inefficient. In this paper, we utilize dimension reduction to be able to make MCMC more efficient in high dimensional and high CPU problems.

2.2 Prior Reduction

The operational GGG algorithm for the FTIR retrieval problem [17] is effectively one dimensional as it only scales the prior mean profile. However, there are about three degrees of freedom in the FTIR signal for the vertical profile information. To construct basis functions that could utilize this information a method that uses prior reduction was developed in [16]. It is based on the singular value decomposition on the prior covariance matrix,

$$\Sigma_{pr} = U \Lambda U^T = \sum_{i=1}^m \lambda_i u_i u_i^T, \quad (6)$$

which allows further decomposition as

$$\Sigma_{pr} = P P^T, \text{ with } P = \left(\sqrt{\lambda_1} u_1 + \dots + \sqrt{\lambda_m} u_m\right). \quad (7)$$

If the prior can be chosen so that most of the singular values are negligible, then the rank of the prior covariance matrix can be reduced by considering only the first r singular values and vectors:

$$\tilde{\Sigma}_{pr} = P_r P_r^T, \text{ with } P_r = \left(\sqrt{\lambda_1} u_1 + \cdots + \sqrt{\lambda_r} u_r \right). \quad (8)$$

The unknown x has an approximate representation by r basis vectors from the columns of P_r and using a reduced dimensional parameter $\alpha \in \mathbb{R}^r$ as

$$x \approx x_0 + P_r \alpha. \quad (9)$$

By the construction, the random vector α has a simple Gaussian prior, $\alpha \sim \mathcal{N}(0, \mathbb{I})$, which allow us to write the approximate posterior as

$$\pi(x|y) \approx \tilde{\pi}(\alpha|y) \propto \exp \left(-\frac{1}{2} \left((y - F(x_0 + P_r \alpha))^T \Sigma_{obs}^{-1} (y - F(x_0 + P_r \alpha)) + \alpha^T \alpha \right) \right). \quad (10)$$

Now, instead running MCMC in the full space defined by x , we can sample the low dimensional parameter α and retain the approximation of the full posterior by Eq. (9).

2.3 Likelihood-Informed Subspace

The prior reduction approach depends on the ability to construct a realistic prior that can be described by only a few principle components. For the FTIR retrieval problem this is possible to some extent [16]. However, there are several possible caveats. We have to manually manipulate the prior covariance matrix to have a lower rank, which can lead to information loss as the solution will be limited to a subspace defined by the reduced prior only.

In atmospheric remote sensing the information content of the measurement is an important concept to be considered when designing the instruments and constructing the retrieval methodology, we refer to book by Rodgers [12].

Consider a linearized version of the inverse problem in Eq. (2),

$$y = J(x - x_0) + \varepsilon, \quad (11)$$

with Gaussian prior and noise. The forward model is assumed to be differentiable, and J denotes the Jacobian matrix of the forward model with elements $J_{ij} = \frac{\partial}{\partial x_j} F_i$. Using Cholesky factorizations for the known prior and error covariances,

$$\Sigma_{pr} = \mathcal{L}_{pr} \mathcal{L}_{pr}^T, \quad \Sigma_{obs} = \mathcal{L}_{obs} \mathcal{L}_{obs}^T, \quad (12)$$

we can perform pre-whitening of the problem by setting

$$\tilde{y} = \mathcal{L}_{obs}^{-1}y, \quad \tilde{J} = \mathcal{L}_{obs}^{-1}J\mathcal{L}_{pr}, \quad \tilde{x} = \mathcal{L}_{pr}^{-1}(x - x_0) \text{ and } \tilde{\varepsilon} = \mathcal{L}_{obs}^{-1}\varepsilon. \quad (13)$$

Now the problem can be written as

$$\tilde{y} = \tilde{J}\tilde{x} + \tilde{\varepsilon}, \quad (14)$$

with $\tilde{\varepsilon} \sim \mathcal{N}(0, \mathbb{I})$ and a priori $\tilde{x} \sim \mathcal{N}(0, \mathbb{I})$.

As the unknown x and the error ε are assumed to be independent, the same holds for the scaled versions. We can compare the prior variability of the observation depending on x and that coming from the noise ε by

$$\tilde{\Sigma}_y = \mathbb{E}[\tilde{y}\tilde{y}^T] = \mathbb{E}[(\tilde{J}\tilde{x} + \tilde{\varepsilon})(\tilde{J}\tilde{x} + \tilde{\varepsilon})^T] = \tilde{J}\tilde{J}^T + \mathbb{I}. \quad (15)$$

The variability in y that depends only on the parameter x depends itself on $\tilde{J}\tilde{J}^T$ and it can be compared to the unit matrix \mathbb{I} that has the contribution from the scaled noise. The directions in $\tilde{J}\tilde{J}^T$ which are larger than unity are those dominated by the signal. Formally this can be seen by diagonalizing the scaled problem by the singular value decomposition,

$$\tilde{J} = W\Lambda V^T, \quad (16)$$

and setting

$$y' = W^T\tilde{y} = W^T\tilde{J}\tilde{x} + W^T\tilde{\varepsilon} = \Lambda V^T\tilde{x} + \tilde{\varepsilon}' = \Lambda\tilde{x}' + \tilde{\varepsilon}'. \quad (17)$$

The transformations ε' and x' conserve the unit covariance matrix. In other words, y' is distributed with covariance $\Lambda^2 + \mathbb{I}$. This is a diagonal matrix, and the elements of vector y' that are not masked by the measurement error are those corresponding to the singular values $\lambda_i \geq 1$ of the pre-whitened Jacobian \tilde{J} . Furthermore, degrees of freedom for signal and noise are invariant under linear transformations [12], so the same result is also valid for the original y .

Another way to compare the information content of the measurement relative to the prior was used in [2]. This is to use the Rayleigh quotient

$$\mathcal{R}(\mathcal{L}_{pr}a) = \frac{a^T \mathcal{L}_{pr}^T H \mathcal{L}_{pr} a}{a^T a}, \quad (18)$$

where $a \in \mathbb{R}^n$ and $H = J^T \Sigma_{obs}^{-1} J$ is the Gauss-Newton approximation of Hessian matrix of the data misfit function

$$\eta(x) = \frac{1}{2} \left((y - F(x))^T \Sigma_{obs}^{-1} (y - F(x)) \right). \quad (19)$$

Directions for which $\mathcal{R}(\mathcal{L}_{pr}a) > 1$ are the ones in which the likelihood contains information relative to the prior. This follows from the fact that the i th eigenvector v_i of the prior-preconditioned Gauss-Newton Hessian

$$\tilde{H} := \mathcal{L}_{pr}^T H \mathcal{L}_{pr} \quad (20)$$

maximizes the Rayleigh quotient over a subspace $\mathbb{R}^n \setminus \text{span}\{v_1, \dots, v_{i-1}\}$ and the r directions v_i for which $\mathcal{R}(\mathcal{L}_{pr}v) > 1$ correspond to the first r eigenvalues of \tilde{H} . We call these vectors the *informative directions of the measurement*.

To see the correspondence for the two approaches for the informative directions we notice that for $\tilde{H}(x)$ it holds that

$$\begin{aligned} \mathcal{L}_{pr}^T H(x) \mathcal{L}_{pr} &= \mathcal{L}_{pr}^T J(x)^T \Sigma_{obs}^{-1} J(x) \mathcal{L}_{pr}^T \\ &= (\mathcal{L}_{obs}^{-1} J(x) \mathcal{L}_{pr})^T (\mathcal{L}_{obs}^{-1} J(x) \mathcal{L}_{pr}) \\ &= \tilde{J}^T(x) \tilde{J}(x). \end{aligned} \quad (21)$$

The eigenvalues λ^2 of matrix $\tilde{H}(x)$ less than unity correspond to the singular values λ less than unity of the scaled Jacobian $\tilde{J}(x)$. The corresponding eigenvectors are the same as the right singular vectors v of \tilde{J} . The informative and non-informative directions for a simple 2-dimensional Gaussian case are illustrated in Fig. 1.

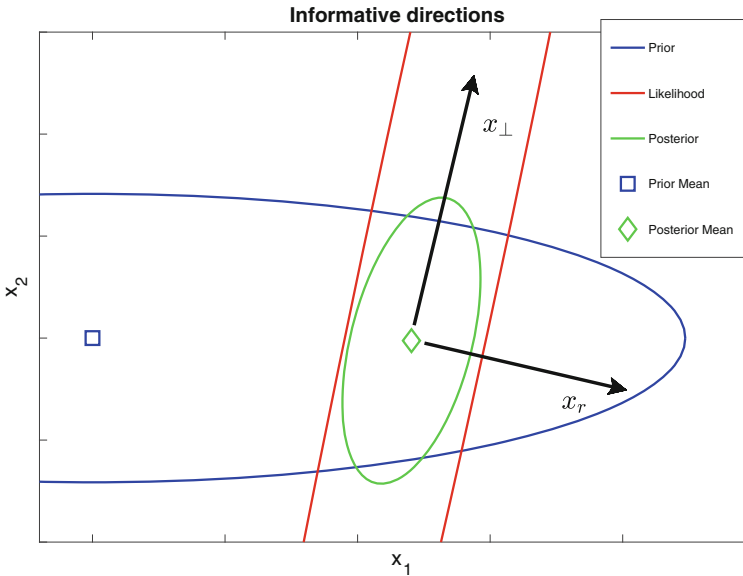


Fig. 1 Illustration of an informative direction x_r and a non-informative direction x_{\perp} using a 2-dimensional Gaussian case. Here, the likelihood has only one informative component, so the remaining direction for the posterior is obtained from the prior

Next, we use the informative directions of the measurement to reduce the dimension of the inverse problem. Consider approximations for the posterior of the form

$$\tilde{\pi}(x|y) \propto \pi(y|\Pi_r x) \pi_{pr}(x), \quad (22)$$

where Π_r is rank r projection matrix. In [2] and [14] it was shown that for any given r , there exists a unique optimal projection Π_r that minimizes the Hellinger distance between the approximative rank r posterior and the full posterior. Furthermore, using the connection to Rodgers' formalism, the optimal projection can be obtained explicitly with the following definition.

Definition 1 (LIS) Let $V_r \in \mathbb{R}^{n \times r}$ be a matrix containing the first r left singular vectors of the scaled Jacobian \tilde{J} . Define

$$\Phi_r := \mathcal{L}_{pr} V_r \text{ and } \Theta_r := \mathcal{L}_{pr}^{-T} V_r. \quad (23)$$

The rank r LIS projection for the posterior approximation (22) is given by

$$\Pi_r = \Phi_r \Theta_r^T. \quad (24)$$

The range \mathbb{X}_r of projection $\Pi_r : \mathbb{R}^n \rightarrow \mathbb{X}_r$ is a subspace of state space \mathbb{R}^n spanned by the column vectors of matrix Φ_r . We call the subspace \mathbb{X}_r the *likelihood-informed subspace (LIS)* for the linear inverse problem, and its complement $\mathbb{R}^n \setminus \mathbb{X}_r$ the *complement subspace (CS)*.

Definition 2 The matrix of singular vectors $V = [V_r V_\perp]$ forms a complete orthonormal system in \mathbb{R}^n and we can define

$$\Phi_\perp := \mathcal{L}_{pr} V_\perp \text{ and } \Theta_\perp := \mathcal{L}_{pr}^{-T} V_\perp \quad (25)$$

and the projection $\mathbb{I} - \Pi_r$ can be written as

$$\mathbb{I} - \Pi_r = \Phi_\perp \Theta_\perp^T. \quad (26)$$

Define the LIS-parameter $x_r \in \mathbb{R}^r$ and the CS-parameter $x_\perp \in \mathbb{R}^{n-r}$ as

$$x_r := \Theta_r^T x, \quad x_\perp := \Theta_\perp^T x. \quad (27)$$

The parameter x can now be naturally decomposed as

$$\begin{aligned} x &= \Pi_r x + (\mathbb{I} - \Pi_r) x \\ &= \Phi_r x_r + \Phi_\perp x_\perp. \end{aligned} \quad (28)$$

Using this decomposition and properties of multivariate Gaussian distributions, we can write the prior as

$$\pi_{pr}(x) = \pi_r(x_r)\pi_{\perp}(x_{\perp}) \quad (29)$$

and approximate the likelihood by using the r informative directions,

$$\pi(y|x) = \pi(y|\Phi_r x_r)\pi(y|\Phi_{\perp} x_{\perp}) \approx \pi(y|\Phi_r x_r), \quad (30)$$

which leads us to the approximate posterior

$$\tilde{\pi}(x|y) = \pi(y|\Phi_r x_r)\pi_r(x_r)\pi_{\perp}(x_{\perp}). \quad (31)$$

When the forward model is not linear, the Jacobian and Hessian matrices depend on the parameter x and the criterion (18) only holds point wise. To extend this local condition into a global one, we consider the expectation of the local Rayleigh quotient $\mathcal{R}(\mathcal{L}_{pr}v; x)$ over the posterior,

$$\mathbb{E}[\mathcal{R}(\mathcal{L}_{pr}v; x)] = \frac{v^T \hat{J}^T \hat{J} v}{v^T v}, \quad \hat{J} = \int_{\mathbb{R}^n} \tilde{J}(x)\pi(x|y)dx. \quad (32)$$

The expectation is with respect to the posterior distribution, which is not available before the analysis. In practice, an estimate is obtained by Monte Carlo,

$$\hat{J}_n = \frac{1}{n} \sum_{k=1}^n \tilde{J}(x^{(k)}), \quad (33)$$

where $x^{(k)}$ is a set of samples from some reference distribution which will be discussed later in this work. We can now use the singular value decomposition $\hat{J}_n = W\Lambda V^T$ to find a basis for the global LIS analogously to the linear case.

The advantage of LIS dimension reduction is that it is sufficient to use MCMC to sample the low-dimensional x_r from the reduced posterior $\pi(y|\Phi_r x_r)\pi_r(x_r)$, and form the full space approximation using the known analytic properties of the Gaussian complement prior $\pi_{\perp}(x_{\perp})$.

3 Results

To solve the inverse problem related to the FTIR measurement [16], we use adaptive MCMC [6, 10] and SWIRLAB [15] toolboxes for Matlab. The results from newly implemented LIS-algorithm as well as from the previous prior reduction method are compared against a full dimensional MCMC simulation using the Hellinger distance of approximations to the full posterior. We use a prior derived from an ensemble of

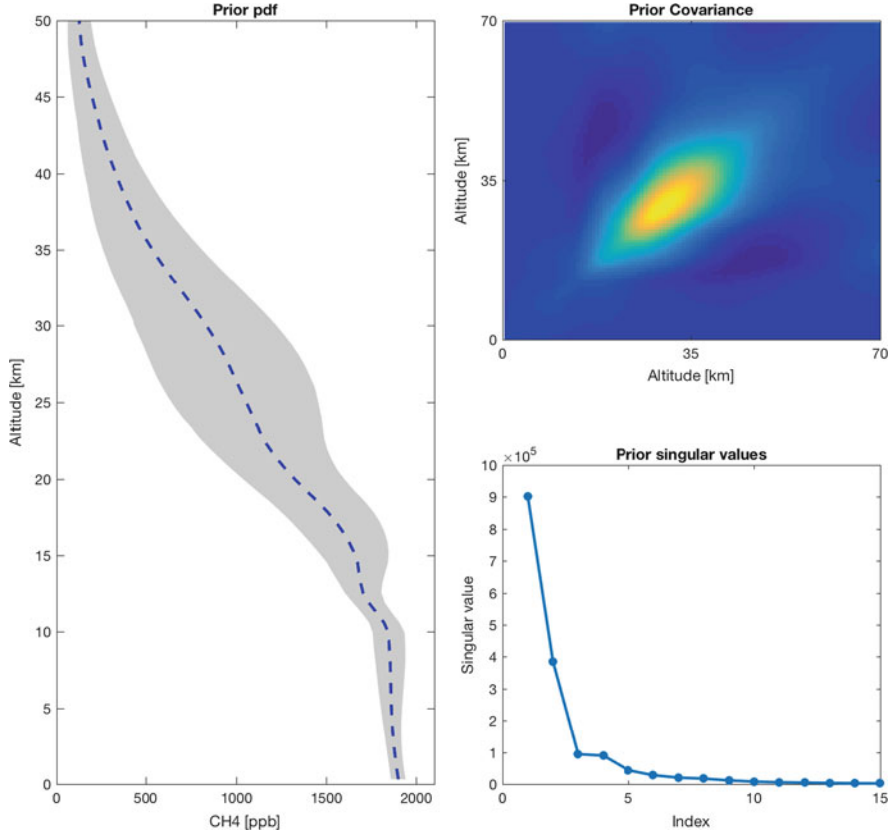


Fig. 2 The prior derived from an ensemble of ACE satellite measurements. Left: Full prior profile, mean with dashed line and 95% probability limits in grey. Top right: covariance matrix derived from the measurements. Bottom right: first 20 singular values of the prior covariance matrix

atmospheric composition measurements by the ACE satellite [1]. The vertical prior distribution, prior covariance and prior singular values are illustrated in Fig. 2.

In Fig. 3, we show the results of our retrievals using full-space MCMC, compared with LIS dimension reduction and prior reduction using four basis vectors in each method. The retrievals are further compared against accurate in-situ measurements made using AirCore balloon soundings [8] which are available for the selected cases, also included in Fig. 3. In this example, the Monte-Carlo estimator (33) for \hat{J}_n in Eq. (33) was computed using 1000 samples drawn from the Laplace approximation $\mathcal{N}(\hat{x}, \hat{\Sigma}_{post})$, where \hat{x} and $\hat{\Sigma}_{post}$ are the posterior MAP and covariance, respectively, obtained using optimal estimation [12].

In order to compare the performance of MCMC methods, we define the *sample speed* of a MCMC run as

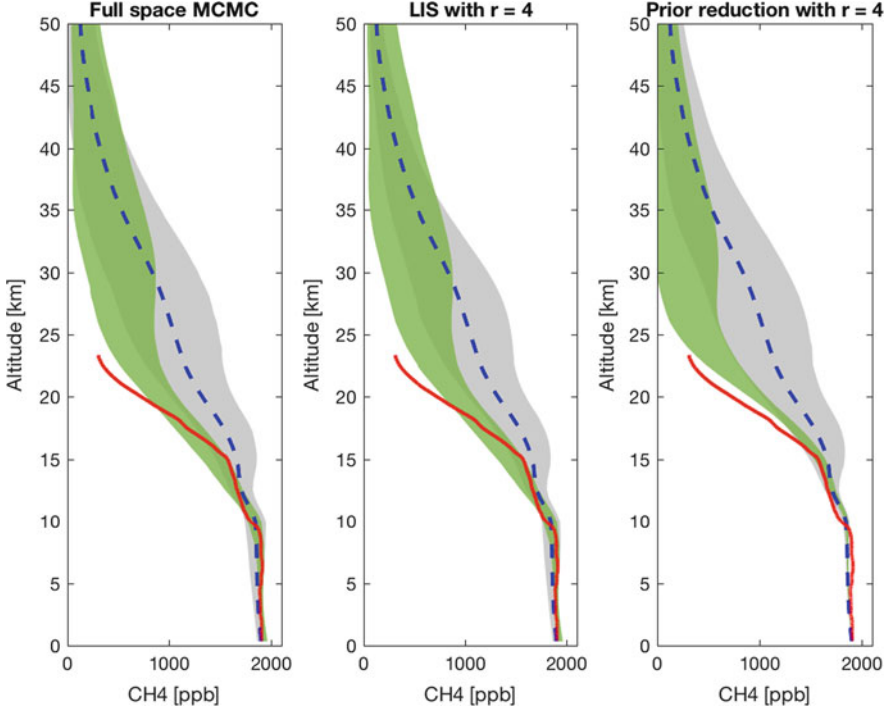


Fig. 3 Atmospheric CH₄ density profile retrieval results. Retrieved posterior in green, prior in gray, and in-situ AirCore measurement in red. The color shading indicates areas where 95% of the profiles are. Right: MCMC with in full space. Middle: MCMC with LIS. Right: MCMC with prior reduction

Definition 3 The *effective sample size* N_{eff} of a MCMC chain is given by

$$N_{\text{eff}} = \frac{N_M}{1 + s \sum_{k=1}^{\infty} \rho_k(x)}, \quad (34)$$

where N_M is the length of the MCMC chain and $\rho_k(x)$ is lag- k autocorrelation for parameter x [11]. Define the *sample speed* of an MCMC chain as

$$\mathbb{V} = \frac{N_{\text{eff}}}{t_M}, \quad (35)$$

where t_M is the total computation time of the MCMC chain.

For the MCMC runs shown in Fig. 3, we get as corresponding sample speeds as samples per second:

$$\mathbb{V}(\text{full}) = 1.56 \text{ s}^{-1}, \quad \mathbb{V}(\text{LIS}) = 19.01 \text{ s}^{-1}, \quad \mathbb{V}(\text{PriRed}) = 19.66 \text{ s}^{-1}. \quad (36)$$

In order to compare the approximate posteriors obtained from prior reduction and LIS-dimension reduction against the full posterior, we use the discrete Hellinger distance,

$$\mathcal{H}(P, Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^k (\sqrt{p_i} - \sqrt{q_i})^2}, \quad (37)$$

where $P = (p_1, \dots, p_k)$ and $Q = (q_1, \dots, q_k)$ are discrete representations of the full and approximate posterior distributions obtained from histograms of corresponding MCMC runs. The Hellinger distances of both approximations to the full posterior can be seen in Fig. 4 together with the corresponding sample speeds, both as a function of the number of singular vectors used. In Fig. 4 we have also visualized the first four singular vectors used in prior reduction and LIS method for the example retrieval in Fig. 3.

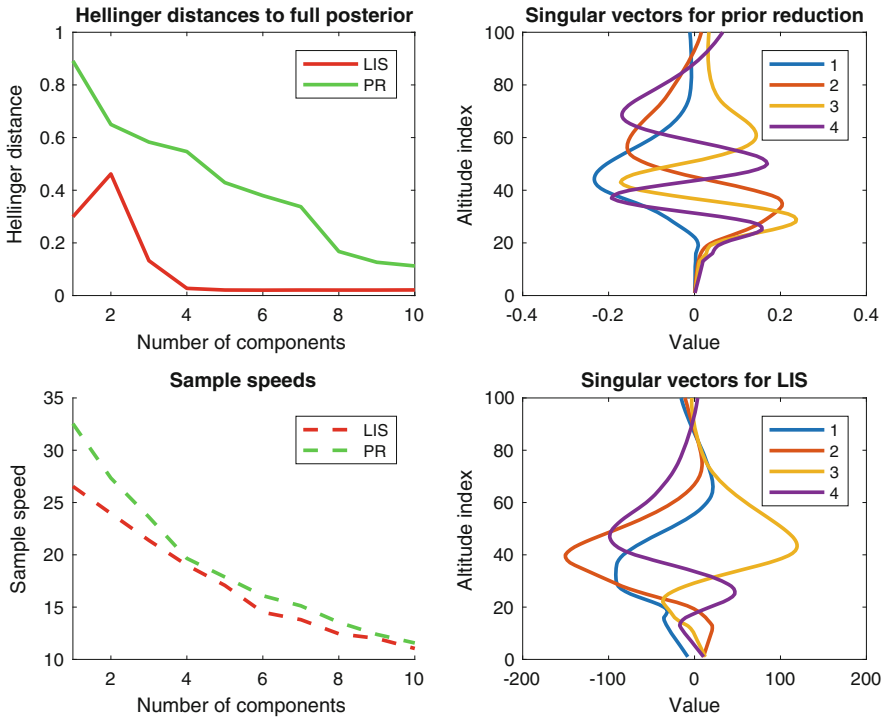


Fig. 4 Left: Hellinger distances to full posterior and sample speeds of corresponding MCMC runs as functions of singular vectors used in the approximation. Top right: first four singular vectors from prior reduction. Bottom right: first four singular vectors of \tilde{J} forming the LIS basis

4 Conclusions

Although both of the discussed dimension reduction methods provide roughly the same computational gains in the performance of the MCMC sampler, we see from Fig. 4 that while using an empirical prior, the prior reduction method requires a lot more singular vectors to achieve the same Hellinger distance from the full posterior as the LIS method, which gets really close already with four singular vectors. We conclude that the LIS method gives an efficient MCMC sampling algorithm to solve the inverse problem arising from the FTIR retrieval, with an additional improvement of allowing the direct usage of an empirical prior.

Acknowledgements We thank Dr. Rigel Kivi from FMI Arctic Research Centre, Sodankylä, Finland for the AirCore and TCCON data. We thank Dr. Tiangang Cui from Monash University and the mathematical research institute MATRIX in Australia for organizing a workshop where a part of this research was performed. This work has been supported by Academy of Finland (projects INQUIRE, IIDA-MARI and CoE in Inverse Modelling and Imaging) and by EU's Horizon 2020 research and innovation programme (project GAIA-CLIM).

References

1. Bernath, P.F., McElroy, C.T., Abrams, M.C., Boone, C.D., Butler, M., Camy-Peyret, C., Carleer, M., Clerbaux, C., Coheur, P.F., Colin, R., DeCola, P., DeMazière, M., Drummond, J.R., Dufour, D., Evans, W.F.J., Fast, H., Fussen, D., Gilbert, K., Jennings, D.E., Llewellyn, E.J., Lowe, R.P., Mahieu, E., McConnell, J.C., McHugh, M., McLeod, S.D., Michaud, R., Midwinter, C., Nassar, R., Nichitiu, F., Nowlan, C., Rinsland, C.P., Rochon, Y.J., Rowlands, N., Semeniuk, K., Simon, P., Skelton, R., Sloan, J.J., Soucy, M.A., Strong, K., Tremblay, P., Turnbull, D., Walker, K.A., Walkty, I., Wardle, D.A., Wehrle, V., Zander, R., Zou, J.: Atmospheric chemistry experiment (ACE): mission overview. *Geophys. Res. Lett.* **32**(15) (2005). <https://doi.org/10.1029/2005GL022386>
2. Cui, T., Martin, J., Marzouk, Y.M., Solonen, A., Spantini, A.: Likelihood-informed dimension reduction for nonlinear inverse problems. *Inverse Prob.* **30**(11), 114015, 28 (2014). <https://doi.org/10.1088/0266-5611/30/11/114015>
3. Dils, B., Buchwitz, M., Reuter, M., Schneising, O., Boesch, H., Parker, R., Guerlet, S., Aben, I., Blumenstock, T., Burrows, J.P., Butz, A., Deutscher, N.M., Frankenberg, C., Hase, F., Hasekamp, O.P., Heymann, J., De Mazière, M., Notholt, J., Sussmann, R., Warneke, T., Griffith, D., Sherlock, V., Wunch, D.: The greenhouse gas climate change initiative (GHG-CCI): comparative validation of GHG-CCI CHY/ENVISAT and TANSO-FTS/GOSAT CO₂ and CH₄ retrieval algorithm products with measurements from the TCCON. *Atmos. Meas. Tech.* **7**(6), 1723–1744 (2014). <https://doi.org/10.5194/amt-7-1723-2014>
4. Feldman, D.R., Collins, W.D., Gero, P.J., Torn, M.S., Mlawer, E.J., Shippert, T.R.: Observational determination of surface radiative forcing by CO₂ from 2000 to 2010. *Nature* **519**, 339–343 (2015). <https://doi.org/10.1038/nature14240>
5. Haario, H., Saksman, E., Tamminen, J.: An adaptive Metropolis algorithm. *Bernoulli* **7**(2), 223–242 (2001). <https://doi.org/10.2307/3318737>
6. Haario, H., Laine, M., Mira, A., Saksman, E.: DRAM: Efficient adaptive MCMC. *Stat. Comput.* **16**(4), 339–354 (2006). <https://doi.org/10.1007/s11222-006-9438-0>
7. Kaipio, J., Somersalo, E.: *Statistical and Computational Inverse Problems*. Springer, New York (2005). <https://doi.org/10.1007/b138659>

8. Karion, A., Sweeney, C., Tans, P., Newberger, T.: Aircore: an innovative atmospheric sampling system. *J. Atmos. Oceanic Technol.* **27**(11), 1839–1853 (2010). <https://doi.org/10.1175/2010JTECHA1448.1>
9. Kivi, R., Heikkinen, P.: Fourier transform spectrometer measurements of column CO₂ at Sodankylä, Finland. *Geosci. Instrum. Methods Data Syst.* **5**(2), 271–279 (2016). <https://doi.org/10.5194/gi-5-271-2016>
10. Laine, M.: MCMC Toolbox for Matlab (2013). <http://helios.fmi.fi/~lainema/mcmc/>
11. Ripley, B.D.: *Stochastic Simulation*. Wiley, New York (1987)
12. Rodgers, C.D.: *Inverse Methods for Atmospheric Sounding: Theory and Practice*. World Scientific, Singapore (2000)
13. Solonen, A., Cui, T., Hakkarainen, J., Marzouk, Y.: On dimension reduction in Gaussian filters. *Inverse Prob.* **32**(4), 045003 (2016). <https://doi.org/10.1088/0266-5611/32/4/045003>
14. Spantini, A., Solonen, A., Cui, T., Martin, J., Tenorio, L., Marzouk, Y.: Optimal low-rank approximations of Bayesian linear inverse problems. *SIAM J. Sci. Comput.* **37**(6), A2451–A2487 (2015). <https://doi.org/10.1137/140977308>
15. Tukiainen, S.: Swirlab Toolbox for Matlab (2017). <https://github.com/tukiains/swirlab>
16. Tukiainen, S., Railo, J., Laine, M., Hakkarainen, J., Kivi, R., Heikkinen, P., Chen, H., Tamminen, J.: Retrieval of atmospheric CH₄ profiles from Fourier transform infrared data using dimension reduction and MCMC. *J. Geophys. Res. Atmos.* **121**, 10,312–10,327 (2016). <https://doi.org/10.1002/2015JD024657>
17. Wunch, D., Toon, G.C., Sherlock, V., Deutscher, N.M., Liu, X., Feist, D.G., Wennberg, P.O.: The total carbon column observing network’s GGG2014 data version. Tech. rep., Oak Ridge, Tennessee, U.S.A., Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory (2015). <https://doi.org/10.14291/tccon.ggg2014.documentation.R0/1221662>

Wider Contours and Adaptive Contours



Shev MacNamara, William McLean, and Kevin Burrage

Abstract Contour integrals in the complex plane are the basis of effective numerical methods for computing matrix functions, such as the matrix exponential and the Mittag-Leffler function. These methods provide successful ways to solve partial differential equations, such as convection–diffusion models. Part of the success of these methods comes from exploiting the freedom to choose the contour, by appealing to Cauchy’s theorem. However, the pseudospectra of non-normal matrices or operators present a challenge for these methods: if the contour is too close to regions where the norm of the resolvent matrix is large, then the accuracy suffers. Important applications that involve non-normal matrices or operators include the Black–Scholes equation of finance, and Fokker–Planck equations for stochastic models arising in biology. Consequently, it is crucial to choose the contour carefully. As a remedy, we discuss choosing a contour that is wider than it might otherwise have been for a normal matrix or operator. We also suggest a semi-analytic approach to adapting the contour, in the form of a parabolic bound that is derived by estimating the field of values.

S. MacNamara (✉)

Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), University of Technology Sydney, Sydney, NSW, Australia
e-mail: shev.macnamara@uts.edu.au

W. McLean

The School of Mathematics and Statistics, University of New South Wales, Sydney, NSW, Australia
e-mail: w.mclean@unsw.edu.au

K. Burrage

Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD, Australia, and Visiting Professor to the University of Oxford, UK
e-mail: kevin.burrage@qut.edu.au

To demonstrate the utility of the approaches that we advocate, we study three models in biology: a monomolecular reaction, a bimolecular reaction and a trimolecular reaction. Modelling and simulation of these reactions is done within the framework of Markov processes. We also consider non-Markov generalisations that have Mittag-Leffler waiting times instead of the usual exponential waiting times of a Markov process.

1 Introduction

We begin with the Chapman–Kolmogorov forward equation, associated with a Markov process on discrete states, for the evolution in continuous time of the probability of being in state j at time t :

$$\frac{d}{dt}p(j, t) = -|a_{jj}|p(j, t) + \sum_{i \neq j} a_{j,i}p(i, t). \quad (1)$$

Here for $j \neq i$, $a_{i,j} \geq 0$, and $a_{i,j}dt$ is approximately the probability to transition from state j to state i in a small time dt . The diagonal entry of an associated matrix $\mathbb{A} = \{a_{i,j}\}$ is defined by the requirement that the matrix has columns that sum to zero, namely $a_{jj} = -\sum_{i \neq j} a_{i,j}$. This equation in the Markov setting (1) can be generalised to a non-Markovian form:

$$\frac{d}{dt}p(j, t) = - \int_0^t K(j, t-u)p(j, u)du + \sum_{i \neq j} \frac{a_{j,i}}{|a_{jj}|} \int_0^t K(i, t-u)p(i, u)du. \quad (2)$$

Here the so-called memory function $K(i, t-u)$ is defined via its Laplace transform $\hat{K}(j, s)$, as the ratio of the Laplace transform of the waiting time to the Laplace transform of the survival time. The process is a Markov process if and only if the waiting times are exponential random variables, in which case the $K(j, t)$ appearing in the convolutions in (2) are Dirac delta distributions and (2) reduces to the usual equation in (1). In the special case of Mittag-Leffler waiting times, (2) can be rewritten as [12, 13]:

$$D_t^\alpha \mathbf{p} = \mathbb{A} \mathbf{p} \quad \text{with solution} \quad \mathbf{p}(t) = E_\alpha(\mathbb{A}t^\alpha) \mathbf{p}(0). \quad (3)$$

Here D_t^α denotes the Caputo fractional derivative, and where the Mittag-Leffler function is

$$E_\alpha(z) = \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(\alpha k + 1)}. \quad (4)$$

Here $0 \leq \alpha \leq 1$. In Sect. 4, we provide a MATLAB code to simulate sample paths of this stochastic process with Mittag-Leffler waiting times. When $\alpha = 1$ the series (4) reduces to the usual exponential function, and the fractional equation in (3) reduces to the original Markov process (1). In Sect. 5, we provide a MATLAB code to compute the solution $E_\alpha(\mathbb{A}t^\alpha)\mathbf{p}(0)$ in (3) directly, via a contour integral.

Next, we introduce the (Markovian) Fokker–Planck partial differential equation (sometimes also known as the forward Kolmogorov equation)

$$\frac{\partial}{\partial t} p(x, t) = -\frac{\partial}{\partial x} (a(x)p(x, t)) + \frac{1}{2} \frac{\partial^2}{\partial x^2} (b(x)p(x, t)) \quad (5)$$

for a probability density $p(x, t)$ in one space dimension and time, with suitable boundary conditions, initial conditions, and smoothness assumptions on the coefficients $a(x)$, and on $b(x) \geq 0$. Later in the work we will use complex-variable methods, so it is worth noting at the outset that our coefficients $a(x)$ and $b(x)$ are always real-valued. It is also worth noting that $a(x)$ here in (5) is not the same as $a_{i,j}$ appearing in the matrix above in (3), although there is a close relationship that allows one to be deduced from the other. This PDE for the density corresponds to a stochastic differential equation for sample paths

$$dX = a(X)dt + \sqrt{b(X)}dW. \quad (6)$$

Both the PDE (5) and SDE (6) are continuous in the spatial variable. An introduction to these models, and their connections with discrete versions, can be found in [1, 5]. Our discrete models do respect important properties such as non-negativity. However, there are issues with the Fokker–Planck PDE model, the Chemical Langevin SDE model, and other diffusion approximations: often these approximations do not maintain positivity. These issues are discussed by Williams in the Kolmogorov Lecture and accompanying paper, where a method that maintains non-negativity is proposed [11].

We do not simulate or solve either of the PDE (5) or the SDE (6). We do however simulate and solve the closely related models that are discrete in space and that are governed by master equations (1) or generalized master equations (2), which can be thought of as finite difference discretizations of (5). In particular, the PDE (5) can be thought of as a continuous-in-space analogue of the discrete process in (1) that involves a matrix \mathbb{A} . The utility of the PDE is that it is easier to find an estimate of the field of values of the spatial differential operator on the right hand side of (5) than of the corresponding matrix \mathbb{A} . We can then use an estimate of one as an approximation for the other.

Developing appropriate methods for modelling and simulation of these important stochastic processes is a necessary first step for more advanced scientific endeavors. A natural next step is an inverse problem, although we do not pursue that in this article. For example, it is of great interest to estimate the rate constants in the models described in Sect. 2, typically based on limited observations of samples resembling the simulation displayed in Fig. 4. It is more challenging to estimate the parameter

α in fractional models, or indeed to address the question of whether or not a single α (the only case considered in this article) is appropriate. Interest in fractional models is growing fast as they are finding wide applications including in cardiac modelling [2], and this includes exciting new applications of Mittag-Leffler functions [3].

Section 2 introduces three models that can be cast in the form of (3) and Sect. 3 uses these models as vignettes to exhibit manifestations of pseudospectra. Next we introduce two methods of simulation for these models. A Monte Carlo approach is presented in Sect. 4. Section 5 presents an alternative method that directly computes the solution of (3) as a probability vector via a contour integral. Finally we suggest a bound on the field of values that is useful when designing contour methods.

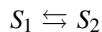
To our knowledge Fig. 3 is the first visualization of the pseudospectra of the Schlögl reactions. The estimate in (24) is also new. In fact (24) is the specialization of our more general estimates appearing in (22) and (23) to the monomolecular model, but it should be possible to likewise adapt our more general estimates to other models such as bimolecular models.

2 Three Fundamental Models

All three models that we present are represented by *tri-diagonal matrices*: $j \notin \{i - 1, i, i + 1\} \Rightarrow \mathbb{A}_{i,j} = 0$. Since all other entries are zero, below, we only specify the non-zero entries on the three main diagonals. In fact, an entry on the main diagonal is determined by the requirement that columns sum to zero (which corresponds to conservation of probability), so it would suffice to specify only the two non-zero diagonals immediately below and above the main diagonal.

2.1 Monomolecular, Bimolecular and Trimolecular Models

A model of a monomolecular reaction



such as arises in chemical isomerisation [5], or in models of ion channels in cardiac-electrophysiology, or neuro-electrophysiology, can be represented by the following $N \times N$ matrix.

Monomolecular model matrix

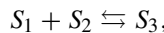
$$\mathbb{A}_{i,j} = \begin{cases} j-1, & i = j-1, \\ -m, & i = j, \\ m-j+1, & i = j+1. \end{cases} \quad (7)$$

Here we have assumed that the rate constants are equal to unity, $c_1 = c_2 = 1$, and we have assumed that $m = N - 1$ where m is the maximum number of molecules. More details, including exact solutions, can be found in [8]. An instance of this matrix when $m = 5$ is

$$\mathbb{A} = \begin{pmatrix} -5 & 1 & & & \\ & 5 & -5 & 2 & \\ & & 4 & -5 & 3 \\ & & & 3 & -5 & 4 \\ & & & & 2 & -5 & 5 \\ & & & & & 1 & -5 \end{pmatrix}. \quad (8)$$

The model is two-dimensional, but the conservation law allows it to be treated as effectively one-dimensional, by letting x represent S_1 , say. Then one possible corresponding continuous model (5) has drift coefficient $a(x) = -c_1x + c_2(m-x)$ and diffusion coefficient $b(x) = c_1x + c_2(m-x)$, for $0 < x < m$. In the discrete Markov case, the exact solution is binomial. When $c_1 = c_2$ the stationary probability density of the continuous Markov model is given by Gillespie [5] as a Gaussian, which highlights the issues associated with the continuous models such as choosing the domain, boundary conditions, and respecting positivity. To enforce positivity, we might instead pose the PDE on the restricted domain $0 \leq x \leq m$.

Next we introduce a model for the bimolecular reaction



via the following $N \times N$ matrix.

Bimolecular model matrix

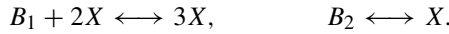
$$\mathbb{A}_{i,j} = \begin{cases} j-1, & i = j-1, \\ -j+1 - (m-j+1)^2 & i = j, \\ (m-j+1)^2, & i = j+1. \end{cases} \quad (9)$$

This matrix and the model are also introduced in [12], where more details can be found. A small example with $m = 5$ is

$$\mathbb{A} = \begin{pmatrix} -25 & 1 & & & \\ & 25 & -17 & 2 & \\ & & 16 & -11 & 3 \\ & & & 9 & -7 & 4 \\ & & & & 4 & -5 & 5 \\ & & & & & 1 & -5 \end{pmatrix}. \quad (10)$$

Here we have assumed that the rate constants are equal to unity, $c_1 = c_2 = 1$, and that the initial condition is $[S_1, S_2, S_3] = [m, m, 0]$, so m is the maximum number of molecules, and $m = N - 1$. The model is three-dimensional, but the conservation law together with this initial condition allow it to be treated as effectively one-dimensional, by letting x represent S_3 , say. Then a possible corresponding continuous model (5) has drift coefficient $a(x) = -c_1x + c_2(m-x)^2$ and diffusion coefficient $b(x) = c_1x + c_2(m-x)^2$.

Finally, we introduce the Schlögl model [13], which consists of two reversible reactions



Here $B_1 = 1 \times 10^5$, $B_2 = 2 \times 10^5$. The associated matrix is given below, where $k_1 = 3 \times 10^{-7}$, $k_2 = 1 \times 10^{-4}$, $k_3 = 1 \times 10^{-3}$, and $k_4 = 3.5$. In theory this matrix is infinite but we truncate to a finite section for numerical computation.

Schlögl model matrix (an example of a **trimolecular model** scheme)

$$\mathbb{A}_{i,j} = \begin{cases} \frac{1}{6}k_2(j-1)(j-2)(j-3) + k_4(j-1), & i = j-1, \\ k_3B_2 + \frac{1}{2}k_1B_1(j-1)(j-2), & i = j+1. \end{cases} \quad (11)$$

For $i = j$, the diagonal entry is $-\left(\frac{1}{6}k_2(j-1)(j-2)(j-3) + k_4(j-1) + k_3B_2 + \frac{1}{2}k_1B_1(j-1)(j-2)\right)$. The first column is indexed by $j = 1$ and corresponds to a state with $0 = j - 1$ molecules. The corresponding continuous model (5) has drift coefficient $a(x) = k_3B_2 + \frac{1}{2}k_1B_1x(x-1) - \frac{1}{6}k_2x(x-1)(x-2) - k_4x$ and diffusion coefficient $b(x) = k_3B_2 + \frac{1}{2}k_1B_1x(x-1) + \frac{1}{6}k_2x(x-1)(x-2) + k_4x$.

3 Pseudospectra Are Important for Stochastic Processes

All three matrices introduced in the previous section exhibit interesting pseudospectra. As an illustration of the way that the pseudospectra manifest themselves, we will now consider numerically computing eigenvalues of the three matrices. This is a first demonstration that one must respect the pseudospectrum when crafting a numerical method. That issue will be important again when we use numerical methods based on contour integrals to compute Mittag-Leffler matrix functions.

The reader can readily verify that using any standard eigenvalue solver, such as `eig` in MATLAB, leads to numerically computed eigenvalues that are complex numbers. However, these numerically computed complex eigenvalues are *wrong*: the *true eigenvalues are purely real*. It is the same story for all three models. See the numerically computed eigenvalues displayed here in Figs. 1 and 2, for example. We suggest this effect happens much more widely for models arising in computational biology.

Figures 1 and 2 make use of the following method to create a diagonal scaling matrix. Here is a MATLAB listing to create a diagonal matrix D that symmetrizes a

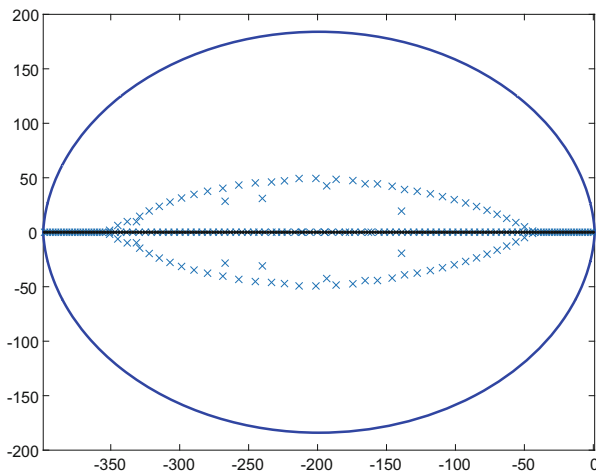


Fig. 1 The field of values (solid line resembling an oval-like shape) for the discrete monomolecular matrix in (7) when $N = 200$ ($m = N - 1$), as computed by Chebfun [4]. The crosses mark numerically computed eigenvalues via `eig`, but complex eigenvalues are *wrong*. The true eigenvalues are purely real and are marked on the real axis by dots (note that the dots are so close together that they may seem to be almost a solid line). These correct values can come by instead computing the eigenvalues of the symmetrized matrix, after using the diagonal scaling matrix created by the MATLAB listing provided here

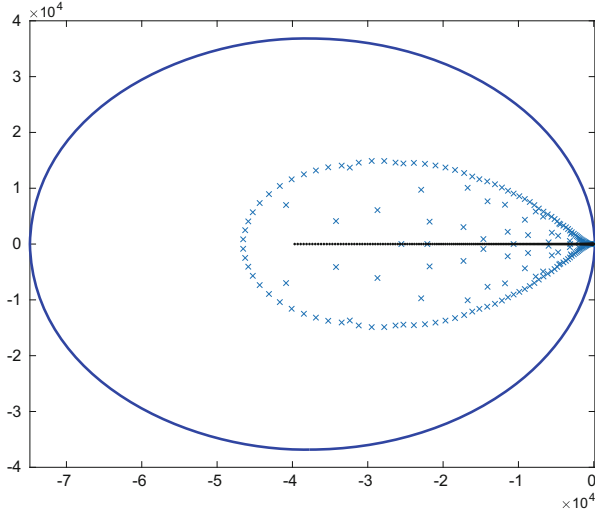


Fig. 2 Same as Fig. 1, for the bimolecular model. The field of values for the discrete bimolecular matrix in (9) when $N = 200$ is the solid line resembling an oval-like shape, as computed by Chebfun [4]

tridiagonal matrix of dimension N of the form described in any of the three models considered here:

```

d(1) = 1;
for i = 1:N-1
    d(i+1) = sqrt(A(i,i+1)/A(i+1,i)) *d(i);
end
D = diag(d);    Asym = D*A*inv(D);

```

This symmetrization by a diagonal matrix in a similarity transform is known to physicists as a gauge transformation, and it is described by Trefethen and Embree [15, Section 12]. A real symmetric matrix has real eigenvalues so the eigenvalues of DAD^{-1} are purely real. The matrix DAD^{-1} and the matrix A share the same eigenvalues because they are similar. This is one way to confirm that the true eigenvalues of A are purely real. Numerical algorithms typically perform well on real symmetric matrices, so it is better to numerically compute the eigenvalues of DAD^{-1} than to compute eigenvalues of A directly.

The ϵ -pseudospectra [15] can be defined as the region of the complex plane where the norm of the resolvent matrix is large: the set $z \in \mathbb{C}$ such that

$$\|(z\mathbb{I} - \mathbb{A})^{-1}\| > \frac{1}{\epsilon}. \quad (12)$$

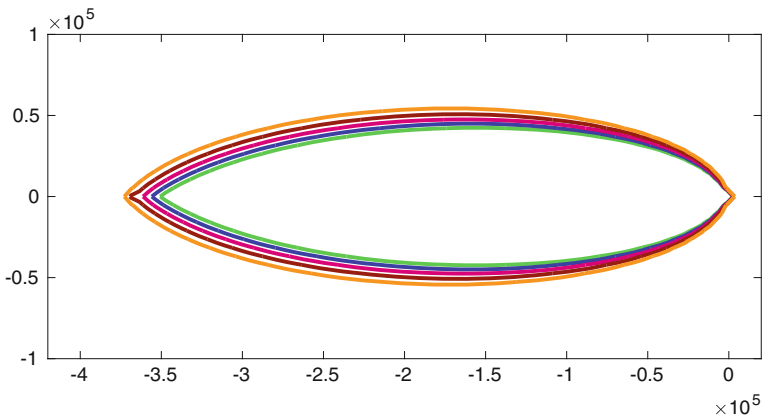


Fig. 3 Pseudospectrum of a 2000×2000 finite section of the matrix in (11) representing the Schlögl reactions, as computed by EigTool. Contours correspond to $\epsilon = 10^{-2}, 10^{-4}, 10^{-6}, 10^{-8}, 10^{-10}$ in (12). The contour closest to the real axis corresponds to 10^{-10}

We use the 2-norm in this article. Equivalently, this is the region of the complex plane for which z is an eigenvalue of $(\mathbb{A} + \mathbb{E})$ for some perturbing matrix that has a small norm: $\|\mathbb{E}\| < \epsilon$. Thus the wrong complex ‘eigenvalues’ appearing as crosses in Figs. 1 and 2 offer a way to (crudely) visualise the pseudospectrum. Numerical visualizations of the pseudospectra of the family of monomolecular matrices defined by (7) can be found in [8], and the pseudospectra of the bimolecular reaction defined by (9) can be found in [12]. Here, as an example of a trimolecular scheme, we present in Fig. 3 the pseudospectra for the Schlögl reactions (11), as computed by EigTool. The resolvent norm is largest near the negative real axis, as we would expect because that is where the true eigenvalues are located. The level curves are not equally spaced in the figure, becoming bunched up together, suggesting a more interesting structure that remains to be elucidated as the dimension of the matrix grows. An interesting experiment is to vary the parameters, namely the product $k_3 B_2$, as is done in [14, Figure 10]. In experiments not shown here, we find that when $k_3 B_2$ is varied, the numerical computation of the eigenvalues becomes less reliable (for example, when $k_3 = 1.4 \times 10^{-3}$ and B_2 is unchanged).

Figures 1 and 2 also display the *numerical range* of a matrix. That is also known as the *field of values*, which we denote by W , and it is defined for a matrix \mathbb{A} , as the set of complex numbers that come from a quadratic form with the matrix

$$W(\mathbb{A}) \equiv \{x^* \mathbb{A} x \in \mathbb{C} : \|x\|_2 = 1\}. \tag{13}$$

The field of values always contains the eigenvalues. A simple algorithm [9] for computing $W(\mathbb{A})$ involves repeatedly ‘rotating’ the matrix and then finding the *numerical abscissa*

$$\frac{1}{2} \max(\text{eig}(\mathbb{A} + \mathbb{A}^*)). \tag{14}$$

The ϵ -pseudospectrum of a matrix is contained in an ϵ -neighbourhood of the field of values, in a sense that can be made a precise theorem [15]. We see an example of this in Figs. 1 and 2.

4 A Mittag-Leffler Stochastic Simulation Algorithm

In this short section we provide a method for Monte Carlo simulation of the solutions of the stochastic processes that we describe. This Monte Carlo method can be considered an alternative to contour integral methods, that we describe later. As the Monte Carlo methods do not seem to fail in the presence of non-normality, they can be a useful cross-check on contour integral methods.

Here is a MATLAB listing to simulate Monte Carlo sample paths of the monomolecular model of isomerization corresponding to the matrix in (7), with Mittag-Leffler waiting times:

```
t_final = 100; m=10; c1=1; c2=1; v = [-1, 1; 1, -1];
initial_state = [m,0]'; alpha = 0.9; all = 1/alpha;
alpi=alpha*pi; sinalpi=sin(alpi); cosalpi=cos(alpi);
t = 0; x = initial_state; T = [0]; X = [initial_state];
while (t < t_final)
    a(1) = c1*x(1); a(2) = c2*x(2); asum = sum(a);
    r1=rand; r2=rand; z=sinalpi/tan(alpi*r2)-cosalpi;
    tau = -(z/asum)^(all)*log(r1);
    r = rand*asum; j = find(r<cumsum(a),1,'first');
    x = x + v(:,j); t =t+tau; T = [T t]; X = [X x];
end
if (t > t_final)
    T(end) = t_final; X(:,end) = X(:,end-1);
end
figure(1); stairs(T, X(1,:), 'LineWidth', 2);
xlabel('Time'); ylabel('molecules of S_1');
title({'Isomerisation: a monomolecular model'; ...
'Mittag-Leffler SSA'; ['\alpha == ', num2str(alpha)]})
```

This is a Markov process with exponential waiting times when $\alpha = 1$, in which case the program reduces to a version of the Gillespie Stochastic Simulation Algorithm. Figure 4 shows a sample path simulated with this MATLAB listing. A histogram of many such Monte Carlo samples is one way to approximate the solution of (3). A different way to compute that solution is via a contour integral, as we describe in the next section, and as displayed in Fig. 5.

A time-fractional diffusion equation in the form of (3) in two space dimensions is solved by such a contour integral method in [16, Figure 16.2]. As an aside, we can modify the code above to offer a method for Monte Carlo simulation of sample paths

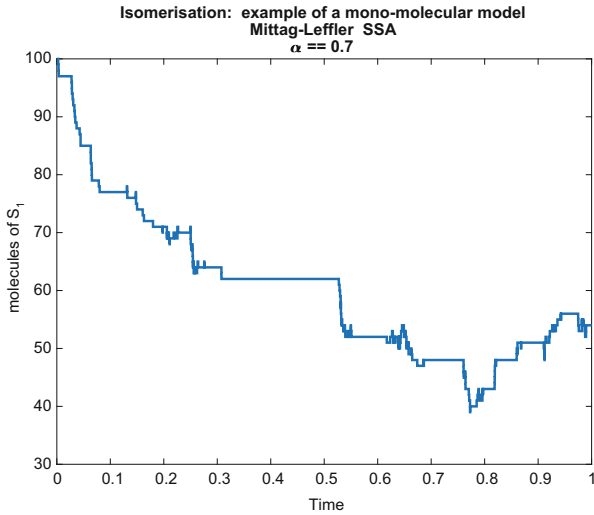


Fig. 4 A sample simulation of the monomolecular reaction, with the Mittag-Leffler SSA provided here in the MATLAB listing. Parameters: $\alpha = 0.7$, $t = 1$, and initial condition a Dirac delta distribution on the state $[S_1, S_2] = [m, 0] = [100, 0]$. Compare with Fig. 5

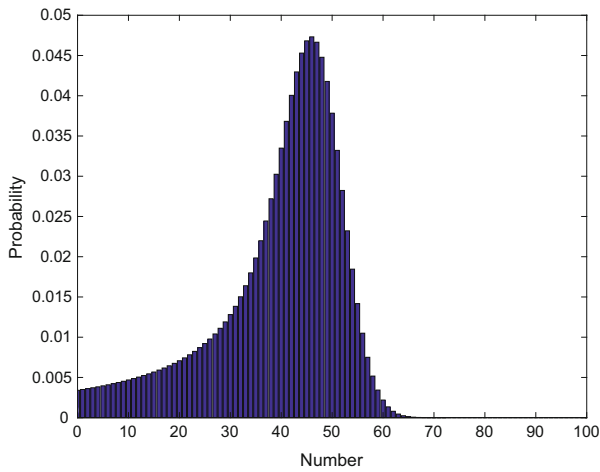


Fig. 5 A discrete probability mass function is the solution of a Mittag-Leffler master Eq. (3) for the monomolecular reaction, and can be computed with the MATLAB listing provided here that uses a hyperbolic contour. Parameters: $\alpha = 0.7$, $t = 1$, and initial condition a Dirac delta distribution on the state $[S_1, S_2] = [m, 0] = [100, 0]$. The x -axis shows the number of molecules of S_2 . Compare with Fig. 4

of a closely related stochastic process, as follows. Simplify to one space dimension (it is also straight forward to simulate in two dimensions), and by supposing that the governing $(m + 1) \times (m + 1)$ matrix is

$$\mathbb{A} = \begin{pmatrix} -1 & 1 & & & \\ & 1 & -2 & 1 & \\ & & & \dots & \\ & & & & 1 & -2 & 1 \\ & & & & & & 1 & -1 \end{pmatrix} \quad (15)$$

with initial vector $p(0)$ being zero everywhere except the middle entry, i.e. the $\text{round}(m/2)$ th entry is one. Then (3) corresponds to a random walk on the integers $0, 1, \dots, m$, beginning in the middle. To simulate, modify the first few lines of the above code segment to

```
t_final = 1; m=10; v = [-1, 1];
initial_state = round(m/2); ...
```

and also the first line in the while loop, to

```
a(1) = (x>0); a(2) = (x<m);
```

leaving the rest unchanged.

5 Computing a Mittag-Leffler Matrix Function

In this section we first establish the utility of contour integral methods by directly computing the desired solution of (3). This is motivation for exploring bounds on the pseudospectrum, so that informed choices can be made when designing contour methods.

5.1 Computing Contour Integrals

Start with (3). Take the Laplace transform. Then take the inverse Laplace transform. Of course those two steps arrive at the same solution we started with. The advantage is the desired solution of (3) represented as a contour integral [12, 13, 16]:

$$p(t) = E_\alpha(\mathbb{A}t^\alpha)p(0) = \frac{1}{2\pi i} \int_\Gamma \exp(zt) \left(z\mathbb{I} - z^{1-\alpha}\mathbb{A} \right)^{-1} p(0) dz. \quad (16)$$

The eigenvalues of \mathbb{A} must lie to the left of the contour Γ . In all our examples, they lie along the negative real axis. Also, we exploit symmetry when \mathbb{A} is real. Apart from those requirements, we are free to choose the contour. Typical choices

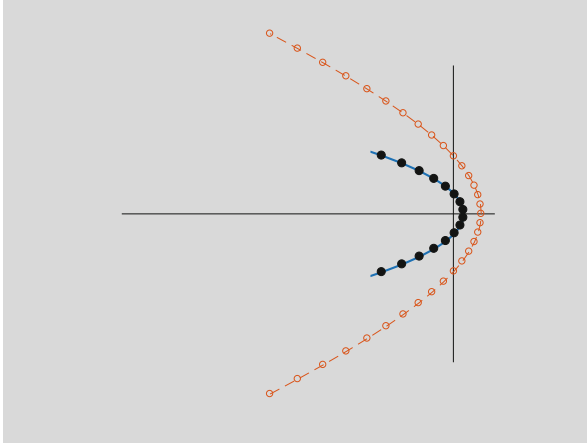


Fig. 6 A parabolic contour (solid) as described in [16], and a hyperbolic contour (dashed, with $M = 16$) used in the MATLAB listing provided here. Nodes for a quadrature rule as in (17) are also marked on the contours

are displayed in Fig. 6. The $\exp(z t)$ factor is nearly zero when the real part of z is sufficiently negative. This motivates choosing Γ to go into the left-half plane because then we can neglect the infinite part of the contour that lies to the left of say, -30 , and still obtain high accuracy. To evaluate the integral on the remaining finite part of the contour, we use quadrature. The trapezoidal rule is a good choice. For a quadrature rule with M nodes z_k on the contour, we approximate (16) by

$$\mathbf{p}(t) \approx \sum_{k=1}^M w_k(t) \mathbf{u}_k \tag{17}$$

where we solve a linear system for the vector \mathbf{u}_k at each node k

$$\left(z\mathbb{I} - z^{1-\alpha} \mathbb{A} \right) \mathbf{u}_k = \mathbf{p}(0).$$

The values $w_k(t)$ (and note that we allow these numbers to depend on t) incorporate both the usual weights coming from the quadrature rule on a line, and also any scalings coming from parameterizing the contour. There are variations of this quadrature scheme. When $\alpha = 1$, this procedure simplifies to compute the familiar matrix exponential solutions of the usual Markov systems.

Here is a MATLAB listing, adapted from Le et al. [10], to compute the Mittag-Leffler solution by applying the quadrature recipe (17) on a *hyperbolic* contour.

```

function p = hyperbola_MittagLeffler(t,A,v,alpha,M)
alpha = 1-alpha; n = length(v); Dxi = 1.08179214/M;
xi = [-M:M]*Dxi; delta=1.17210423; mu =4.49207528*M/t;
z = mu * ( 1 - sin(complex(delta,-xi)) );
dz = 1i * mu * cos(complex(delta,-xi)) ;
c = Dxi * dz .* exp(z*t) / (2*pi*1i);
I = speye(n); p = zeros(size(v));
for k = 1:M
    p = p + c(k) * ((z(k)*I-z(k)^(alpha)*A)\v);
end
p = 2*real(p); k = M+1;
p = p + real(c(k) * ((z(k)*I-z(k)^(alpha)*A)\v));

```

We choose $M = 16$ quadrature points. An example of the hyperbolic contour being used here is displayed in Fig. 6. Figure 5 shows a solution of (3) computed with this MATLAB listing.

5.2 Estimating the Field of Values

One strategy to choose the contour, Γ , is to first compute a pseudospectrum of the matrix with `EigTool` [17] (as in Fig. 3 for example). Then choose Γ so that $\|(z\mathbb{I} - \mathbb{A})^{-1}\|$ is never too large for $z \in \Gamma$. For example, we might choose the contour so that the bound $\|(z\mathbb{I} - \mathbb{A})^{-1}\| < 10^3$ holds (which corresponds to $\epsilon = 10^{-3}$ in (12)). Arguably, the value 10^3 could be replaced by something smaller, say $\mathcal{O}(1)$. The particular value would depend on the application. In numerical experiments with these models, the issue seems to matter only when $\|(z\mathbb{I} - \mathbb{A})^{-1}\|$ is significantly larger than, say, 100. Such an unfavourable situation can certainly arise. It is demonstrated to happen on the parabolic contour displayed here in Fig. 6 for bimolecular reactions [12]. The issue is also addressed by In't Hout and Weideman [7] for Black–Scholes models. Figure 6 compares a parabolic contour used for a bimolecular model [12] with a hyperbolic contour used here. A similar comparison of these contours and discussion can be found in a survey article by Trefethen and Weideman [16, Figure 15.2].

However, a drawback of this procedure is that it is expensive to first compute the pseudospectrum. It would therefore be preferable to instead find an estimate of the region where the resolvent is large by some more efficient means. We find one such estimate next.

Trefethen and Embree [15] discuss various ways to bound the pseudospectrum. One approach uses the fact that if z is not inside the field of values of a matrix, then the norm of the resolvent is bounded by the distance to the field of values:

$$\|(z\mathbb{I} - \mathbb{A})^{-1}\|_2 < \frac{1}{\text{dist}(z, W(\mathbb{A}))}.$$

This result suggests an idea for adapting the contour: choose the contour to be outside of the field of values.

We now estimate the field of values of the spatial operator in the Fokker–Planck PDE (5); a similar approach has been applied to the Black–Scholes equation of finance [7, Theorem 3.1]. We focus on the particular example of the monomolecular model, but we believe this approach will be extended to the other models in future work.

Begin by writing the Fokker–Planck equation in the form of a conservation law,

$$u_t + Au = 0 \quad \text{for } 0 < x < m \text{ and } t > 0,$$

where, using a dash for $\partial/\partial x$,

$$Au = -\frac{1}{2}(bu)'' + (au)' = -\left(\frac{1}{2}bu' + Bu\right)' \quad \text{and} \quad B(x) = -a(x) + \frac{1}{2}b'(x). \quad (18)$$

For the monomolecular model, the coefficients are

$$b(x) = c_1x + c_2(m - x) \quad \text{and} \quad a(x) = -c_1x + c_2(m - x).$$

Here, c_1, c_2 and m are positive constants. Note that A is uniformly elliptic because $b(x) \geq \min(c_1, mc_2)$ for $0 < x < m$.

We impose either homogeneous Dirichlet boundary conditions,

$$u(0) = 0 = u(m), \quad (19)$$

or else zero-flux boundary conditions,

$$\frac{1}{2}bu' + Bu = 0 \quad \text{for } x \in \{0, m\}. \quad (20)$$

The domain of A is then the complex vector space $D(A)$ of C^2 functions $v : [0, m] \rightarrow \mathbb{C}$ satisfying the chosen boundary conditions.

Denote the numerical range (field of values) of A by

$$W(A) = \{ \langle Au, u \rangle : u \in D(A) \text{ with } \langle u, u \rangle = 1 \}, \quad (21)$$

where $\langle u, v \rangle = \int_0^m u \bar{v}$. Compare this definition of $W(A)$ in the continuous setting with the definition (13) in the discrete setting. For either (19) or (20), integration by parts gives

$$\langle Au, u \rangle = \frac{1}{2} \int_0^m b|u'|^2 + \int_0^m Bu\bar{u}'.$$

We write

$$X + iY \equiv \langle Au, u \rangle = \frac{1}{2}P - Q \quad \text{where} \quad P = \int_0^m b|u'|^2 \quad \text{and} \quad Q = - \int_0^m B u \bar{u}',$$

and assume that

$$\frac{B(x)^2}{b(x)} \leq K \quad \text{and} \quad 0 < \beta_0 \leq \frac{B'(x)}{2} \leq \beta_1 \quad \text{for } 0 \leq x \leq m.$$

In the case of zero-flux boundary conditions (20), we require the additional assumption

$$B(0) \leq 0 \leq B(m).$$

Then we claim that for Dirichlet boundary conditions (19),

$$Y^2 \leq 2K(X + \beta_1) - \beta_0^2, \tag{22}$$

whereas for zero-flux boundary conditions (20),

$$Y^2 \leq 2K(X + \beta_1). \tag{23}$$

To derive these estimates, first observe that

$$2\Re Q = Q + \bar{Q} = - \int_0^m B(u\bar{u}' + \bar{u}u') = - \int_0^m B(u\bar{u})' = -[B|u|^2]_0^m + \int_0^m B'|u|^2.$$

Assume that $\langle u, u \rangle = 1$. The bounds on B' give

$$\beta_0 - \left[\frac{1}{2}B|u|^2\right]_0^m \leq \Re Q \leq \beta_1 - \left[\frac{1}{2}B|u|^2\right]_0^m,$$

and by the Cauchy–Schwarz inequality,

$$|Q|^2 \leq \left(\int_0^m B^2|u'|^2\right)\left(\int_0^m |u|^2\right) = \int_0^m \frac{B^2}{b} b|u'|^2 \leq \left(\max_{[0,m]} \frac{B^2}{b}\right) \int_0^m b|u'|^2 \leq KP;$$

thus,

$$Y^2 = (-\Im Q)^2 = |Q|^2 - (\Re Q)^2 \leq KP - (\Re Q)^2.$$

For Dirichlet boundary conditions we have $[B|u|^2]_0^m = 0$, so $\beta_0 \leq \Re Q \leq \beta_1$ and hence

$$P = 2X + 2\Re Q \leq 2X + 2\beta_1 \quad \text{and} \quad Y^2 \leq KP - \beta_0^2,$$

implying (22). For zero-flux boundary conditions, the extra assumptions on B ensure that $[B|u|^2]_0^m \geq 0$, so $\Re Q \leq \beta_1$ and hence

$$P = 2X + 2\Re Q \leq 2X + 2\beta_1 \quad \text{and} \quad Y^2 \leq KP,$$

implying (23).

In the simple case $c_1 = c_2 = c$ we have

$$b(x) = cm \quad \text{and} \quad B(x) = -a(x) = c(2x - m)$$

and therefore

$$\frac{B(x)^2}{b(x)} = \frac{c}{m}(2x - m)^2 \leq cm \quad \text{and} \quad B'(x) = 2c \quad \text{for } 0 \leq x \leq m,$$

giving $K = cm$ and $\beta_0 = \beta_1 = c$. In addition, $B(0) = -cm \leq 0$ and $B(m) = cm \geq 0$. Thus, for zero-flux boundary conditions,

$$Y^2 \leq 2cm(X + c).$$

Note that the sign convention in the notation in (18) makes the operator A positive definite (though not symmetric), whereas our model matrices are negative definite, so $W(A)$ provides an approximation for $W(-\mathbb{A}) = -W(\mathbb{A})$. We therefore have to flip signs in (22) and (23) to get estimates for $W(\mathbb{A})$, as in the following bound.

Parabolic estimate of the field of values for the **monomolecular model** matrix when $c_1 = c_2 = c$ in (7):

$$Y^2 \leq 2cm(c - X). \tag{24}$$

This bound is displayed in Fig. 7. As in the figure, our matrix examples typically have a unique zero eigenvalue, with all other eigenvalues having negative real part, and the numerical abscissa (14) is typically strictly positive. Having derived a bound in the continuous setting, we can only regard it as an approximation in the discrete setting. Nonetheless, in these numerical experiments it does indeed seem to offer a useful bound for the discrete case.

Discussion A natural next step is to incorporate this bound into the design of the contour for methods such as the MATLAB listing provided in Sect. 5. That will be pursued elsewhere. This article has thus laid the foundations for such attractive future research.

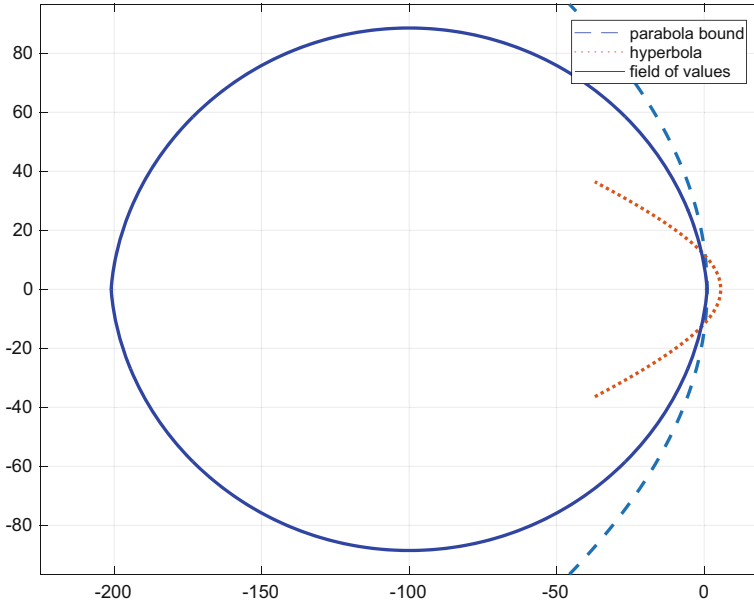


Fig. 7 The field of values for the discrete matrix in (7) when $c_1 = c_2 = 1$ and $m = 100$, as computed by `Chebfun`. The dashed line is a parabolic bound in (24) for the field of values of the corresponding continuous PDE (5). This parabolic bound is semi-analytic and requires some analysis of the equation ‘by hand.’ Also displayed is a hyperbolic contour, as in Fig. 6 as used in the MATLAB listing provided here

It is worth commenting that if a matrix is real symmetric (unlike the examples in this article) then the pseudospectrum is not an issue so it would not be good to make the contour wider (and thus also longer), and instead previously proposed (and often optimized and shorter) contours such as surveyed in [16] would be good choices. Making the contour wider and thus longer does impact the numerical method, but this is unavoidable in applications where the behaviour of the pseudospectra is an issue, such as the examples discussed here. It is also worth commenting on numerical methods for computing the field of values. Here we used `Chebfun`, which in turn is based on Johnson’s algorithm [9]. It is conceivable that a different algorithm might be devised to estimate the field of values more cheaply for the purposes of contour integrals, but we do not explore such an approach here. However, we do observe in these numerical experiments, as displayed in the figures for example, that the field of values—and therefore also the estimate that we derive from it—seems to be too conservative. That is especially noticeable in regions where the real part is large and negative. It might therefore also be worth considering methods of more cheaply, directly, estimating the pseudospectra. We began with the continuous PDE so using a coarse mesh in a PDE-solver, such as a finite difference method, might be one way to obtain a cheap estimate of the pseudospectra.

6 Conclusion

We have discussed, and illustrated by example, the significance of the pseudospectra for stochastic models in biology, including both Markovian and non-Markovian generalisations. Although we focused exclusively on the 2-norm, the pseudospectra of these stochastic processes are perhaps more naturally addressed in the 1-norm. In that regard, future work will explore ways to incorporate the 1-norm methods described by Higham and Tisseur [6], to adaptively choosing the contour.

Numerical methods to compute the matrix exponential functions and matrix Mittag-Leffler functions via contour integrals must take the pseudospectrum of the matrix into account. In particular, such methods must choose contours that are wide enough, and that adapt to avoid regions of the complex plane where the norm of the resolvent matrix is too large. We have derived a simple, parabolic bound on the field of values of an associated Fokker–Planck PDE, which can be used as an approximation to the field of values of the corresponding discrete matrix model. We believe this bound can help inform us when making a good choice for the contour. Ultimately, how to devise contours in a truly adaptive fashion, so that they can be efficiently computed, automatically and without a priori analysis, remains an important open question.

Acknowledgements The authors thank the organisers of the Computational Inverse Problems theme at the MATRIX, 2017.

References

1. Anderson, D.F., Kurtz, T.G.: Continuous time Markov chain models for chemical reaction networks. In: Design and Analysis of Biomolecular Circuits. Springer, New York (2011)
2. Bueno-Orovio, A., Kay, D., Grau, V., Rodriguez, B., Burrage, K.: Fractional diffusion models of cardiac electrical propagation: role of structural heterogeneity in dispersion of repolarization. *J. R. Soc. Interface* **11**(97), 20140352 (2014)
3. Bueno-Orovio, A., Teh, I., Schneider, J.E., Burrage, K., Grau, V.: Anomalous diffusion in cardiac tissue as an index of myocardial microstructure. *IEEE Trans. Med. Imaging* **35**(9), 2200–2207 (2016)
4. Driscoll, T.A., Hale, N., Trefethen, L.N.: Chebfun Guide. Pafnuty Publications (2014). <http://www.chebfun.org/docs/guide/>
5. Gillespie, D.T.: The chemical Langevin and Fokker–Planck equations for the reversible isomerization reaction. *J. Phys. Chem. A* **106**(20), 5063–5071 (2002). <https://doi.org/10.1021/jp0128832>
6. Higham, N.J., Tisseur, F.: A block algorithm for matrix 1-norm estimation, with an application to 1-norm pseudospectra. *SIAM J. Matrix Anal. Appl.* **21**(4), 1185–1201. <https://doi.org/10.1137/S0895479899356080>
7. in 't Hout, K.J., Weideman, J.A.C.: A contour integral method for the Black–Scholes and Heston equations. *SIAM J. Sci. Comput.* **33**, 763–785 (2011). <https://doi.org/10.1137/090776081>
8. Iserles, A., MacNamara, S.: Magnus expansions and pseudospectra of Markov processes (2017). <https://arxiv.org/abs/1701.02522>

9. Johnson, C.: Numerical determination of the field of values of a general complex matrix. *SIAM J. Numer. Anal.* **15**, 595–602 (1978)
10. Le, K.N., Mclean, W., Lamichhane, B.: Finite element approximation of a time-fractional diffusion problem for a domain with a re-entrant corner. *ANZIAM J.* **59**, 1–22 (2017)
11. Leite, S.C., Williams, R.J.: A constrained Langevin approximation for chemical reaction networks. Kolmogorov Lecture, Ninth World Congress In Probability and Statistics, Toronto (2016). <http://www.math.ucsd.edu/~williams/biochem/biochem.pdf>
12. Macnamara, S.: Cauchy integrals for computational solutions of master equations. *ANZIAM J.* **56**, 32–51 (2015). <https://doi.org/10.21914/anziamj.v56i0.9345>
13. MacNamara, S., Henry, B.I., McLean, W.: Fractional Euler limits and their applications. *SIAM J. Appl. Math.* **54**, 1763–1784 (2016)
14. Sargsyan, K., Debusschere, B., Najm, H., Marzouk, Y.: Bayesian inference of spectral expansions for predictability assessment in stochastic reaction networks. *J. Comput. Theor. Nanosci.* **6**, 2283–2297 (2009)
15. Trefethen, L.N., Embree, M.: *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton (2005)
16. Trefethen, L.N., Weideman, J.A.C.: The exponentially convergent trapezoidal rule. *SIAM Review* **56**(3), 385–458 (2014)
17. Wright, T.G.: *Eigtool* (2002). <http://www.comlab.ox.ac.uk/pseudospectra/eigtool/>

Bayesian Point Set Registration



Adam Spannaus, Vasileios Maroulas, David J. Keffer, and Kody J. H. Law

Abstract Point set registration involves identifying a smooth invertible transformation between corresponding points in two point sets, one of which may be smaller than the other and possibly corrupted by observation noise. This problem is traditionally decomposed into two separate optimization problems: (1) assignment or correspondence, and (2) identification of the optimal transformation between the ordered point sets. In this work, we propose an approach solving both problems simultaneously. In particular, a coherent Bayesian formulation of the problem results in a marginal posterior distribution on the transformation, which is explored within a Markov chain Monte Carlo scheme. Motivated by Atomic Probe Tomography (APT), in the context of structure inference for high entropy alloys (HEA), we focus on the registration of noisy sparse observations of rigid transformations of a known reference configuration. Lastly, we test our method on synthetic data sets.

1 Introduction

In recent years, a new class of materials has emerged, called High Entropy Alloys (HEAs). The resulting HEAs possess unique mechanical properties and have shown marked resistance to high-temperature, corrosion, fracture and fatigue [5, 18]. HEAs

A. Spannaus · V. Maroulas
Department of Mathematics, University of Tennessee, Knoxville, TN, USA
e-mail: spannaus@utk.edu; maroulas@math.utk.edu

D. J. Keffer
Department of Materials Science and Engineering, University of Tennessee, Knoxville, TN, USA
e-mail: dkeffer@utk.edu

K. J. H. Law (✉)
Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN,
USA

School of Mathematics, University of Manchester, Manchester, UK

demonstrate a ‘cocktail’ effect [7], in which the mixing of many components results in properties not possessed by any single component individually. Although these metals hold great promise for a wide variety of applications, the greatest impediment in tailoring the design of HEAs to specific applications is the inability to accurately predict their atomic structure and chemical ordering. This prevents Materials Science researchers from constructing structure-property relationships necessary for targeted materials discovery.

An important experimental characterization technique used to determine local structure of materials at the atomic level is Atomic Probe Tomography (APT) [8, 10]. APT provides an identification of the atom type and its position in space within the sample. APT has been successfully applied to the characterization of the HEA, AlCoCrCuFeNi [16]. Typically, APT data sets consist of 10^6 – 10^7 atoms. Sophisticated reconstruction techniques are employed to generate the coordinates based upon the construction of the experimental apparatus. APT data has two main drawbacks: (1) up to 66% of the data is missing and (2) the recovered data is corrupted by noise. The challenge is to uncover the true atomic level structure and chemical ordering amid the noise and missing data, thus giving material scientists an unambiguous description of the atomic structure of these novel alloys. Ultimately, our goal is to infer the correct spatial alignment and chemical ordering of a dataset, herein referred to as a configuration, containing up to 10^7 atoms. This configuration will be probed by individual registrations of the observed point sets in a neighborhood around each atom.

In this paper we outline our approach to this unique registration problem of finding the correct chemical ordering and atomic structure in a noisy and sparse dataset. While we do not solve the problem in full generality here, we present a Bayesian formulation of the model and a general algorithmic approach, which allows us to confront the problem with a known reference, and can be readily generalized to the full problem of an unknown reference.

In Sect. 2 we describe the problem and our Bayesian formulation of the statistical model. In Sect. 3, we describe Hamiltonian Monte Carlo, a sophisticated Markov chain Monte Carlo technique used to sample from multimodal densities, which we use in our numerical experiments in Sect. 4. Lastly, we conclude with a summary of the work presented here and directions for future research.

2 Problem Statement and Statistical Model

An alloy consists of a large configuration of atoms, henceforth “points”, which are rotated and translated instances of a reference collection of points, denoted $\mathbf{X} = (X_1, \dots, X_N)$, $X_i \in \mathbb{R}^d$ for $1 \leq i \leq N$ which is the matrix representation of the reference points. The tomographic observation of this configuration is missing some percentage of the points and is subject to noise, which is assumed additive and Gaussian. The sample consists of a single point and its M nearest neighbors, where M is of the order 10. If $p \in [0, 1]$ is the percent observed, i.e. $p = 1$ means all points

are observed and $p = 0$ means no points are observed, then the reference point set will be comprised of $N = \lceil M/p \rceil$ points. We write the matrix representation of the noisy data point as $\mathbf{Y} = (Y_1, \dots, Y_M)$, $Y_i \in \mathbb{R}^d$, for $1 \leq i \leq M$.

The observed points have labels, but the reference points do not. We seek to register these noisy and sparse point sets, onto the reference point set. The ultimate goal is to identify the ordering of the labels of the points (types of atoms) in a configuration. We will find the best assignment and rigid transformation between the observed point set and the reference point set. Having completed the registration process for all observations in the configuration, we may then construct a three dimensional distribution of labeled points around each reference point, and the distribution of atomic composition is readily obtained.

The point-set registration problem has two crucial elements. The first is the correspondence, or assignment of each point in the observed set to the reference set. The second is the identification of the optimal transformation from within an appropriate class of transformations. If the transformation class is taken to be the rigid transformations, then each of the individual problems is easily solved by itself, and naive methods simply alternate the solution of each individually until convergence.

One of the most frequently used point set registration algorithms is the iterative closest point method, which alternates between identifying the optimal transformation for a given correspondence, and then corresponding closest points [1]. If the transformation is rigid, then both problems are uniquely solvable. If instead we replace the naive closest point strategy with the assignment problem, so that any two observed points correspond to two different reference points, then again the problem can be solved with a linear program [9]. However, when these two solvable problems are combined into one, the resulting problem is non-convex [14], and no longer admits a unique solution, even for the case of rigid transformations as considered here. The same strategy has been proposed with more general non-rigid transformations [3], where identification of the optimal transformation is no longer analytically solvable. The method in [11] minimizes an upper bound on their objective function, and is thus also susceptible to getting stuck in a local basin of attraction. We instead take a Bayesian formulation of the problem that will simultaneously find the transformation and correspondence between point sets. Most importantly, it is designed to avoid local basins of attraction and locate a global minimum.

We will show how alternating between finding correspondences and minimizing distances can lead to an incorrect registration. Consider now the setup in Fig. 1. If we correspond closest points first, then all three green points would be assigned to the blue ‘1’. Then, identifying the single rigid transformation to minimize the distances

Fig. 1 Setup for incorrect registration; alternating assignment and ℓ^2 minimization



between all three green and the blue ‘1’ would yield a local minimum, with no correct assignments. If we consider instead *assignments*, so that no two observation points can correspond to the same reference point, then again it is easy then to see two equivalent solutions with the eye. The first is a pure translation, and the second can be obtained for example by one of two equivalent rotations around the mid-point between ‘1’s, by π or $-\pi$. The first only gets the assignment of ‘2’ correct, while the second is correct. Note that in reality the reference labels are unknown, so both are equivalent for us. Here it is clear what the solutions are, but once the problem grows in scale, the answer is not always so clear. This simple illustration of degenerate (equal energy) multi-modality of the registration objective function arises from physical symmetry of the reference point-set. This will be an important consideration for our reference point sets, which will arise as a unit cell of a lattice, hence with appropriate symmetry. We will never be able to know the registration beyond these symmetries, but this will nonetheless not be the cause of concern, as symmetric solutions will be considered equivalent. The troublesome multi-modality arises in the presence of noisy and partially observed point sets, where there may be local minima with higher energy than the global minima.

The multi-modality of the combined problem, in addition to the limited information in the noisy and sparse observations, motivates the need for a global probabilistic notion of solution for this problem. It is illustrated in the following subsection that the problem lends itself naturally to a flexible Bayesian formulation which circumvents the intrinsic shortcomings of deterministic optimization approaches for non-convex problems. Indeed at an additional computational cost, we obtain a distribution of solutions, rather than a point estimate, so that general quantities of interest are estimated and uncertainty is quantified. In case a single point estimate is required we define an appropriate optimal one (for example the global energy minimizer or probability maximizer).

2.1 Bayesian Formulation

We seek to compute the registration between the observation set and reference set. We are concerned primarily with rigid transformations of the form

$$\mathcal{T}(X; \theta) = R_\theta X + t_\theta, \quad (1)$$

where $R_\theta \in \mathbb{R}^{d \times d}$ is a rotation and $t_\theta \in \mathbb{R}^d$ is a translation vector.

Write $[\mathbb{T}(X; \theta)]_{ki} = \mathcal{T}_k(X_i)$ for $1 \leq i \leq N$, $1 \leq k \leq d$, and where X_i is the i th column of X . Now let $\xi \in \mathbb{R}^{d \times M}$ with entries $\xi_{ij} \sim N(0, \gamma^2)$, and assume the following statistical model

$$Y = \mathbb{T}(X; \theta)C + \xi, \quad (2)$$

for ξ , θ and C independent.

The matrix of correspondences $C \in \{0, 1\}^{N \times M}$, is such that $\sum_{i=1}^N C_{ij} = 1$, $1 \leq j \leq M$, and each observation point corresponds to only one reference point. So if X_i matches Y_j then $C_{ij} = 1$, otherwise, $C_{ij} = 0$. We let C be endowed with a prior, $\pi_0(C_{ij} = 1) = \pi_{ij}$ for $1 \leq i \leq N$ and $1 \leq j \leq M$. Furthermore, assume a prior on the transformation parameter θ given by $\pi_0(\theta)$. The posterior distribution then takes the form

$$\pi(C, \theta | \mathbf{X}, \mathbf{Y}) \propto \mathcal{L}(\mathbf{Y} | \mathbf{X}, C, \theta) \pi_0(C) \pi_0(\theta), \quad (3)$$

where \mathcal{L} is the likelihood function associated with Eq. (1).

For a given $\tilde{\theta}$, an estimate \hat{C} can be constructed *a posteriori* by letting $\hat{C}_{i^*(j),j} = 1$ for $j = 1, \dots, M$ and zero otherwise, where

$$i^*(j) = \underset{1 \leq i \leq N}{\operatorname{argmin}} |Y_j - \mathcal{T}(X_i; \tilde{\theta})|^2. \quad (4)$$

For example, $\tilde{\theta}$ may be taken as the maximum a posteriori (MAP) estimator or the mean. We note that \hat{C} can be constructed either with a closest point approach, or via assignment to avoid multiple registered points assigned to the same reference.

Lastly, we assume the j th observation only depends on the j th column of the correspondence matrix, and so Y_i, Y_j are conditionally independent with respect to the matrix C for $i \neq j$. This does not exclude the case where multiple observation points are assigned to the same reference point, but as mentioned above such scenario should have zero probability.

To that end, instead of considering the full joint posterior in Eq. (3) we will focus on the marginal of the transformation

$$\pi(\theta | \mathbf{X}, \mathbf{Y}) \propto \mathcal{L}(\mathbf{Y} | \mathbf{X}, \theta) \pi_0(\theta). \quad (5)$$

Let C_j denote the j th column of C . Since C_j is completely determined by the single index i at which it takes the value 1, the marginal likelihood takes the form

$$\begin{aligned} \sum_C p(Y_j | \mathbf{X}, \theta, C) \pi_0(C) &= \sum_{i=1}^N p(Y_j | \mathbf{X}, \theta, C_{ij} = 1) \pi_0(C_{ij} = 1) \\ &= \sum_{i=1}^N \pi_{ij} p(Y_j | \mathbf{X}, \theta, C_{ij} = 1) \\ &\propto \pi_{ij} \exp \left\{ -\frac{1}{2\gamma^2} |Y_j - T(X_i; \theta)|^2 \right\}. \end{aligned} \quad (6)$$

The above marginal together with the conditional independence assumption allows us to construct the likelihood function of the marginal posterior (5) as follows

$$\begin{aligned} \mathcal{L}(\mathbf{Y} \mid \mathbf{X}, \theta) &= \prod_{j=1}^M p(Y_j \mid \mathbf{X}, \theta) \\ &\propto \prod_{j=1}^M \sum_{i=1}^N \pi_{ij} \exp \left\{ -\frac{1}{2\gamma^2} |Y_j - T(X_i; \theta)|^2 \right\}. \end{aligned} \quad (7)$$

Thus the posterior in question is

$$\begin{aligned} \pi(\theta \mid \mathbf{X}, \mathbf{Y}) &\propto \mathcal{L}(\mathbf{Y} \mid \mathbf{X}, \theta) \pi_0(\theta) \\ &= \prod_{j=1}^M \sum_{i=1}^N \pi_{ij} \exp \left\{ -\frac{1}{2\gamma^2} |Y_j - T(X_i; \theta)|^2 \right\} \pi_0(\theta). \end{aligned} \quad (8)$$

Consider a prior on θ such that $\pi_0(\theta) \propto \exp(-\lambda R(\theta))$, where $\lambda > 0$. Then we have the following objective function

$$E(\theta) = - \sum_{j=1}^M \log \sum_{i=1}^N \pi_{ij} \exp \left\{ -\frac{1}{2\gamma^2} |Y_j - T(X_i; \theta)|^2 \right\} + \lambda R(\theta). \quad (9)$$

The minimizer, θ^* , of the above, Eq. (9) is also the maximizer of a posteriori probability under Eq. (8). It is called the maximum a posteriori estimator. This can also be viewed as maximum likelihood estimation regularized by $\lambda R(\theta)$.

By sampling consistently from the posterior, we may estimate quantities of interest, such as moments, together with quantified uncertainty. Additionally, we may recover other point estimators, such as local and global modes.

3 Hamiltonian Monte Carlo

Monte Carlo Markov chain (MCMC) methods are a natural choice for sampling from distributions which can be evaluated pointwise up to a normalizing constant, such as the posterior (8). Furthermore, MCMC comprises the workhorse of Bayesian computation, often appearing as crucial components of more sophisticated sampling algorithms. Formally, an MCMC simulates a distribution μ over a state space Ω by producing an ergodic Markov chain $\{w_k\}_{k \in \mathbb{N}}$ that has μ as its invariant

distribution, i.e.

$$\frac{1}{K} \sum_{k=1}^K g(w_k) \rightarrow \int_{\Omega} g(w) \mu(dw) = \mathbb{E}_{\mu} g(w), \quad (10)$$

with probability 1, for $g \in L^1(\Omega)$.

The Metropolis-Hastings method is a general MCMC method defined by choosing $\theta_0 \in \text{supp}(\pi)$ and iterating the following two steps for $k \geq 0$

- (1) Propose: $\theta^* \sim Q(\theta_k, \cdot)$.
- (2) Accept/reject: Let $\theta_{k+1} = \theta^*$ with probability

$$\alpha(\theta_k, \theta^*) = \min \left\{ 1, \frac{\pi(\theta^*) Q(\theta_k, \theta^*)}{\pi(\theta_k) Q(\theta^*, \theta_k)} \right\},$$

and $\theta_{k+1} = \theta_k$ otherwise.

In general, random-walk proposals Q can result in MCMC chains which are slow to explore the state space and susceptible to getting stuck in local basins of attraction. Hamiltonian Monte Carlo (HMC) is designed to improve this shortcoming. HMC is a Metropolis-Hastings method [4, 13] which incorporates gradient information of the log density with a simulation of Hamiltonian dynamics to efficiently explore the state space and accept large moves of the Markov chain. Heuristically, the gradient yields d pieces of information, for a \mathbb{R}^d -valued variable and scalar objective function, as compared with one piece of information from the objective function alone. Our description here of the HMC algorithm follows that of [2] and the necessary foundations of Hamiltonian dynamics for the method can be found in [17].

Our objective here is to sample from a specific target density

$$\pi(\theta) \propto \exp(-E(\theta)) \quad (11)$$

over θ , where $E(\theta)$ is as defined in Eq. (9) and $\pi(\theta)$ is of the form given by Eq. (8).

First, an artificial momentum variable $p \sim N(0, \Gamma)$, independent of θ , is included into Eq. (11), for a symmetric positive definite mass matrix Γ , that is usually a scalar multiple of the identity matrix. Define a Hamiltonian now by

$$H(p, \theta) = E(\theta) + \frac{1}{2} p^T \Gamma^{-1} p$$

where $E(\theta)$ is the ‘‘potential energy’’ and $\frac{1}{2} p^T \Gamma^{-1} p$ is the ‘‘kinetic energy’’.

Hamilton's equations of motion for $p, \theta \in \mathbb{R}^d$ are, for $i = 1, \dots, d$:

$$\begin{aligned}\frac{d\theta_i}{dt} &= \frac{\partial H}{\partial p_i} \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial \theta_i}\end{aligned}$$

In practice, the algorithm creates a Markov chain on the joint position-momentum space \mathbb{R}^{2d} , by alternating between independently sampling from the marginal Gaussian on momentum p , and numerical integration of Hamiltonian dynamics along an energy contour to update the position. If the initial condition $\theta \sim \pi$ and we were able to perfectly simulate the dynamics, this would give samples from π because the Hamiltonian H remains constant along trajectories. Due to errors in numerical approximation, the value of H will vary. To ensure the samples are indeed drawn from the correct distribution, a Metropolis-Hastings accept/reject step is incorporated into the method.

In particular, after a new momentum is sampled, suppose the chain is in the state (p, θ) . Provided the numerical integrator is reversible, the probability of accepting the proposed point (p^*, θ^*) takes the form

$$\alpha((p, \theta), (p^*, \theta^*)) = \min \{1, \exp \{H(p, \theta) - H(p^*, \theta^*)\}\}. \quad (12)$$

If (p^*, θ^*) is rejected, the next state remains unchanged from the previous iteration. However, note that a fresh momentum variable is drawn each step, so only θ remains fixed. Indeed the momentum variables can be discarded, as they are only auxiliary variables. To be concrete, the algorithm requires an initial state θ_0 , a reversible numerical integrator, integration step-size h , and number of steps L . Note that reversibility of the integrator is crucial such that the proposal integration $Q((p, \theta), (p^*, \theta^*))$ is symmetric and drops out of the acceptance probability in Eq. (12). The parameters h and L are tuning parameters, and are described in detail [2, 13].

The HMC algorithm then proceeds as follows:

for $k \geq 0$ **do** **HMC**:

$p_k \leftarrow \xi$ for $\xi \sim \mathcal{N}(0, \Gamma)$

function INTEGRATOR(p_k, θ_k, h) **return** (p^*, θ^*)

end function

$\alpha \leftarrow \min \{1, \exp \{H(p_k, \theta_k) - H(p^*, \theta^*)\}\}$

$\theta_{k+1} \leftarrow \theta^*$ with probability α **otherwise**

$\theta_{k+1} \leftarrow \theta_k$

end for

Under appropriate assumptions [13], this method will provide samples $\theta_k \sim \pi$, such that for bounded $g : \mathbb{R}^n \rightarrow \mathbb{R}$

$$\frac{1}{K} \sum_{k=1}^K g(\theta_k) \rightarrow \int_{\mathbb{R}^n} g(\theta) d\theta \quad \text{as } K \rightarrow \infty.$$

4 Numerical Experiments

To illustrate our approach, we consider numerical experiments on synthetic datasets in \mathbb{R}^2 and \mathbb{R}^3 , with varying levels of noise and percentage of observed data. We focus our attention to rigid transformations of the form Eq. (1).

For all examples here, the M observation points are simulated as $Y_i \sim N(R_\varphi X_{j(i)} + t, \gamma^2 I_d)$, for a rotation matrix R_φ parameterized by φ , and some t and γ . So, $\theta = (\varphi, t)$. To simulate the unknown correspondence between the reference and observation points, for each $i = 1, \dots, M$, the corresponding index $j(i) \in [1, \dots, N]$ is chosen randomly and without replacement. Recall that we define percentage of observed points here as $p = \frac{M}{N} \in [0, 1]$. We tested various percentages of observed data and noise γ on the observation set, then computed the mean square error (MSE), given by Eq. (13), between the reference points and the registered observed points,

$$\mathcal{E}(\theta) = \frac{1}{M} \sum_{i=1}^M \min_{X \in \mathbf{X}} |R_\varphi^T(Y_i - t) - X|^2. \quad (13)$$

4.1 Two Dimensional Registration

First we consider noise-free data, i.e. $\gamma = 0$ (however in the reconstruction some small $\gamma > 0$ is used). The completed registration for the 2-dimensional ‘fish’ set is shown in Figs. 2 and 3. The ‘fish’ set is a standard benchmark test case for registration algorithms in \mathbb{R}^2 [6, 12]. Our methodology, employing the HMC sampler described in Sect. 3 allows for a correct registration, even in the case where we have only 33% of the initial data, see Fig. 3.

As a final experiment with the ‘fish’ dataset, we took 25 i.i.d. realizations of the reference, all having the same transformation, noise, and percent observed. Since we have formulated the solution of our registration problem as a density, we may compute moments, and find other quantities of interest. In this experiment we evaluate $\bar{\theta} = \frac{1}{25} \sum_{k=1}^{25} \hat{\theta}_k$, where $\hat{\theta}$ is our MAP estimator of θ from the HMC algorithm. We then evaluated the transformation under $\bar{\theta}$. The completed registration

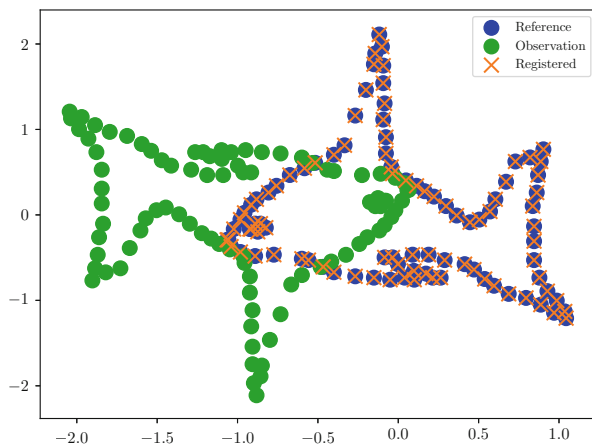


Fig. 2 Full data

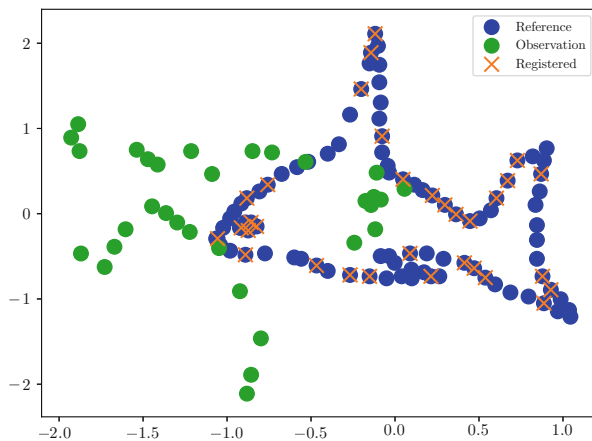


Fig. 3 33% Observed data

is shown in Fig. 4. With a relatively small number of configurations, we are able to accurately reconstruct the data, despite the noisy observations.

4.2 Synthetic APT Data

The datasets from APT experiments are perturbed by additive noise on each of the points. The variance of this additive noise is not known in general, and so in practice it should be taken as a hyper-parameter, endowed with a hyper-prior, and inferred or optimized. It is known that the size of the displacement is on the order of several Å (Angstroms), so that provides a good basis for choice of hyper prior.

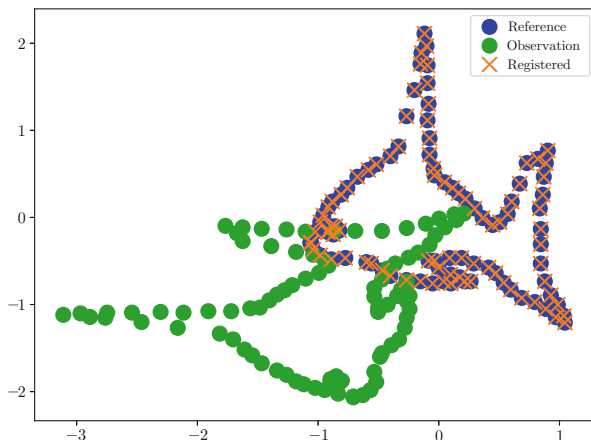


Fig. 4 Full data, $\gamma = 0.5$, average of 25 registrations

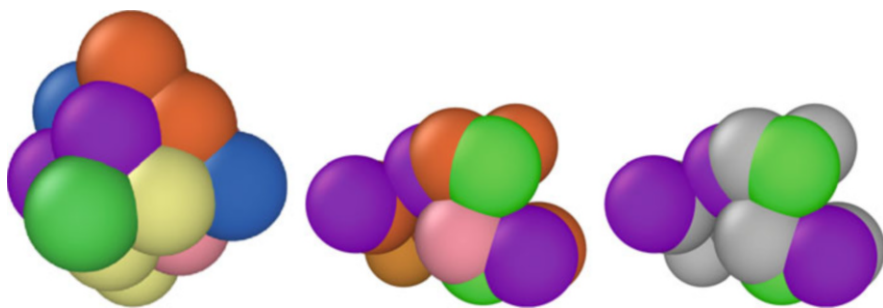


Fig. 5 Example APT data: Left: Hidden truth, Center: Noise added, Right: Missing atoms colored grey

In order to simulate this uncertainty in our experiments, we incorporated additive noise in the form of a truncated Gaussian, to keep all the mass within several \AA . The experiments consider a range of variances in order to measure the impact of noise on our registration process.

In our initial experiments with synthetic data, we have chosen percentages of observed data and additive noise similar to what Materials Scientist experimentalists have reported in their APT datasets. The percent observed of these experimental datasets is approximately 33%. The added noise of these APT datasets is harder to quantify. Empirically, we expect the noise to be Gaussian in form, truncated to be within $1\text{--}3 \text{\AA}$. The standard deviation of the added noise is less well-known, so we will work with different values to assess the method's performance. With respect to the size of the cell, a displacement of 3\AA is significant. Consider the cell representing the hidden truth in Fig. 5. The distance between the front left and right corners is on the scale of 3\AA . Consequently a standard deviation of 0.5 for the additive noise represents a significant displacement of the atoms.

Table 1 $\mathcal{E}(\theta)$ registration errors

Standard deviation	Percent observed	Registration error
0.0	75%	3.49368609883352e-11
0.0	45%	4.40071892313178e-11
0.25	75%	0.1702529649951198
0.25	45%	0.1221555853433331
0.5	75%	0.3445684328735114
0.5	45%	0.3643178111314804

As a visual example, the images in Fig. 5 are our synthetic test data used to simulate the noise and missing data from the APT datasets. The leftmost image in Fig. 5 is the hidden truth we seek to uncover. The middle image is the first with noise added to the atom positions. Lastly, in the right-most image we have ‘ghosted’ some atoms, by coloring them grey, to give a better visual representation of the missing data. In these representations of HEAs, a color different from grey denotes a distinct type of atom. What we seek is to infer the chemical ordering and atomic structure of the left image, from transformed versions of the right, where $\gamma = 0.5$.

For our initial numerical experiments with simulated APT data, we choose a single reference and observation, and consider two different percentages of observed data, 75% and 45%. For both levels of observations in the data, we looked at results with three different levels of added noise on the atomic positions: no noise, and Gaussian noise with standard deviation of 0.25 and 0.5. The MSE of the processes are shown in Table 1. We initially observe the method is able, within an appreciably small tolerance, find the exact parameter θ in the case of no noise, with both percentages of observed data. In the other cases, as expected, the error scales with the noise. This follows from our model, as we are considering a rigid transformation between the observation and reference, which is a volume preserving transformation. If the exact transformation is used with an infinite number of points, then the RMSE (square root of Eq. (13)) is γ .

Now we make the simplifying assumption that the entire configuration corresponds to the same reference, and each observation in the configuration corresponds to the same transformation applied to the reference, with independent, identically distributed (i.i.d.) noise added to it. This enables us to approximate the mean and variance of Eq. (13) over these observation realizations, i.e. we obtain a collection $\{\mathcal{E}^l(\theta^l)\}_{l=1}^L$ of errors, where $\mathcal{E}^l(\theta^l)$ is the MSE corresponding to replacing \mathbf{Y}^l and its estimated registration parameters θ^l into Eq. (13), where L is the total number of completed registrations. The statistics of this collection of values provide robust estimates of the expected error for a single such registration, and the variance we can expect over realizations of the observational noise. In other words

$$\mathbb{E}^L \mathcal{E}(\theta) := \frac{1}{L} \sum_{l=1}^L \mathcal{E}^l(\theta^l) \quad \text{and} \quad \mathbb{V}^L \mathcal{E}(\theta) := \frac{1}{L} \sum_{l=1}^L (\mathcal{E}^l(\theta^l) - \mathbb{E}^L \mathcal{E}(\theta))^2. \quad (14)$$

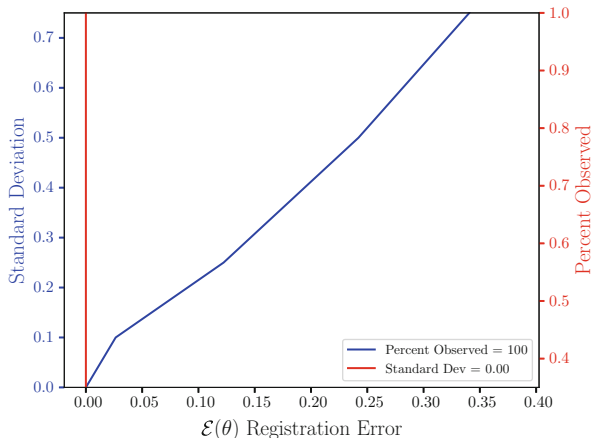


Fig. 6 Blue: Full data, Red: Noiseless data

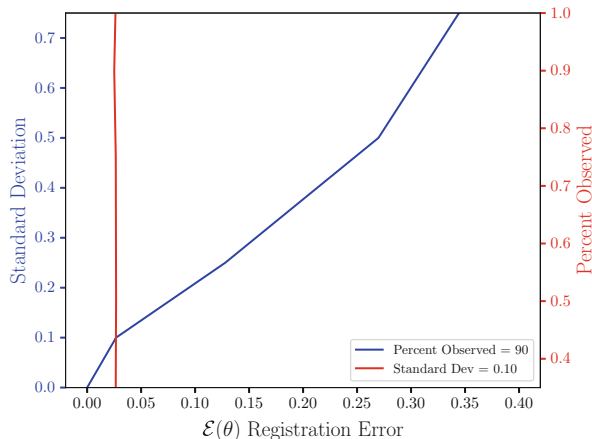


Fig. 7 Blue: 90% Observed, Red: $\gamma = 0.1$

We have confidence intervals as well, corresponding to a central limit theorem approximation based on these L samples.

In Figs. 6, 7, 8, and 9 we computed the registration for $L = 125$ i.i.d. observation sets corresponding to the same reference, for each combination of noise and percent observed data. We then averaged all 125 registration errors for a fixed noise/percent observed combination, as in Eq. (14), and compared the values. What we observe in Figs. 6, 7, 8, and 9 is the registration error scaling with the noise, which is expected. What is interesting to note here is that the registration error is essentially constant with respect to the percentage of observed data, for a fixed standard deviation of the noise. More information will lead to a lower variance in the posterior on the transformation θ , following from standard statistical intuition. However, the

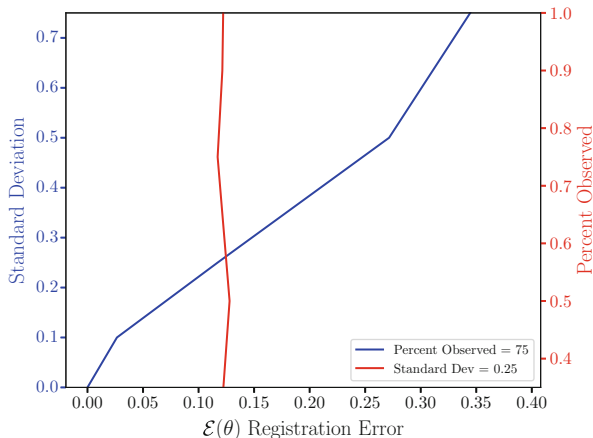


Fig. 8 Blue: 75% Observed, Red: $\gamma = 0.25$

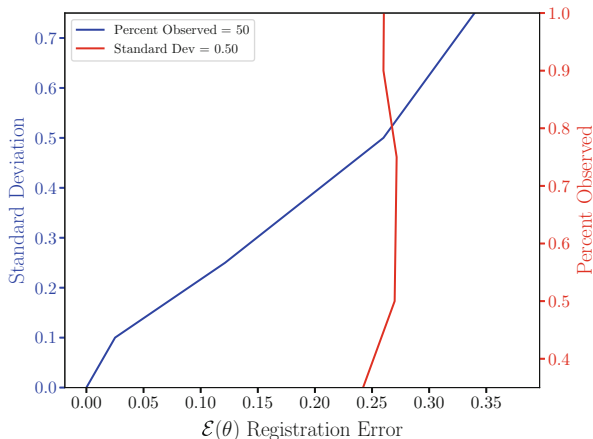


Fig. 9 Blue: 50% Observed, Red: $\gamma = 0.5$

important point to note is that, as mentioned above, for exact transformation, and infinite points, (13) will equal γ^2 . So, for sufficiently accurate transformation, one can expect a sample approximation thereof. Sufficient accuracy is found here with very few observed points, which is reasonable considering that in the zero noise case 2 points is sufficient to fit the 6 parameters exactly.

The MSE registration errors shown in Figs. 6, 7, 8, and 9, show the error remains essentially constant with respect to the percent observed. Consequently, if we consider only Fig. 7, we observe that the blue and red lines intersect, when the blue has a standard deviation of 0.1, and the associated MSE is approximately 0.05. This same error estimate holds for all tested percentages of observed data having a standard deviation of 0.1. Similar results hold for other combinations of noise and percent observed, when the noise is fixed.

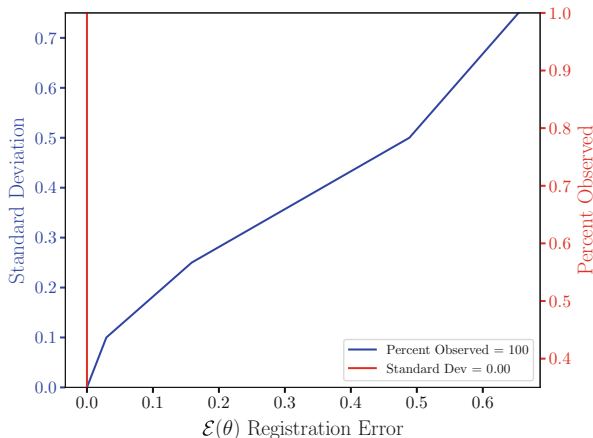


Fig. 10 Blue: Full data, Red: Noiseless data (MALA)

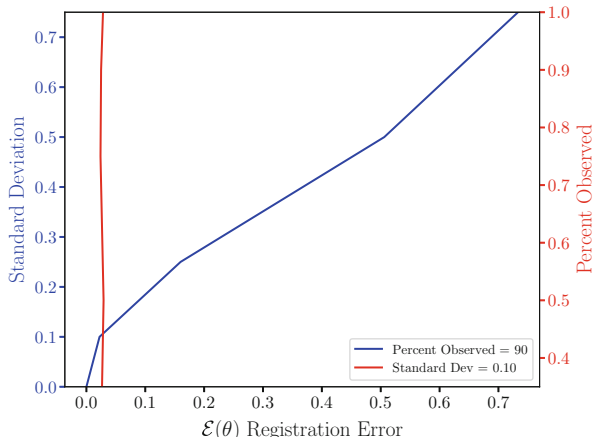


Fig. 11 Blue: 90% Observed, Red: $\gamma = 0.1$ (MALA)

Furthermore, the results shown in Figs. 6, 7, 8, and 9 are independent of the algorithm, as the plots in Figs. 10 and 11 show. For the latter, we ran a similar experiment with 125 i.i.d. observation sets, but to compute the registration, we used the Metropolis Adjusted Langevin Algorithm (MALA) [15], as opposed to HMC in Figs. 6, 7, 8, and 9. Both algorithms solve the same problem and use information from the gradient of the log density. In the plots shown in Figs. 6, 7, 8, and 9, we see the same constant error with respect to the percent observed and the error increasing with the noise, for a fixed percent observed. The MSE also appears to be proportional to γ^2 , which is expected, until some saturation threshold of $\gamma \geq 0.5$ or so. This can be understood as a threshold beyond which the observed points will tend to get assigned to the wrong reference point.

To examine the contours of our posterior described by Eq. (8), we drew 10^5 samples from the density using the HMC methodology described previously. For this simulation we set the noise to have standard deviation of 0.25 and the percent observed was 35%, similar values to what we expect from real APT datasets. The rotation matrix R is constructed via Euler angles denoted: $\varphi_x, \varphi_y, \varphi_z$, where $\varphi_x \in [0, 2\pi)$, $\varphi_y \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and $\varphi_z \in [0, 2\pi)$. These parameters are especially important to making the correct atomic identification, which is crucial to the success of our method.

In Figs. 12, 13, and 14, we present marginal single variable histograms and all combinations of marginal two-variable joint histograms for the individual components of θ . We observe multiple modes in a number of the marginals. In Figs. 15, 16, 17, 18, 19, and 20 we present autocorrelation and trace plots for the rotation parameters from the same instance of the HMC algorithm as presented in the histograms above in Figs. 12, 13, and 14. We focus specifically on the rotation angles, to ensure efficient mixing of the Markov chain as these have thus far been more difficult for the algorithm to optimize. We see the chain is mixing well with respect to these parameters and appears not to become stuck in local basins of attraction.

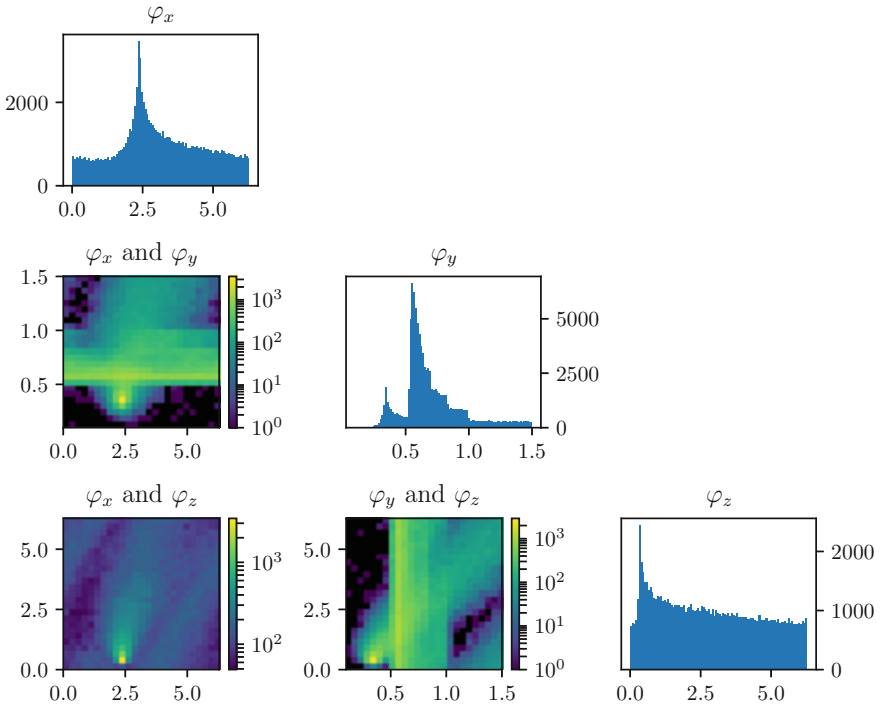


Fig. 12 Histograms of φ parameters, 100,000 samples, $\gamma = 0.25$, Observed = 35%

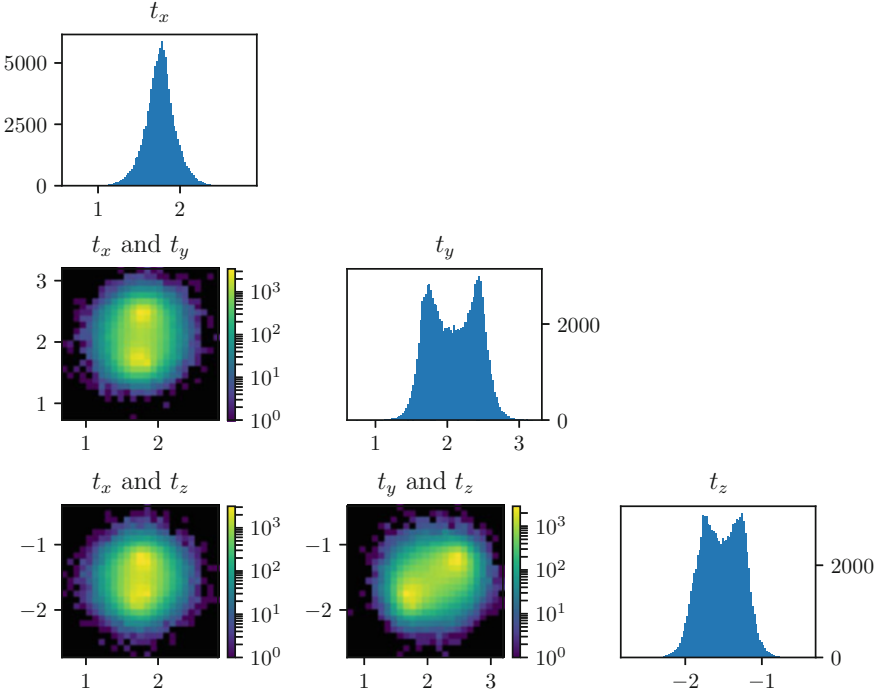


Fig. 13 Histograms of θ parameters, 100,000 samples, $\gamma = 0.25$, Observed = 35%

Additionally, we consider the following. Define null sets A_1, \dots, A_N . For each $j = 1, \dots, M$ and $l = 1, \dots, L$, let $i^*(j, l) := \operatorname{argmin}_{i \in \{1, \dots, N\}} |R_{\phi^l}^T(Y_j^l - t^l) - X_i|^2$, and increment $A_{i^*(j,l)} = A_{i^*(j,l)} \cup Y_j^l$. This provides a distribution of registered points for each index i , A_i , from which we estimate various statistics such as mean and variance. However, note that the cardinality varies between $|A_i| \in \{0, \dots, L\}$. We are only concerned with statistics around reference points i such that $|A_i| > L/10$ or so, assuming that the other reference points correspond to outliers which were registered to by accident. Around each of these $N' \leq N$ reference points X_i , we have a distribution of some $K \leq L$ registered points. We then computed the mean of these K points, denoted by \bar{X}_i and finally we compute the MSE $\frac{1}{N'} \sum_{i=1}^{N'} |X_i - \bar{X}_i|^2$. The RMSE is reported in Table 2. Here we note that a lower percentage observed p is correlated with a larger error. Coupling correct inferences about spatial alignment with an ability to find distributions of atoms around each lattice point is a transformative tool for understanding High Entropy Alloys.

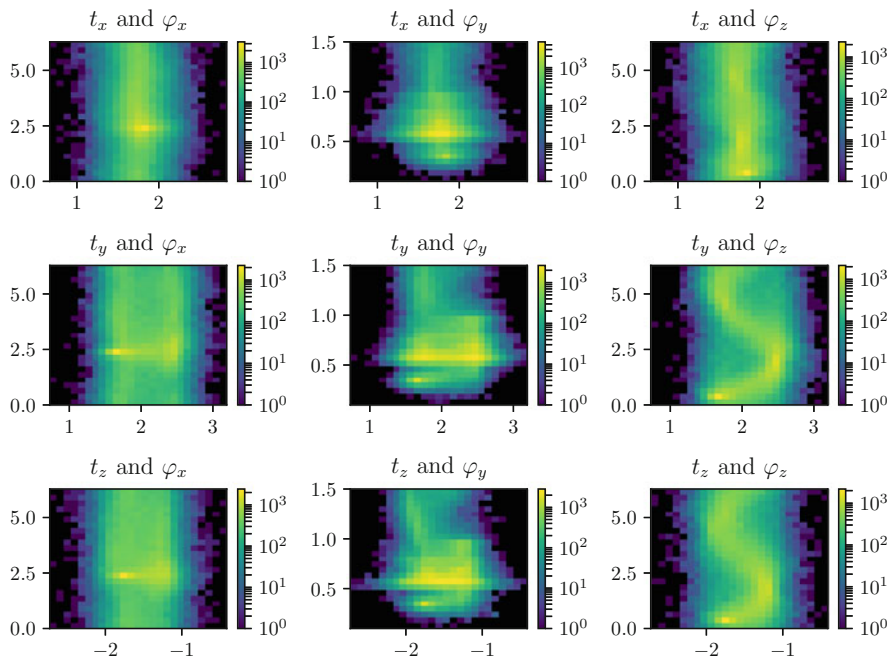


Fig. 14 Histograms of θ parameters, 100,000 samples, $\gamma = 0.25$, Observed = 35%

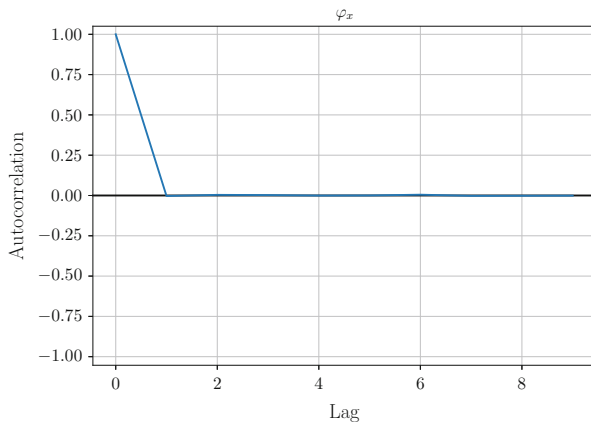


Fig. 15 Autocorrelation plot, φ_x

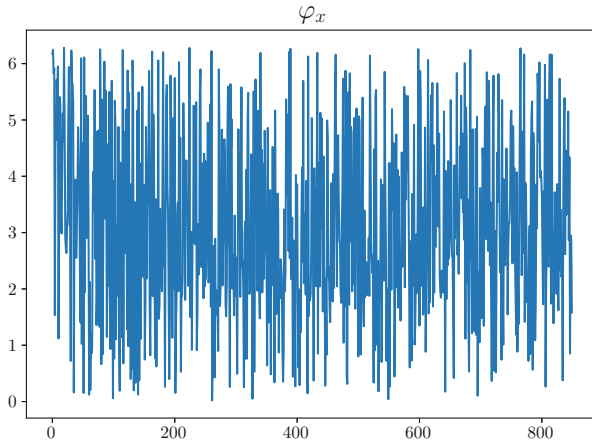


Fig. 16 Trace plot, φ_x

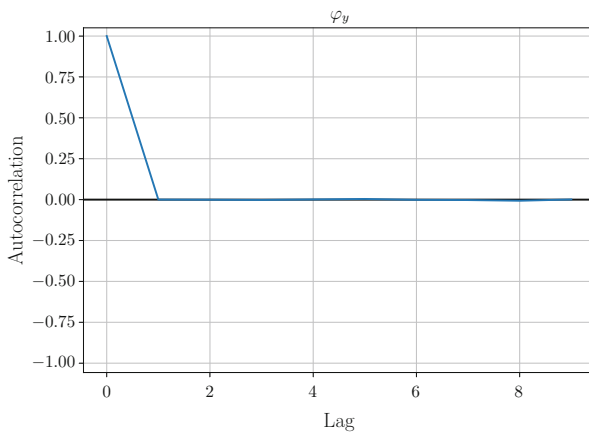


Fig. 17 Autocorrelation plot, φ_y

5 Conclusion

We have presented a statistical model and methodology for point set registration. We are able to recover a good estimate of the correspondence and spatial alignment between point sets in \mathbb{R}^2 and \mathbb{R}^3 despite missing data and added noise. As a continuation of this work, we will extend the Bayesian framework presented in Sect. 2.1 to incorporate the case of an unknown reference. In such a setting, we will seek not only the correct spatial alignment and correspondence, but the reference point set, or crystal structure. The efficiency of our algorithm could be improved through a tempering scheme, allowing for easier transitions between modes, or an

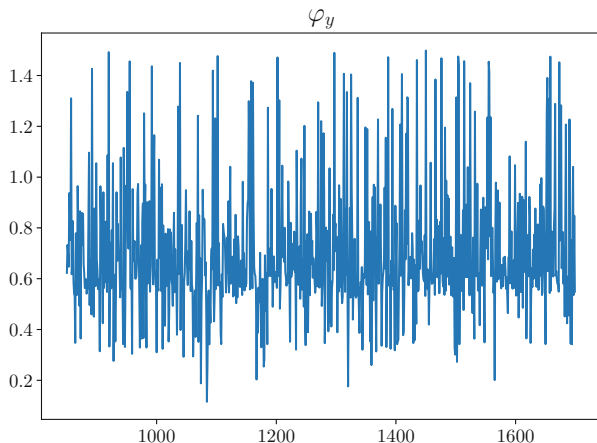


Fig. 18 Trace plot, φ_y

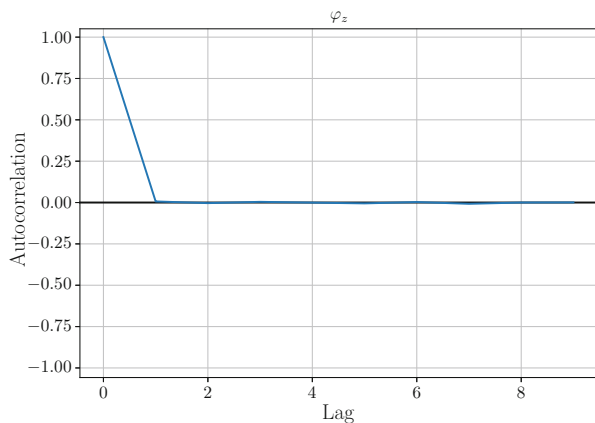


Fig. 19 Autocorrelation plot, φ_z

adaptive HMC scheme, where the chain learns about the sample space in order to make more efficient moves.

Being able to recover the alignment and correspondences with an unknown reference will give Materials Science researchers an unprecedented tool in making accurate predictions about High Entropy Alloys and allow them to develop the necessary tools for classical interaction potentials. Researchers working in the field will be able to determine the atomic level structure and chemical ordering of High Entropy Alloys. From such information, the Material Scientists will have the necessary tools to develop interaction potentials, which is crucial for molecular dynamics simulations and designing these complex materials.

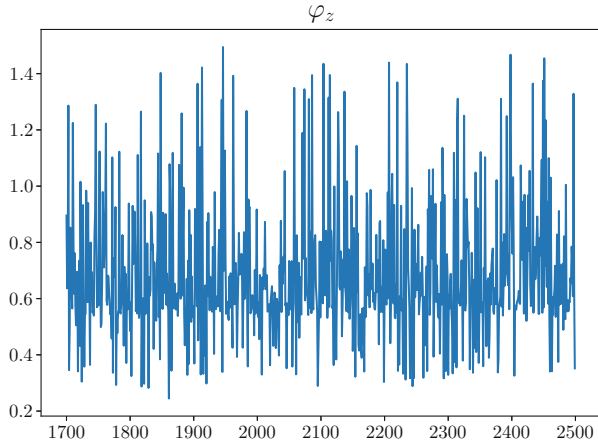


Fig. 20 Trace plot, φ_z

Table 2 Errors for 125 completed registrations

Standard deviation	Percent observed	Error
0.25	75%	0.04909611134835241
0.5	75%	0.07934531875006196
0.25	45%	0.07460005923988245
0.5	45%	0.11978598998930728

Acknowledgements A.S. would like to thank ORISE as well as Oak Ridge National Laboratory (ORNL) Directed Research and Development funding. In addition, he thanks the CAM group at ORNL for their hospitality. K.J.H.L. gratefully acknowledges the support of Oak Ridge National Laboratory Directed Research and Development funding.

References

1. Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **14**(2), 239–256 (1992)
2. Brooks, S., Gelman, A., Jones, G., Meng, X.-L.: *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton (2011)
3. Chui, H., Rangarajan, A.: A new point matching algorithm for non-rigid registration. *Comput. Vis. Image Underst.* **89**(2), 114–141 (2003)
4. Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D.: Hybrid Monte Carlo. *Phys. Lett. B* **195**(2), 216–222 (1987)
5. Gao, M.C., Yeh, J.-W., Liaw, P.K., Zhang, Y.: *High-Entropy Alloys: Fundamentals and Applications*. Springer, Berlin (2016)
6. Jian, B., Vemuri, B.C.: Robust point set registration using gaussian mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(8), 1633–1645 (2011)

7. Jien-Wei, Y.: Recent progress in high entropy alloys. *Ann. Chim. Sci. Mater.* **31**(6), 633–648 (2006)
8. Larson, D.J., Prosa, T.J., Ulfig, R.M., Geiser, B.P., Kelly, T.F.: *Local Electrode Atom Probe Tomography: A User's Guide*. Springer, Berlin (2013)
9. Li, H., Hartley, R.: The 3d-3d registration problem revisited. In: *IEEE 11th International Conference on Computer vision, 2007 (ICCV 2007)*, pp. 1–8 (2007)
10. Miller, M.K., Forbes, R.G.: *Atom-Probe Tomography: The Local Electrode Atom Probe*. Springer, Berlin (2014)
11. Myronenko, A., Song, X.: Point set registration: coherent point drift. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(12), 2262–2275 (2010)
12. Myronenko, A., Song, X., Carreira-Perpinán, M.A., et al.: Non-rigid point set registration: coherent point drift. *Adv. Neural Inf. Proces. Syst.* **19**, 1009 (2007)
13. Neal, R.M.: *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics (Springer-Verlag), vol. 118. Springer, New York (1996)
14. Papazov, C., Burschka, D.: Stochastic global optimization for robust point set registration. *Comput. Vis. Image Underst.* **115**(12), 1598–1609 (2011)
15. Roberts, G.O., Rosenthal, J.S.: Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **60**(1), 255–268 (1998)
16. Santodonato, L.J., Zhang, Y., Feygenson, M., Parish, C.M., Gao, M.C., Weber, R.J.K., Neufeind, J.C., Tang, Z., Liaw, P.K.: Deviation from high-entropy configurations in the atomic distributions of a multi-principal-element alloy. *Nat. Commun.* **6**, 5964 (2015)
17. Teschl, G.: *Ordinary Differential Equations and Dynamical Systems*, vol. 140. American Mathematical Society, Providence (2012)
18. Zhang, Y., Zuo, T.T., Tang, Z., Gao, M.C., Dahmen, K.A., Liaw, P.K., Lu, Z.P.: Microstructures and properties of high-entropy alloys. *Prog. Mater. Sci.* **61**, 1–93 (2014)

Optimization Methods for Inverse Problems



Nan Ye, Farbod Roosta-Khorasani, and Tiangang Cui

Abstract Optimization plays an important role in solving many inverse problems. Indeed, the task of inversion often either involves or is fully cast as a solution of an optimization problem. In this light, the mere non-linear, non-convex, and large-scale nature of many of these inversions gives rise to some very challenging optimization problems. The inverse problem community has long been developing various techniques for solving such optimization tasks. However, other, seemingly disjoint communities, such as that of machine learning, have developed, almost in parallel, interesting alternative methods which might have stayed under the radar of the inverse problem community. In this survey, we aim to change that. In doing so, we first discuss current state-of-the-art optimization methods widely used in inverse problems. We then survey recent related advances in addressing similar challenges in problems faced by the machine learning community, and discuss their potential advantages for solving inverse problems. By highlighting the similarities among the optimization challenges faced by the inverse problem and the machine learning communities, we hope that this survey can serve as a bridge in bringing together these two communities and encourage cross fertilization of ideas.

1 Introduction

Inverse problems arise in many applications in science and engineering. The term “inverse problem” is generally understood as the problem of finding a specific physical property, or properties, of the medium under investigation, using indirect measurements. This is a highly important field of applied mathematics and scientific computing, as to a great extent, it forms the backbone of modern science and

N. Ye · F. Roosta-Khorasani (✉)
University of Queensland, Brisbane, QLD, Australia
e-mail: n.ye@qut.edu.au; fred.roosta@uq.edu.au

T. Cui
School of Mathematical Sciences, Monash University, Clayton, VIC, Australia

engineering. Examples of inverse problems can be found in various fields within medical imaging [6, 7, 12, 71, 95] and several areas of geophysics including mineral and oil exploration [8, 18, 74, 96].

In general, an inverse problem aims at recovering the unknown underlying parameters of a physical system which produces the available observations/measurements. Such problems are generally ill-posed [52]. This is often solved via two approaches: a Bayesian approach which computes a posterior distribution of the models given prior knowledge and the data, or a regularized data fitting approach which chooses an optimal model by minimizing an objective that takes into account both fitness to data and prior knowledge. The Bayesian approach can be used for a variety of downstream inference tasks, such as credible intervals for the parameters; it is generally more computationally expensive than the data fitting approach. The computational attractiveness of data fitting comes at a cost: it can only produce a “point” estimate of the unknown parameters. However, in many applications, such a point estimate can be more than adequate.

In this review, we focus on the data fitting approach. The approach consists of the four building blocks: a parametric model of the underlying physical phenomenon, a forward solver that predicts the observation given the model parameters, an objective function measuring how well a model fits the observation, and an optimization algorithm for finding model parameters optimizing the objective function. The first three components together conceptually defines what an optimal model is, and the optimization algorithm provides a computational means to find the optimal model (usually requires solving the forward problem during optimization). Each of these four building blocks is an active area of research. This paper focuses on the optimization algorithms. While numerous works have been done on the subject, there are still many challenges remaining, including scaling up to large-scale problems, dealing with non-convexity. On the other hand, optimization also constitutes a backbone of machine learning [17, 32]. Consequently, there are many related developments in optimization from the machine learning community. However, thus far and rather independently, the machine learning and the inverse problems communities have largely developed their own sets of tools and algorithms to address their respective optimization challenges. It only stands to reason that many of the recent advances by machine learning can be potentially applicable for addressing challenges in solving inverse problems. We aim to bring out this connection and encourage permeation of ideas across these two communities.

In Sect. 2, we present general formulations for the inverse problem, some typical inverse problems, and optimization algorithms commonly used to solve the data fitting problem. We discuss recent advances in optimization in Sect. 3. We then discuss areas in which cross-fertilization of optimization and inverse problems can be beneficial in Sect. 4. We conclude in Sect. 5. We remark that our review of these recent developments focus on iterative algorithms using gradient and/or Hessian information to update current solution. We do not examine global optimization methods, such as genetic algorithms, simulated annealing, particle swarm optimization, which have also received increasing attention recently (e.g. see [98]).

2 Inverse Problems

An inverse problem can be seen as the reverse process of a forward problem, which concerns with predicting the outcome of some measurements given a complete description of a physical system. Mathematically, a physical system is often specified using a set of model parameters \mathbf{m} whose values completely characterize the system. The model space \mathcal{M} is the set of possible values of \mathbf{m} . While \mathbf{m} usually arises as a function, in practice it is often discretized as a parameter vector for the ease of computation, typically using the finite element method, the finite volume method, or the finite difference method. The forward problem can be denoted as

$$\mathbf{m} \rightarrow \mathbf{d} = \mathbf{f}(\mathbf{m}), \tag{1}$$

where \mathbf{d} are the error-free predictions, and the above notation is a shorthand for $\mathbf{d} = (\mathbf{d}_1, \dots, \mathbf{d}_s) = (\mathbf{f}_1(\mathbf{m}), \dots, \mathbf{f}_s(\mathbf{m}))$, with $\mathbf{d}_i \in \mathbb{R}^l$ being the i -th measurement. The function \mathbf{f} represents the physical theory used for the prediction and is called the forward operator. The observed outcomes contain noises and relate to the system via the following the observation equation

$$\mathbf{d} = \mathbf{f}(\mathbf{m}) + \boldsymbol{\eta}, \tag{2}$$

where $\boldsymbol{\eta}$ are the noises occurred in the measurements. The inverse problem aims to recover the model parameters \mathbf{m} from such noisy measurements.

The inverse problem is almost always ill-posed, because the same measurements can often be predicted by different models. There are two main approaches to deal with this issue. The Bayesian approach assumes a prior distribution $P(\mathbf{m})$ on the model and a conditional distribution $P(\boldsymbol{\eta} \mid \mathbf{m})$ on noise given the model. The latter is equivalent to a conditional distribution $P(\mathbf{d} \mid \mathbf{m})$ on measurements given the model. Given some measurements \mathbf{d} , a posterior distribution $P(\mathbf{m} \mid \mathbf{d})$ on the models is then computed using the Bayes rule

$$P(\mathbf{m} \mid \mathbf{d}) \propto P(\mathbf{m})P(\mathbf{d} \mid \mathbf{m}). \tag{3}$$

Another approach sees the inverse problem as a data fitting problem that finds an parameter vector \mathbf{m} that gives predictions $\mathbf{f}(\mathbf{m})$ that best fit the observed outcomes \mathbf{d} in some sense. This is often cast as an optimization problem

$$\min_{\mathbf{m} \in \mathcal{M}} \psi(\mathbf{m}, \mathbf{d}), \tag{4}$$

where the misfit function ψ measures how well the model \mathbf{m} fits the data \mathbf{d} . When there is a probabilistic model of \mathbf{d} given \mathbf{m} , a typical choice of $\psi(\mathbf{m}, \mathbf{d})$ is the negative log-likelihood. Regularization is often used to address the issue of multiple solutions, and additionally has the benefit of stabilizing the solution, that is, the solution is less likely to change significantly in the presence of outliers

[5, 36, 111]. Regularization incorporates some *a priori* information on \mathbf{m} in the form of a regularizer $R(\mathbf{m})$ and solves the regularized optimization problem

$$\min_{\mathbf{m} \in \mathcal{M}} \psi_{R,\alpha}(\mathbf{m}, \mathbf{d}) := \psi(\mathbf{m}, \mathbf{d}) + \alpha R(\mathbf{m}), \quad (5)$$

where $\alpha > 0$ is a constant that controls the tradeoff between prior knowledge and the fitness to data. The regularizer $R(\mathbf{m})$ encodes a preference over the models, with preferred models having smaller R values. The formulation in Eq. (5) can often be given a *maximum a posteriori (MAP)* interpretation within the Bayesian framework [97]. Implicit regularization also exists in which there is no explicit term $R(\mathbf{m})$ in the objective [53, 54, 86, 87, 107, 109].

The misfit function often has the form $\phi(\mathbf{f}(\mathbf{m}), \mathbf{d})$, which measures the difference between the prediction $\mathbf{f}(\mathbf{m})$ and the observation \mathbf{d} . For example, ϕ may be chosen to be the Euclidean distance between $\mathbf{f}(\mathbf{m})$ and \mathbf{d} . In this case, the regularized problem takes the form

$$\min_{\mathbf{m} \in \mathcal{M}} \phi_{R,\alpha}(\mathbf{m}, \mathbf{d}) := \phi(\mathbf{f}(\mathbf{m}), \mathbf{d}) + \alpha R(\mathbf{m}), \quad (6)$$

This can also be equivalently formulated as choosing the most preferred model satisfying constraints on its predictions

$$\min_{\mathbf{m} \in \mathcal{M}} R(\mathbf{m}), \quad \text{s.t.} \quad \phi(\mathbf{f}(\mathbf{m}), \mathbf{d}) \leq \rho. \quad (7)$$

The constant ρ usually relates to noise and the maximum discrepancy between the measured and the predicted data, and can be more intuitive than α .

2.1 PDE-Constrained Inverse Problems

For many inverse problems in science and engineering, the forward model is not given explicitly via a forward operator $\mathbf{f}(\mathbf{m})$, but often conveniently specified via a set of partial differential equations (PDEs). For such problems, Eq. (6) has the form

$$\min_{\mathbf{m} \in \mathcal{M}, \mathbf{u}} \phi(P \cdot \mathbf{u}, \mathbf{d}) + \alpha R(\mathbf{m}), \quad \text{s.t.} \quad c_i(\mathbf{m}, \mathbf{u}_i) = 0, \quad i = 1, \dots, s, \quad (8)$$

where $P \cdot \mathbf{u} = (P_1, \dots, P_s) \cdot (\mathbf{u}_1, \dots, \mathbf{u}_s) = (P_1 \mathbf{u}_1, \dots, P_s \mathbf{u}_s)$ with \mathbf{u}_i being the field in the i -th experiment, P_i being the projection operator that selects fields at measurement locations in \mathbf{d}_i (that is, $P_i \mathbf{u}_i$ are the predicted values at locations measured in \mathbf{d}_i), and $c_i(\mathbf{m}, \mathbf{u}_i) = 0$ corresponds to the forward model in the i -th experiment. In practice, the forward model can often be written as

$$\mathcal{L}_i(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i, \quad i = 1, \dots, s, \quad (9)$$

where $\mathcal{L}_i(\mathbf{m})$ is a differential operator, and \mathbf{q}_i is a term that incorporates source terms and boundary values.

The fields $\mathbf{u}_1, \dots, \mathbf{u}_s$ in Eqs. (8) and (9) are generally functions in two or three dimensional spaces, and finding closed-form solutions is usually not possible. Instead, the PDE-constrained inverse problem is often solved numerically by discretizing Eqs. (8) and (9) using the finite element method, the finite volume method, or the finite difference method. Often the discretized PDE-constrained inverse problem takes the form

$$\min_{\mathbf{m} \in \mathcal{M}, \mathbf{u}} \phi(P\mathbf{u}, \mathbf{d}) + \alpha R(\mathbf{m}), \quad \text{s.t.} \quad L_i(\mathbf{m})\mathbf{u}_i = \mathbf{q}_i, \quad i = 1, \dots, s, \quad (10)$$

where P is a block-diagonal matrix consisting of diagonal blocks P_1, \dots, P_s representing the discretized projection operators, \mathbf{u} is the concatenation of the vectors $\mathbf{u}_1, \dots, \mathbf{u}_s$ representing the discretized fields, and each $L_i(\mathbf{m})$ is a square, non-singular matrix representing the differential operator $\mathcal{L}_i(\mathbf{m})$. Each $L_i(\mathbf{m})$ is typically large and sparse. We abuse the notations P, \mathbf{u} to represent both functions and their discretized versions, but the meanings of these notations will be clear from context.

The constrained problem in Eq. (10) can be written in an unconstrained form by eliminating \mathbf{u} using $\mathbf{u}_i = L_i^{-1}\mathbf{q}_i$,

$$\min_{\mathbf{m} \in \mathcal{M}} \phi(PL^{-1}(\mathbf{m})\mathbf{q}, \mathbf{d}) + \alpha R(\mathbf{m}), \quad (11)$$

where L is the block-diagonal matrix with L_1, \dots, L_s as the diagonal blocks, and \mathbf{q} is the concatenation of $\mathbf{q}_1, \dots, \mathbf{q}_s$. Note that, as in the case of (6), here we have $\mathbf{f}(\mathbf{m}) = PL^{-1}(\mathbf{m})\mathbf{q}$.

Both the constrained and unconstrained formulations are used in practice. The constrained formulation can be solved using the method of Lagrangian multipliers. This does not require explicitly solving the forward problem as in the unconstrained formulation. However, the problem size increases, and the problem becomes one of finding a saddle point of the Lagrangian, instead of finding a minimum as in the constrained formulation.

2.2 Image Reconstruction

Image reconstruction studies the creation of 2-D and 3-D images from sets of 1-D projections. The 1-D projections are generally line integrals of a function representing the image to be reconstructed. In the 2-D case, given an image function $f(x, y)$, the integral along the line at a distance of s away from the origin and having a normal which forms an angle ϕ with the x -axis is given by the Radon transform

$$p(s, \phi) = \int_{-\infty}^{\infty} f(z \sin \phi + s \cos \phi, -z \cos \phi + s \sin \phi) dz. \quad (12)$$

Reconstruction is often done via back projection, filtered back projection, or iterative methods [55, 77]. Back projection is the simplest but often results in a blurred reconstruction. Filtered back projection (FBP) is the analytical inversion of the Radon transform and generally yields reconstructions of much better quality than back projection. However, FBP may be infeasible in the presence of discontinuities or noise. Iterative methods take noise into account, by assuming a distribution for the noise. The objective function is often chosen to be a regularized likelihood of the observation, which is then iteratively optimized using the expectation maximization (EM) algorithm.

2.3 Objective Function

One of the most commonly used objective function is the least squares criterion, which uses a quadratic loss and a quadratic regularizer. Assume that the noise for each experiment in (2) is independently but normally distributed, i.e., $\boldsymbol{\eta}_i \sim \mathcal{N}(0, \boldsymbol{\Sigma}_i), \forall i$, where $\boldsymbol{\Sigma}_i \in \mathbb{R}^{l \times l}$ is the covariance matrix. Let $\boldsymbol{\Sigma}$ be the block-diagonal matrix with $\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_s$ as the diagonal blocks. The standard *maximum likelihood* (ML) approach [97], leads to minimizing the least squares (LS) misfit function

$$\phi(\mathbf{m}) := \|\mathbf{f}(\mathbf{m}) - \mathbf{d}\|_{\boldsymbol{\Sigma}^{-1}}^2, \quad (13)$$

where the norm $\|x\|_A = \sqrt{x^\top A x}$ is a generalization of the Euclidean norm (assuming the matrix A is positive definite, which is true in the case of $\boldsymbol{\Sigma}_i^{-1}$). In the above equation, we simply write the general misfit function $\phi(\mathbf{f}(\mathbf{m}), \mathbf{d})$ as $\phi(\mathbf{m})$ by taking the measurements \mathbf{d} as fixed and omitting it from the notation. As previously discussed, we often minimize a regularized misfit function

$$\phi_{R,\alpha}(\mathbf{m}) := \phi(\mathbf{m}) + \alpha R(\mathbf{m}). \quad (14)$$

The prior $R(\mathbf{m})$ is often chosen as a Gaussian regularizer $R(\mathbf{m}) = (\mathbf{m} - \mathbf{m}_{\text{prior}})^\top \boldsymbol{\Sigma}_m^{-1} (\mathbf{m} - \mathbf{m}_{\text{prior}})$. We can also write the above optimization problem as minimizing $R(\mathbf{m})$ under the constraints

$$\sum_{i=1}^s \|\mathbf{f}_i(\mathbf{m}) - \mathbf{d}_i\| \leq \rho. \quad (15)$$

The least-squares criterion belongs to the class of ℓ_p -norm criteria, which contain two other commonly used criteria: the least-absolute-values criterion and the minimax criterion [104]. These correspond to the use of the ℓ_1 -norm and the ℓ_∞ -norm for the misfit function, while the least squares criterion uses the ℓ_2 -norm. Specifically, the least-absolute-values criterion takes $\phi(\mathbf{m}) := \|\mathbf{f}(\mathbf{m}) - \mathbf{d}\|_1$, and the

minimax criterion takes $\phi(\mathbf{m}) := \|\mathbf{f}(\mathbf{m}) - \mathbf{d}\|_\infty$. More generally, each coordinate in the difference may be weighted. The ℓ_1 solution is more robust (that is, less sensitive to outliers) than the ℓ_2 solution, which is in turn more robust than the ℓ_∞ solution [25]. The ℓ_∞ norm is desirable when outliers are uncommon but the data are corrupted by uniform noise such as the quantization errors [26].

Besides the ℓ_2 regularizer discussed above, the ℓ_1 -norm is often used too. The ℓ_1 regularizer induces sparsity in the model parameters, that is, heavier ℓ_1 regularization leads to fewer non-zero model parameters.

2.4 Optimization Algorithms

Various optimization techniques can be used to solve the regularized data fitting problem. We focus on iterative algorithms for nonlinear optimization below as the objective functions are generally nonlinear. In some cases, the optimization problem can be transformed to a linear program. For example, linear programming can be used to solve the least-absolute-values criterion or the minimax criterion. However, linear programming are considered to have no advantage over gradient-based methods (see Section 4.4.2 in [104]), and thus we do not discuss such methods here. Nevertheless, there are still many optimization algorithms that can be covered here, and we refer the readers to [13, 80].

For simplicity of presentation, we consider the problem of minimizing a function $g(\mathbf{m})$. We consider iterative algorithms which start with an iterate \mathbf{m}_0 , and compute new iterates using

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \lambda_k p_k, \quad (16)$$

where p_k is a search direction, and λ_k a step size. Unless otherwise stated, we focus on unconstrained optimization. These algorithms can be used to directly solve the inverse problem in Eq. (5). We only present a selected subset of the algorithms available and have to omit many other interesting algorithms.

Newton-Type Methods The classical Newton's method starts with an initial iterate \mathbf{m}_0 , and computes new iterates using

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \left(\nabla^2 g(\mathbf{m}_k)\right)^{-1} \nabla g(\mathbf{m}_k), \quad (17)$$

that is, the search direction is $p_k = -\left(\nabla^2 g(\mathbf{m}_k)\right)^{-1} \nabla g(\mathbf{m}_k)$, and the step length is $\lambda_k = 1$. The basic Newton's method has quadratic local convergence rate at a small neighborhood of a local minimum. However, computing the search direction p_k can be very expensive, and thus many variants have been developed. In addition, in non-convex problems, classical Newton direction might not exist (if the Hessian

matrix is not invertible) or it might not be an appropriate direction for descent (if the Hessian matrix is not positive definite).

For non-linear least squares problems, where the objective function $g(\mathbf{m})$ is a sum of squares of nonlinear functions, the Gauss-Newton (GN) method is often used [101]. Extensions to more general objective functions as in Eq. (13) with covariance matrix Σ and arbitrary regularization as in Eq. (14) is considered in [94]. Without loss of generality, assume $g(\mathbf{m}) = \sum_{i=1}^s (\mathbf{f}_i(\mathbf{m}) - \mathbf{d}_i)^2$. At iteration k , the GN search direction p_k is given by

$$\left(\sum_{i=1}^s J_i^\top J_i \right) p_k = -\nabla g, \quad (18)$$

where the sensitivity matrix J_i and the gradient ∇g are given by

$$J_i = \frac{\partial \mathbf{f}_i}{\partial \mathbf{m}}(\mathbf{m}_k), \quad i = 1, \dots, s, \quad (19)$$

$$\nabla g = 2 \sum_{i=1}^s J_i^\top (\mathbf{f}_i(\mathbf{m}_k) - \mathbf{d}_i), \quad (20)$$

The Gauss-Newton method can be seen as an approximation of the basic Newton's method obtained by replacing $\nabla^2 g$ by $\sum_{i=1}^s J_i^\top J_i$. The step length $\lambda_k \in [0, 1]$ can be determined by a weak line search [80] (using, say, the Armijo algorithm starting with $\lambda_k = 1$) ensuring sufficient decrease in $g(\mathbf{m}_{k+1})$ as compared to $g(\mathbf{m}_k)$.

Often several nontrivial modifications are required to adapt this prototype method for different applications, e.g., dynamic regularization [53, 86, 87, 108] and more general *stabilized GN* studied [30, 93]. This method replaces the solution of the linear systems defining p_k by r preconditioned conjugate gradient (PCG) inner iterations, which costs $2r$ solutions of the forward problem per iteration, for a moderate integer value r . Thus, if K outer iterations are required to obtain an acceptable solution then the total work estimate (in terms of the number of PDE solves) is approximated *from below* by $2(r + 1)Ks$.

Though Gauss-Newton is arguable the method of choice within the inverse problem community, other Newton-type methods exist which have been designed to suitably deal with the non-convex nature of the underlying optimization problem include Trust Region [27, 114] and the Cubic Regularization [23, 114]. These methods have recently found applications in machine learning [115]. Studying the advantages/disadvantages of these non-convex methods for solving inverse problems can be indeed a useful undertaking.

Quasi-Newton Methods An alternative method to the above Newton-type methods is the quasi-Newton variants including the celebrated limited memory BFGS (L-BFGS) [68, 79]. BFGS iteration is closely related to conjugate gradient (CG) iteration. In particular, BFGS applied to a strongly convex quadratic objective, with exact line search as well as initial Hessian P , is equivalent to preconditioned CG

with preconditioner P . However, as the objective function departs from being a simple quadratic, the number of iterations of L-BFGS could be significantly higher than that of GN or trust region. In addition, it has been shown that the performance of BFGS and its limited memory version is greatly negatively affected by the high degree of ill-conditioning present in such problems [90, 91, 113]. These two factors are among the main reasons why BFGS (and L-BFGS) can be less effective compared with other Newton-type alternatives in many inversion applications [44].

Krylov Subspace Method A Krylov subspace method iteratively finds the optimal solution to an optimization in a larger subspace by making use of the previous solution in a smaller subspace. One of the most commonly used Krylov subspace methods is the conjugate gradient (CG) method. CG was originally designed to solve convex quadratic minimization problems of the form $g(\mathbf{m}) = \frac{1}{2} \mathbf{m}^\top A \mathbf{m} - b^\top \mathbf{m}$. Equivalently, this solves the positive definite linear system $A \mathbf{m} = b$. It computes a sequence of iterates $\mathbf{m}_0, \mathbf{m}_1, \dots$ converging to the minimum through the following two sets of equations.

$$\mathbf{m}_0 = 0, \quad r_0 = b, \quad p_0 = r_0, \quad (21)$$

$$\mathbf{m}_{k+1} = \mathbf{m}_k + \frac{\|r_k\|_2^2}{p_k^\top A p_k} p_k, \quad r_{k+1} = r_k - \frac{\|r_k\|_2^2}{p_k^\top A p_k} A p_k, \quad p_{k+1} = r_{k+1} + \frac{\|r_{k+1}\|_2^2}{\|r_k\|_2^2} p_k, \quad k \geq 0. \quad (22)$$

This can be used to solve the forward problem of the form $L_i(\mathbf{m}) \mathbf{u}_i = \mathbf{q}_i$, provided that $L_i(\mathbf{m})$ is positive definite, which is true in many cases.

CG can be used to solve the linear system for the basic Newton direction. However, the Hessian is not necessarily positive definite and modification is needed [80].

In general, CG can be generalized to minimize a nonlinear function $g(\mathbf{m})$ [28, 39]. It starts with an arbitrary \mathbf{m}_0 , and $p_1 = -\nabla g(\mathbf{m}_0)$, and computes a sequence of iterates $\mathbf{m}_1, \mathbf{m}_2, \dots$ using the equations below: for $k \geq 0$,

$$\mathbf{m}_{k+1} = \arg \min_{\mathbf{m} \in \{\mathbf{m}_k + \lambda p_k, \lambda \in \mathbb{R}\}} g(\mathbf{m}), \quad (23)$$

$$p_{k+1} = -\nabla g(\mathbf{m}_{k+1}) + \beta_k p_k, \quad \text{where } \beta_k = \frac{\|\nabla g(\mathbf{m}_{k+1})\|_2^2}{\|\nabla g(\mathbf{m}_k)\|_2^2}. \quad (24)$$

The above formula for β_k is known as the Fletcher-Reeves formula. Other choices of β_k exist. The following two formulas are known as the Polak-Ribiere and Hestenes-Stiefel formula respectively.

$$\beta_k = \frac{\langle \nabla g(\mathbf{m}_{k+1}) - \nabla g(\mathbf{m}_k), \nabla g(\mathbf{m}_{k+1}) \rangle}{\|\nabla g(\mathbf{m}_k)\|_2^2}, \quad (25)$$

$$\beta_k = \frac{\langle \nabla g(\mathbf{m}_{k+1}) - \nabla g(\mathbf{m}_k), \nabla g(\mathbf{m}_{k+1}) \rangle}{p_k^\top (\nabla g(\mathbf{m}_{k+1}) - \nabla g(\mathbf{m}_k))}. \quad (26)$$

In practice, nonlinear CG does not seem to work well, and is mainly used together with other methods, such as in the Newton CG method [80].

Lagrangian Method of Multipliers The above discussion focuses on unconstrained optimization algorithms, which are suitable for unconstrained formulations of inverse problems, or unconstrained auxiliary optimization problems in methods which solve the constrained formulations directly. The Lagrangian method of multipliers is often used to directly solve the constrained version. Algorithms have been developed to offset the heavier computational cost and slow convergence rates of standard algorithms observed on the Lagrangian, which is a larger problem than the constrained problem. For example, such algorithm may reduce the problem to a smaller one, such as working with the reduced Hessian of the Lagrangian [47], or preconditioning [10, 45]. These methods have shown some success in certain PDE-constrained optimization problems.

Augmented Lagrangian methods have also been developed (e.g. [1, 57]). Such method constructs a series of penalized Lagrangians with vanishing penalty, and finds an optimizer of the Lagrangian by successively optimizing the penalized Lagrangians.

2.5 Challenges

Scaling up to Large Problems The discretized version of an inverse problem is usually of very large scale, and working with fine resolution or discretized problems in high dimension is still an active area of research.

Another challenge is to scale up to large number of measurements, which is widely believed to be helpful for quality reconstruction of the model in practice, with some theoretical support. While recent technological advances makes many big datasets available, existing algorithms cannot efficiently cope with such datasets. Examples of such problems include electromagnetic data inversion in mining exploration [33, 48, 78, 81], seismic data inversion in oil exploration [38, 56, 88], diffuse optical tomography (DOT) [6, 14], quantitative photo-acoustic tomography (QPAT) [42, 117], direct current (DC) resistivity [30, 49, 50, 83, 100], and electrical impedance tomography (EIT) [16, 24, 110].

It has been suggested that many well-placed experiments yield practical advantage in order to obtain reconstructions of acceptable quality. For the special case where the measurement locations as well as the discretization matrices do not change from one experiment to another, various approximation techniques have been proposed to reduce the effective number of measurements, which in turn implies a smaller scale optimization problem, under the unifying category of “simultaneous sources inversion” [46, 63, 89, 93, 94]. Under certain circumstances, even if the P_i 's are different across experiments (but L_i 's are fixed), there are methods to transform the existing data set into the one where all sources share the same receivers, [92].

Dealing with Non-convexity Another major source of difficulty in solving many inverse problems, is the high-degree of non-linearity and non-convexity in (1). This is most often encountered in problems involving PDE-constrained optimization where each \mathbf{f}_i corresponds to the solution of a PDE. Even if the output of the PDE model itself, i.e., the “right-hand side”, is linear in the sought-after parameter, the solution of the PDE, i.e., the forward problem, shows a great deal of non-linearity. This coupled with a great amount of non-convexity can have significant consequences in the quality of inversion and the obtained parameter. Indeed, in presence of non-convexity, the large-scale computational challenges are exacerbated, multiple folds over, by the difficulty of avoiding (possibly degenerate) *saddle-points* as well as finding (at least) a *local minimum*.

Dealing with Discontinuity While the parameter function of the model is often smooth, the parameter function can be discontinuous in some cases. Such discontinuities arise very naturally as a result of the physical properties of the underlying physical system, e.g., EIT and DC resistivity, and require non-trivial modifications to optimization algorithms, e.g., [30, 93]. Ignoring such discontinuities can lead to unsatisfactory recovery results [30, 31, 103]. The level set method [82] is often used to model discontinuous parameter function. This reparametrizes the discontinuous parameter function as a differentiable one, and thus enabling more stable optimization [31].

3 Recent Advances in Optimization

Recent successes in using machine learning to deal with challenging perception and natural language understanding problems have spurred many advances in the study of optimization algorithms as optimization is a building block in machine learning. These new developments include efficient methods for large-scale optimization, methods designed to handle non-convex problems, methods incorporating the structural constraints, and finally the revival of second-order methods. While these developments address a different set of applications in machine learning, they address similar issues as encountered in inverse optimization and could be useful. We highlight some of the works below. We keep the discussion brief because numerous works have been done behind these developments and an indepth and comprehensive discussion is beyond the scope of this review. Our objective is thus to delineate the general trends and ideas, and provide references for interested readers to dig on relevant topics.

Stochastic Optimization The development in large-scale optimization methods is driven by the availability of many large datasets, which are made possible by the rapid development and extensive use of computers and information technology. In machine learning, a model is generally built by optimizing a sum of misfit on the examples. This finite-sum structure naturally invites the application of stochastic optimization algorithms. This is mainly due to the fact that stochastic algorithms

recover the sought-after models more efficiently by employing small batches of data in each iteration, as opposed to the whole data-set. The most well-known stochastic gradient based algorithm is the stochastic gradient descent (SGD). To minimize a finite-sum objective function

$$g(\mathbf{m}) = \frac{1}{n} \sum_{i=1}^n g_i(\mathbf{m}), \quad (27)$$

in the big data regime where $n \gg 1$, the vanilla SGD performs an update

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \lambda_k \nabla g_{i_k}(\mathbf{m}_k), \quad (28)$$

where i_k is randomly sampled from $1, \dots, n$. As compared to gradient descent, SGD replaces the full gradient $\nabla g(\mathbf{m})$ by a stochastic gradient $g_{i_k}(\mathbf{m}_k)$ with its expectation being the full gradient. The batch version of SGD constructs a stochastic gradient by taking the average of several stochastic gradients.

Vanilla SGD is inexpensive per iteration, but suffers from a slow rate of convergence. For example, while full gradient descent achieves a linear convergence rate for smooth strongly convex problems, SGD only converges at a sublinear rate. The slow convergence rate can be partly accounted by the variance in the stochastic gradient. Recently, variance reduction techniques have been developed, e.g. SVRG [61] and SDCA [99]. Perhaps surprisingly, such variants can achieve linear convergence rates on convex smooth problems as full gradient descent does, instead of sublinear rates achieved by the vanilla SGD. There are also a number of variants with no known linear rates but have fast convergence rates for non-convex problems in practice, e.g., AdaGrad [34], RMSProp [105], ESGD [29], Adam [62], and Adadelta [118]. Indeed, besides efficiency, stochastic optimization algorithms also seem to be able to cope with the nonconvex objective functions well, and play a key role in the revival of neural networks as deep learning [43, 60, 66].

Recently, it has also been shown that SGD can be used as a variational algorithm for computing the posterior distribution of parameters given observations [72]. This can be useful in the Bayesian approach for solving inverse problems.

Nonconvex Optimization There is also an increasing interest in non-convex optimization in the machine learning community recently. Nonconvex objectives not only naturally occur in deep learning, but also occur in problems such as tensor decomposition, variable selection, low-rank matrix completion, e.g. see [43, 59, 73] and references therein.

As discussed above, stochastic algorithms have been found to be capable of effectively escaping local minima. There are also a number of studies which adapt well-known acceleration techniques for convex optimization to accelerate the convergence rates of both stochastic and non-stochastic optimization algorithms for nonconvex problems, e.g., [4, 67, 85, 102].

Dealing with Structural Constraints Many problems in machine learning come with complex structural constraints. The Frank-Wolfe algorithm (a.k.a. conditional gradient) [41] is an algorithm for optimizing over a convex domain. It has gained a revived interest due to its ability to deal with many structural constraints efficiently. It requires solving a linear minimization problem over the feasible set, instead of a quadratic program as in the case of proximal gradient algorithms or projected gradient descent. Domains suitable for the Frank-Wolfe algorithm include simplices, ℓ_p -balls, matrix nuclear norm ball, matrix operator norm ball [58].

The Frank-Wolfe algorithm belongs to the class of linear-optimization-based algorithms [64, 65]. These algorithms share with the Frank-Wolfe algorithm the characteristic of requiring a first-order oracle for gradient computation and an oracle for solving a linear optimization problem over the constraint set.

Second-Order Methods The great appeal of the second-order methods lies mainly in the observed empirical performance as well as some very appealing theoretical properties. For example, it has been shown that stochastic Newton-type methods in general, and Gauss-Newton in particular, can not only be made scalable and have low per-iteration cost [30, 47, 51, 92–94], but more importantly, and unlike first-order methods, are very *resilient* to many adversarial effects such as *ill-conditioning* [90, 91, 113]. As a result, for moderately to very ill-conditioned problems, commonly found in scientific computing, while first-order methods make effectively no progress at all, second-order counterparts are not affected by the degree of ill-conditioning. A more subtle, yet potentially more severe draw-back in using first-order methods, is that their success is tightly intertwined with *fine-tuning* (often many) *hyper-parameters*, most importantly, the step-size [11]. In fact, it is highly unlikely that many of these methods exhibit acceptable performance on first try, and it often takes many trials and errors before one can see reasonable results. In contrast, second-order optimization algorithms involve much less parameter tuning and are less sensitive to the choice of hyper-parameters [11, 115].

Since for the finite-sum problem (27) with $n \gg 1$, the operations with the Hessian/gradient constitute major computational bottlenecks, a rather more recent line of research is to construct the inexact Hessian information using the application of *randomized methods*. Specifically, for convex optimization, the stochastic approximation of the full Hessian matrix in the classical Newton's method has been recently considered in [3, 11, 15, 19, 20, 35, 37, 75, 76, 84, 90, 91, 112, 113, 116]. In addition to inexact Hessian, a few of these methods study the fully stochastic case in which the gradient is also approximated, e.g., [15, 90, 91]. For non-convex problems, however, the literature on methods that employ randomized Hessian approximation is significantly less developed than that of convex problems. A few recent examples include the stochastic trust region [114], stochastic cubic regularization [106, 114], and noisy negative curvature method [69]. Empirical performance of many of these methods for some non-convex machine learning applications has been considered in [115].

3.1 A Concrete Success Story

The development of optimization methods in the machine learning community has been fueled by the need to obtain better generalization performance on future “unseen” data. This is in contrast with typical inverse problem applications where fitting the model to the observations on hand make up of all that matters. These rather strikingly different goals have led the ML community to develop optimization methods that can address ML specific challenges. This, in part, has given rise to scalable algorithms that can often deliver far beyond what the most widely used optimization methods in the inverse problem community can.

As a concrete example, consider L-BFGS and Gauss-Newton, which are, arguably, among the most popular optimization techniques used by the scientific computing community in a variety of inverse problem applications. In fact, unlike Gauss-Newton method, L-BFGS, due to its low per-iteration costs, has found significant attraction within the machine learning community as well. Nevertheless, due to the resurgence of non-convex deep learning problems in ML, there is an increasing demand for scalable optimization algorithms that can avoid saddle points and converge to a local minimum. This demand has driven the development algorithms that can surpass the performance of L-BFGS and Gauss-Newton when applied to deep learning applications, e.g., [115].

These results are not unexpected. Indeed, contrary to popular belief, BFGS is not quite a “full-fledged” second-order method as it merely employs first-order information, i.e. gradients, to approximate the curvature. Similar in spirit, Gauss-Newton also does not fully utilize the Hessian information. In particular, in exchange for obtaining a positive definite approximation matrix, GN completely ignores the information from *negative curvature*, which is critical for allowing to escape from regions with small gradient. Escaping saddle points and converging to a local minimum with lower objective values have surprisingly not been a huge concern for the inverse problem community. This is in sharp contrast to the machine learning applications where obtaining lower training errors with deep learning models typically translates to better generalization performance.

4 Discussion

Optimization is not only used in the data fitting approach to inverse problems, but also used in the Bayesian approach. An important problem in the Bayesian approach is the choice of the parameters for the prior. While these were often chosen in a somewhat ad hoc way, there are studies which use sampling [2, 40], hierarchical prior models [21, 22], and optimization [9, 70] methods to choose the parameters. While choosing the prior parameters through optimization has found some success, such optimization is hard and it remains a challenge to develop effective algorithms to solve these problems.

For inverse problems with large number of measurements, solving each forward problem can be expensive, and the mere evaluation of the misfit function may become computationally prohibitive. Stochastic optimization algorithms might be beneficial in this case, because the objective function is often a sum of misfits over different measurements.

The data fitting problem is generally non-convex and thus optimization algorithms may be trapped in a local optimum. Stochastic optimization algorithms also provide a means to escape the local optima. Recent results in nonconvex optimization, such as those on accelerated methods, may provide more efficient alternatives to solve the data fitting problem.

While box constraints are often used in inverse problems because they are easier to deal with, simplex constraint can be beneficial. The Frank-Wolfe algorithm provides a efficient way to deal with the simplex constraint, and can be a useful tool to add on to the toolbox of an inverse problem researcher.

5 Conclusion

State-of-the-art optimization methods in the inverse problem community struggle to cope with important issues such as large-scale problems and nonconvexity. At the same time, many progresses in optimization have been made in the machine learning community. Our discussion on the connections has been brief. Nevertheless, we have highlighted the valuable potential synergies that are to be reaped by bringing these two communities closer together.

Acknowledgements We thank the anonymous reviewers for their helpful comments. The work was carried out when the author was affiliated with ACEMS & Queensland University of Technology.

References

1. Abdoulaev, G.S., Ren, K., Hielscher, A.H.: Optical tomography as a PDE-constrained optimization problem. *Inverse Prob.* **21**(5), 1507–1530 (2005)
2. Agapiou, S., Bardsley, J.M., Papaspiliopoulos, O., Stuart, A.M.: Analysis of the Gibbs sampler for hierarchical inverse problems. *SIAM/ASA J. Uncertain. Quantif.* **2**(1), 511–544 (2014)
3. Agarwal, N., Bullins, B., Hazan, E.: Second order stochastic optimization in linear time. Preprint, arXiv:1602.03943 (2016)
4. Allen-Zhu, Z., Hazan, E.: Variance reduction for faster non-convex optimization. Preprint, arXiv:1603.05643 (2016)
5. Archer, G., Titterton, D.: On some Bayesian/regularization methods for image restoration. *IEEE Trans. Image Process.* **4**(7), 989–995 (1995)
6. Arridge, S.R.: Optical tomography in medical imaging. *Inverse Prob.* **15**(2), R41 (1999)

7. Arridge, S.R., Hebden, J.C.: Optical imaging in medicine: Ii. Modelling and reconstruction. *Phys. Med. Biol.* **42**(5), 841 (1997)
8. Aster, R.C., Borchers, B., Thurber, C.H.: *Parameter Estimation and Inverse Problems*. Academic, London (2013)
9. Bardsley, J.M., Calvetti, D., Somersalo, E.: Hierarchical regularization for edge-preserving reconstruction of pet images. *Inverse Prob.* **26**(3), 035010 (2010)
10. Benzi, M., Haber, E., Taralli, L.: A preconditioning technique for a class of PDE-constrained optimization problems. *Adv. Comput. Math.* **35**(2), 149–173 (2011)
11. Berahas, A.S., Bollapragada, R., Nocedal, J.: An investigation of Newton-sketch and subsampled Newton methods. Preprint, arXiv:1705.06211 (2017)
12. Bertero, M., Boccacci, P.: *Introduction to Inverse Problems in Imaging*. CRC Press, Boca Raton (2010)
13. Björck, Å.: *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia (1996)
14. Boas, D., Brooks, D., Miller, E., DiMarzio, C.A., Kilmer, M., Gaudette, R., Zhang, Q.: Imaging the body with diffuse optical tomography. *IEEE Signal Process. Mag.* **18**(6), 57–75 (2001)
15. Bollapragada, R., Byrd, R., Nocedal, J.: Exact and inexact subsampled Newton methods for optimization. Preprint, arXiv:1609.08502 (2016)
16. Borcea, L., Berryman, J.G., Papanicolaou, G.C.: High-contrast impedance tomography. *Inverse Prob.* **12**, 835–858 (1996)
17. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. Preprint, arXiv:1606.04838 (2016)
18. Bunks, C., Saleck, F.M., Zaleski, S., Chavent, G.: Multiscale seismic waveform inversion. *Geophysics* **60**(5), 1457–1473 (1995)
19. Byrd, R.H., Chin, G.M., Neveitt, W., Nocedal, J.: On the use of stochastic Hessian information in optimization methods for machine learning. *SIAM J. Optim.* **21**(3), 977–995 (2011)
20. Byrd, R.H., Chin, G.M., Nocedal, J., Wu, Y.: Sample size selection in optimization methods for machine learning. *Math. Program.* **134**(1), 127–155 (2012)
21. Calvetti, D., Somersalo, E.: A gaussian hypermodel to recover blocky objects. *Inverse Prob.* **23**(2), 733 (2007)
22. Calvetti, D., Somersalo, E.: Hypermodels in the Bayesian imaging framework. *Inverse Prob.* **24**(3), 034013 (2008)
23. Cartis, C., Gould, N.I., Toint, P.L.: Evaluation complexity of adaptive cubic regularization methods for convex unconstrained optimization. *Optim. Methods Softw.* **27**(2), 197–219 (2012)
24. Cheney, M., Isaacson, D., Newell, J.C.: Electrical impedance tomography. *SIAM Rev.* **41**, 85–101 (1999)
25. Claerbout, J.F., Muir, F.: Robust modeling with erratic data. *Geophysics* **38**(5), 826–844 (1973)
26. Clason, C.: L_∞ fitting for inverse problems with uniform noise. *Inverse Prob.* **28**(10), 104007 (2012)
27. Conn, A.R., Gould, N.I., Toint, P.L.: *Trust Region Methods*, vol. 1. SIAM, Philadelphia (2000)
28. Dai, Y.: Nonlinear conjugate gradient methods. In: *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, New York (2011)
29. Dauphin, Y., de Vries, H., Bengio, Y.: Equilibrated adaptive learning rates for non-convex optimization. In: *Advances in Neural Information Processing Systems*, pp. 1504–1512 (2015)
30. Doel, K.v.d., Ascher, U.: Adaptive and stochastic algorithms for EIT and DC resistivity problems with piecewise constant solutions and many measurements. *SIAM J. Scient. Comput.* **34** (2012). <https://doi.org/10.1137/110826,692>
31. Doel, K.v.d., Ascher, U., Leita, A.: Multiple level sets for piecewise constant surface reconstruction in highly ill-posed problems. *J. Sci. Comput.* **43**(1), 44–66 (2010)
32. Domingos, P.: A few useful things to know about machine learning. *Commun. ACM* **55**(10), 78–87 (2012)

33. Dorn, O., Miller, E.L., Rappaport, C.M.: A shape reconstruction method for electromagnetic tomography using adjoint fields and level sets. *Inverse Prob.* **16**, 1119–1156 (2000)
34. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12**, 2121–2159 (2011)
35. Eisen, M., Mokhtari, A., Ribeiro, A.: Large scale empirical risk minimization via truncated adaptive Newton method. Preprint, arXiv:1705.07957 (2017)
36. Engl, H.W., Hanke, M., Neubauer, A.: *Regularization of Inverse Problems*. Kluwer, Dordrecht (1996)
37. Erdogdu, M.A., Montanari, A.: Convergence rates of sub-sampled Newton methods. In: *Advances in Neural Information Processing Systems*, vol. 28, pp. 3034–3042 (2015)
38. Fichtner, A.: *Full Seismic Waveform Modeling and Inversion*. Springer, Berlin (2011)
39. Fletcher, R.: *Practical Methods of Optimization*. Wiley, New York (2013)
40. Fox, C., Norton, R.A.: Fast sampling in a linear-gaussian inverse problem. *SIAM/ASA J. Uncertain. Quantif.* **4**(1), 1191–1218 (2016)
41. Frank, M., Wolfe, P.: An algorithm for quadratic programming. *Nav. Res. Logist. Q.* **3**(1–2), 95–110 (1956)
42. Gao, H., Osher, S., Zhao, H.: Quantitative photoacoustic tomography. In: *Mathematical Modeling in Biomedical Imaging II*, pp. 131–158. Springer, Berlin (2012)
43. Ge, R., Huang, F., Jin, C., Yuan, Y.: Escaping from saddle points-online stochastic gradient for tensor decomposition. In: *Proceedings of COLT*, pp. 797–842 (2015)
44. Haber, E.: Quasi-Newton methods for large-scale electromagnetic inverse problems. *Inverse Prob.* **21**(1), 305 (2004)
45. Haber, E., Ascher, U.M.: Preconditioned all-at-once methods for large, sparse parameter estimation problems. *Inverse Prob.* **17**(6), 1847 (2001)
46. Haber, E., Chung, M.: Simultaneous source for non-uniform data variance and missing data. Preprint, arXiv:1404.5254 (2014)
47. Haber, E., Ascher, U.M., Oldenburg, D.: On optimization techniques for solving nonlinear inverse problems. *Inverse Prob.* **16**(5), 1263 (2000)
48. Haber, E., Ascher, U., Oldenburg, D.: Inversion of 3D electromagnetic data in frequency and time domain using an inexact all-at-once approach. *Geophysics* **69**, 1216–1228 (2004)
49. Haber, E., Heldmann, S., Ascher, U.: Adaptive finite volume method for distributed non-smooth parameter identification. *Inverse Prob.* **23**, 1659–1676 (2007)
50. Haber, E., Chung, M., Herrmann, F.: An effective method for parameter estimation with PDE constraints with multiple right-hand sides. *SIAM J. Optim.* **22**, 739–757 (2012)
51. Haber, E., Chung, M., Herrmann, F.: An effective method for parameter estimation with PDE constraints with multiple right-hand sides. *SIAM J. Optim.* **22**(3), 739–757 (2012)
52. Hadamard, J.: Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pp. 49–52 (1902)
53. Hanke, M.: Regularizing properties of a truncated Newton-CG algorithm for nonlinear inverse problems. *Numer. Funct. Anal. Optim.* **18**, 971–993 (1997)
54. Hansen, P.C.: *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, Philadelphia (1998)
55. Herman, G.T.: *Fundamentals of Computerized Tomography: Image Reconstruction from Projections*. Springer Science & Business Media, London (2009)
56. Herrmann, F., Erlangga, Y., Lin, T.: Compressive simultaneous full-waveform simulation. *Geophysics* **74**, A35 (2009)
57. Ito, K., Kunisch, K.: The augmented Lagrangian method for parameter estimation in elliptic systems. *SIAM J. Control Optim.* **28**(1), 113–136 (1990)
58. Jaggi, M.: Revisiting Frank-Wolfe: projection-free sparse convex optimization. In: *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 427–435 (2013)
59. Jain, P., Netrapalli, P., Sanghavi, S.: Low-rank matrix completion using alternating minimization. In: *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, pp. 665–674. ACM, New York (2013)

60. Jin, C., Ge, R., Netrapalli, P., Kakade, S.M., Jordan, M.I.: How to escape saddle points efficiently. Preprint, arXiv:1703.00887 (2017)
61. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: *Advances in Neural Information Processing Systems*, pp. 315–323 (2013)
62. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. Preprint, arXiv:1412.6980 (2014)
63. Kumar, R., Silva, C.D., Akalin, O., Aravkin, A.Y., Mansour, H., Recht, B., Herrmann, F.J.: Efficient matrix completion for seismic data reconstruction. *Geophysics* **80**(5), V97–V114 (2015)
64. Lan, G., Zhou, Y.: Conditional gradient sliding for convex optimization. *SIAM J. Optim.* **26**(2), 1379–1409 (2016)
65. Lan, G., Pokutta, S., Zhou, Y., Zink, D.: Conditional accelerated lazy stochastic gradient descent. In: *Proceedings of ICML. PMLR* (2017). <http://proceedings.mlr.press/v70/lan17a.html>
66. Levy, K.Y.: The power of normalization: faster evasion of saddle points. Preprint, arXiv:1611.04831 (2016)
67. Li, H., Lin, Z.: Accelerated proximal gradient methods for nonconvex programming. In: *Advances in Neural Information Processing Systems*, pp. 379–387 (2015)
68. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. *Math. Program.* **45**(1–3), 503–528 (1989)
69. Liu, M., Yang, T.: On noisy negative curvature descent: competing with gradient descent for faster non-convex optimization. Preprint, arXiv:1709.08571 (2017)
70. Liu, W., Li, J., Marzouk, Y.M.: An approximate empirical Bayesian method for large-scale linear-gaussian inverse problems. Preprint, arXiv:1705.07646 (2017)
71. Louis, A.: Medical imaging: state of the art and future development. *Inverse Prob.* **8**(5), 709 (1992)
72. Mandt, S., Hoffman, M., Blei, D.: A variational analysis of stochastic gradient algorithms. In: *International Conference on Machine Learning*, pp. 354–363 (2016)
73. Mazumder, R., Friedman, J.H., Hastie, T.: Sparsenet: Coordinate descent with nonconvex penalties. *J. Am. Stat. Assoc.* **106**(495), 1125–1138 (2011)
74. Menke, W.: *Geophysical Data Analysis: Discrete Inverse Theory*. Academic, London (2012)
75. Mutn̄y, M.: Stochastic second-order optimization via von Neumann series. Preprint, arXiv:1612.04694 (2016)
76. Mutn̄y, M., Richtárik, P.: Parallel stochastic Newton method. Preprint, arXiv:1705.02005 (2017)
77. Natterer, F., Wübbeling, F.: *Mathematical Methods in Image Reconstruction*. SIAM, Philadelphia (2001)
78. Newman, G.A., Alumbaugh, D.L.: Frequency-domain modelling of airborne electromagnetic responses using staggered finite differences. *Geophys. Prospect.* **43**, 1021–1042 (1995)
79. Nocedal, J.: Updating quasi-Newton matrices with limited storage. *Math. Comput.* **35**(151), 773–782 (1980)
80. Nocedal, J., Wright, S.: *Numerical Optimization*. Springer Science & Business Media, New York (2006)
81. Oldenburg, D., Haber, E., Shekhtman, R.: 3D inversion of multi-source time domain electromagnetic data. *J. Geophys.* **78**(1), E47–E57 (2013)
82. Osher, S., Sethian, J.: Fronts propagating with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations. *J. Comp. Phys.* **79**, 12–49 (1988)
83. Pidlisecky, A., Haber, E., Knight, R.: RESINVM3D: a MATLAB 3D resistivity inversion package. *Geophysics* **72**(2), H1–H10 (2007)
84. Pilanci, M., Wainwright, M.J.: Newton sketch: a linear-time optimization algorithm with linear-quadratic convergence. Preprint, arXiv:1505.02250 (2015)
85. Reddi, S.J., Hefny, A., Sra, S., Póczos, B., Smola, A.: Stochastic variance reduction for nonconvex optimization. Preprint, arXiv:1603.06160 (2016)

86. Rieder, A.: Inexact Newton regularization using conjugate gradients as inner iteration. *SIAM J. Numer. Anal.* **43**, 604–622 (2005)
87. Rieder, A., Lechleiter, A.: Towards a general convergence theory for inexact Newton regularizations. *Numer. Math.* **114**(3), 521–548 (2010)
88. Rohmberg, J., Neelamani, R., Krohn, C., Krebs, J., Deffenbaugh, M., Anderson, J.: Efficient seismic forward modeling and acquisition using simultaneous random sources and sparsity. *Geophysics* **75**(6), WB15–WB27 (2010)
89. Roosta-Khorasani, F.: Randomized algorithms for solving large scale nonlinear least squares problems. Ph.D. thesis, University of British Columbia (2015)
90. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods I: globally convergent algorithms. Preprint, arXiv:1601.04737 (2016)
91. Roosta-Khorasani, F., Mahoney, M.W.: Sub-sampled Newton methods II: local convergence rates. Preprint, arXiv:1601.04738 (2016)
92. Roosta-Khorasani, F., van den Doel, K., Ascher, U.: Data completion and stochastic algorithms for PDE inversion problems with many measurements. *Electron. Trans. Numer. Anal.* **42**, 177–196 (2014)
93. Roosta-Khorasani, F., van den Doel, K., Ascher, U.: Stochastic algorithms for inverse problems involving PDEs and many measurements. *SIAM J. Sci. Comput.* **36**(5), S3–S22 (2014). <https://doi.org/10.1137/130922756>
94. Roosta-Khorasani, F., Székely, G.J., Ascher, U.: Assessing stochastic algorithms for large scale nonlinear least squares problems using extremal probabilities of linear combinations of gamma random variables. *SIAM/ASA J. Uncertain. Quantif.* **3**(1), 61–90 (2015)
95. Rundell, W., Engl, H.W.: *Inverse Problems in Medical Imaging and Nondestructive Testing*. Springer, New York (1997)
96. Russell, B.H.: *Introduction to Seismic Inversion Methods*, vol. 2. Society of Exploration Geophysicists, Tulsa (1988)
97. Scharf, L.L.: *Statistical Signal Processing*, vol. 98. Addison-Wesley, Reading (1991)
98. Sen, M.K., Stoffa, P.L.: *Global Optimization Methods in Geophysical Inversion*. Cambridge University Press, Cambridge (2013)
99. Shalev-Shwartz, S., Zhang, T.: Stochastic dual coordinate ascent methods for regularized loss. *J. Mach. Learn. Res.* **14**(1), 567–599 (2013)
100. Smith, N.C., Vozoff, K.: Two dimensional DC resistivity inversion for dipole dipole data. *IEEE Trans. Geosci. Remote Sens.* **GE 22**, 21–28 (1984)
101. Sun, W., Yuan, Y.X.: *Optimization Theory and Methods: Nonlinear Programming*, vol. 1. Springer Science & Business Media, New York (2006)
102. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: *International Conference on Machine Learning*, pp. 1139–1147 (2013)
103. Tai, X.C., Li, H.: A piecewise constant level set method for elliptic inverse problems. *Appl. Numer. Math.* **57**, 686–696 (2007)
104. Tarantola, A.: *Inverse Problem Theory and Methods for Model Parameter Estimation*. SIAM, Philadelphia (2005)
105. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. In: *COURSERA: Neural Networks for Machine Learning*, vol. 4 (2012)
106. Tripuraneni, N., Stern, M., Jin, C., Regier, J., Jordan, M.I.: Stochastic cubic regularization for fast nonconvex optimization. Preprint, arXiv:1711.02838 (2017)
107. van den Doel, K., Ascher, U.M.: On level set regularization for highly ill-posed distributed parameter estimation problems. *J. Comp. Phys.* **216**, 707–723 (2006)
108. van den Doel, K., Ascher, U.M.: Dynamic level set regularization for large distributed parameter estimation problems. *Inverse Prob.* **23**, 1271–1288 (2007)
109. van den Doel, K., Ascher, U.M.: Dynamic regularization, level set shape optimization, and computed myography. In: *Control and Optimization with Differential-Algebraic Constraints*, vol. 23, p. 315. SIAM, Philadelphia (2012)

110. Van Den Doel, K., Ascher, U., Haber, E.: The lost honour of ℓ_2 -based regularization. In: Large Scale Inverse Problems. Radon Series on Computational and Applied Mathematics, vol. 13, pp. 181–203. De Gruyter (2012)
111. Vogel, C.: Computational Methods for Inverse Problem. SIAM, Philadelphia (2002)
112. Wang, C.C., Huang, C.H., Lin, C.J.: Subsampled Hessian Newton methods for supervised learning. *Neural Comput.* **27**(8), 1766–1795 (2015)
113. Xu, P., Yang, J., Roosta-Khorasani, F., Ré, C., Mahoney, M.W.: Sub-sampled Newton methods with non-uniform sampling. In: Advances in Neural Information Processing Systems (NIPS), pp. 2530–2538 (2016)
114. Xu, P., Roosta-Khorasani, F., Mahoney, M.W.: Newton-type methods for non-convex optimization under inexact hessian information. Preprint, arXiv:1708.07164 (2017)
115. Xu, P., Roosta-Khorasani, F., Mahoney, M.W.: Second-order optimization for non-convex machine learning: an empirical study. Preprint, arXiv:1708.07827 (2017)
116. Ye, H., Luo, L., Zhang, Z.: Revisiting sub-sampled Newton methods. Preprint, arXiv:1608.02875 (2016)
117. Yuan, Z., Jiang, H.: Quantitative photoacoustic tomography: recovery of optical absorption coefficient maps of heterogeneous media. *Appl. Phys. Lett.* **88**(23), 231101 (2006)
118. Zeiler, M.D.: Adadelta: an adaptive learning rate method. Preprint, arXiv:1212.5701 (2012)

Diagonal Form Factors from Non-diagonal Ones



Zoltan Bajnok and Chao Wu

Abstract We prove the asymptotic large volume expression of diagonal form factors in integrable models by evaluating carefully the diagonal limit of a non-diagonal form factor in which we send the rapidity of the extra particle to infinity.

1 Introduction

Two dimensional integrable quantum field theories are useful toy models of statistical and particle physics as they provide many interesting observables, which can be calculated exactly [12]. These models are first solved in infinite volume, where the scattering matrix [4, 21], which connects asymptotic multiparticle states, are determined together with the form factors which are the matrix elements of local operators sandwiched between the same asymptotic states [19]. These form factors then can be used to build up the correlation functions, which define the theory in the Wightman sense [1].

In the relevant practical applications, however, quantum field theories are confined to a finite volume and the calculation of finite size corrections is unavoidable. Fortunately, all these finite size corrections can be expressed in terms of the infinite volume characteristics, such as masses, scattering matrices and form factors [10, 11, 15]. We can distinguish three domains in the volume according to the nature of the corrections. The leading finite size corrections are polynomial in the inverse power of the volume, while the sub-leading corrections are exponentially volume-suppressed.

Concerning the finite volume energy spectrum the domain when only polynomial corrections are kept is called the Bethe-Yang (BY) domain. We there merely need to take into account the finite volume quantization of the momenta, which originates from the periodicity requirement and explicitly includes the scattering

Z. Bajnok (✉) · C. Wu

MTA Lendület Holographic QFT Group, Wigner Research Centre for Physics, Budapest, Hungary
e-mail: bajnok.zoltan@wigner.mta.hu; chao.wu@wigner.mta.hu

phase-shifts [11]. The exponentially small corrections are due to virtual particles traveling around the world and the domain in which we keep only the leading exponential correction is called the Luscher domain [10]. In a small volume, when all exponentials contribute the same way, we have to sum them up leading to a description given by the Thermodynamic Bethe Ansatz (TBA) [20].

The situation for the form factors are not understood at the same level yet. The BY domain was investigated in [15, 16]. It was proven for non-diagonal form factors that all polynomial finite size effects come only from the finite volume (Kronecker-delta) normalization of states. The authors also conjectured the BY form of diagonal finite volume form factors, which they derived for two particle-states. The leading exponential finite size corrections for generic form factors are not known, except for the diagonal ones, for which exact conjectures exist. The LeClair-Mussardo (LM) conjecture expresses the exact finite volume/temperature one-point functions in terms of infinite volume diagonal connected form factors, and densities of mirror states determined by the TBA equation [9]. Actually it was shown in [13, 14] that the BY form of diagonal form factors implies the LM formula and vice versa. Using analytical continuation a 'la [5] Pozsgay extended the LM formula for finite volume diagonal matrix elements [17]. The aim of the present paper is to prove the conjectured BY form of diagonal form factors [16, 18] from the already proven non-diagonal BY form factors [15] by carefully calculating the diagonal limit, in which we send one particle's rapidity to infinity. By this way our result also leads to the proof of the LM formula. Here we focus on theories with one type of particles.

The paper is organized such that in the next section we summarize the known facts about the BY form of diagonal and non-diagonal form factors. We then in Sect. 3 prove the diagonal conjecture and conclude in Sect. 4.

2 The Conjecture for Diagonal Large Volume Form Factors

In this section we introduce the infinite volume form factors and their properties and use them later on to describe the finite volume form factors in the BY domain.

2.1 Infinite Volume Form Factors

Infinite volume form factors are the matrix elements of local operators sandwiched between asymptotic states $\langle \theta'_1, \dots, \theta'_m | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle$. We use the rapidity θ to parametrize the momenta as $p = m \sinh \theta$. The crossing formula

$$\langle \theta'_1, \dots, \theta'_m | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle = \langle \theta'_1, \dots, \theta'_{m-1} | \mathcal{O} | \bar{\theta}'_m - i\epsilon, \theta_n, \dots, \theta_1 \rangle + \sum_{i=1}^n 2\pi \delta(\theta'_m - \theta_i) \prod_{j=i+1}^n S(\theta_j - \theta_i) \langle \theta'_1, \dots, \theta'_{m-1} | \mathcal{O} | \theta_{n-1}, \dots, \theta_1 \rangle \quad (1)$$

can be used to express every matrix element in terms of the elementary form factors

$$\langle 0 | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle = F_n(\theta_n, \dots, \theta_1) \tag{2}$$

where $\bar{\theta} = \theta + i\pi$ denotes the crossed rapidity and the two particle S-matrix satisfies $S(\theta) = S(i\pi - \theta) = S(-\theta)^{-1}$. Infinite volume states are normalized to Dirac δ -functions: as $\langle \theta' | \theta \rangle = 2\pi\delta(\theta - \theta')$. The elementary form factor satisfies the permutation and periodicity axiom

$$\begin{aligned} F_n(\theta_1, \theta_2, \dots, \theta_i, \theta_{i+1} \dots, \theta_n) &= S(\theta_i - \theta_{i+1}) F_n(\theta_1, \theta_2, \dots, \theta_{i+1}, \theta_i \dots, \theta_n) \\ &= F_n(\theta_2, \dots, \theta_i, \theta_{i+1} \dots, \theta_n, \theta_1 - 2i\pi) \end{aligned} \tag{3}$$

together with the kinematical singularity relation

$$-i \text{Res}_{\theta'=\theta} F_{n+2}(\theta' + i\pi, \theta, \theta_1, \dots, \theta_n) = (1 - \prod_{i=1}^n S(\theta - \theta_i)) F_n(\theta_1, \dots, \theta_n) \tag{4}$$

For scalar operators, when properly normalized, the form factor also satisfies the cluster property

$$\lim_{\Lambda \rightarrow \infty} F_{n+m}(\theta_1 + \Lambda, \dots, \theta_n + \Lambda, \theta_{n+1}, \dots, \theta_{n+m}) = F_n(\theta_1, \dots, \theta_n) F_m(\theta_{n+1}, \dots, \theta_{n+m}) \tag{5}$$

which will be used to analyze the diagonal limit of $\langle \theta, \theta'_1, \dots, \theta'_n | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle$ via $\theta \rightarrow \infty$ in finite volume.

The diagonal form factors $\langle \theta_1, \dots, \theta_n | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle$ are singular due to the $\delta(\theta)$ terms coming from the normalization of the states and also from poles related to the kinematical singularity axiom. Actually, $F_{2n}(\bar{\theta}_1 + \epsilon_1, \dots, \bar{\theta}_n + \epsilon_n, \theta_n, \dots, \theta_1)$ is not singular when all ϵ_i go to zero simultaneously, but depends on the direction of the limit. The *connected* diagonal form factor is defined as the finite ϵ -independent part:

$$F_{2n}^c(\theta_1, \dots, \theta_k) = \text{Fp} (F_{2n}(\bar{\theta}_1 + \epsilon_1, \dots, \bar{\theta}_n + \epsilon_n, \theta_n, \dots, \theta_1)) \tag{6}$$

while the *symmetric* evaluation is simply

$$F_{2n}^s(\theta_1, \dots, \theta_k) = \lim_{\epsilon \rightarrow 0} F_{2n}(\bar{\theta}_1 + \epsilon, \dots, \bar{\theta}_n + \epsilon, \theta_n, \dots, \theta_1) \tag{7}$$

where Fp means the finite part, in order to understand the singularity structure of the diagonal limit we note that the singular part can very nicely be visualized by graphs [16]:

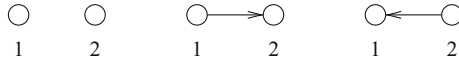
$$F_{2n}(\bar{\theta}_1 + \epsilon_1, \dots, \bar{\theta}_n + \epsilon_n, \theta_n, \dots, \theta_1) = \sum_{\text{allowed graphs}} F(\text{graph}) + O(\epsilon_i) \tag{8}$$

where an allowed graph is an oriented tree-like (no-loop) graph in which at each vertex there is at most one outgoing edge. The contribution of a graph, $F(\text{graph})$, can be evaluated as follows: points (i_1, \dots, i_k) with no outgoing edges contribute a factor, $F_{2k}^c(\theta_{i_1}, \dots, \theta_{i_k})$, while for each edge from i to j we associate a factor $\frac{\epsilon_j}{\epsilon_i} \phi(\theta_i - \theta_j)$, where $\phi(\theta) = -i \partial_\theta \log S(\theta) = -i \frac{S'(\theta)}{S(\theta)}$. We recall the proof of (8) from [16] as similar argumentations will be used later on. The proof goes in induction in n and evaluates the residue at $\epsilon_n = 0$ keeping all other ϵ s finite. Clearly such singular term can come only from graphs in which n has only an outgoing edge and no incoming ones. The contributions of such terms are

$$\frac{1}{\epsilon_n} (\epsilon_1 \phi_{1n} + \dots + \epsilon_{n-1} \phi_{n-1n}) F_{2n-2}(\bar{\theta}_1 + \epsilon_1, \dots, \bar{\theta}_{n-1} + \epsilon_{n-1}, \theta_{n-1}, \dots, \theta_1) \tag{9}$$

where $\phi_{jk} = \phi_{kj} = \phi(\theta_i - \theta_j)$. Now comparing this expression to the kinematical singularity axiom and using the definition of $\phi(\theta)$ together with the properties of the scattering matrix we can see that they completely agree. The formula (8) can be used to define connected form factors recursively by subtracting the singular terms and taking the diagonal limit. Observe also that taking all ϵ to be the same makes the lhs. of (8) the symmetric form factor, which is expressed by (8) in terms of the connected ones.

In particular, for the 2-particle form factor we have only three graphs:



which give

$$F_4(\bar{\theta}_1 + \epsilon_1, \bar{\theta}_2 + \epsilon_2, \theta_2, \theta_1) = F_4^c(\theta_1, \theta_2) + \frac{\epsilon_1}{\epsilon_2} \phi_{12} F_2^c(\theta_1) + \frac{\epsilon_2}{\epsilon_1} \phi_{21} F_2^c(\theta_2) + O(\epsilon_i) \tag{10}$$

This equation on the one hand can be used to define $F_4^c(\theta_1, \theta_2)$, once $F_2^c(\theta)$ has been already defined, and on the other hand, it connects the symmetric form factor to the connected one:

$$F_4^s(\theta_1, \theta_2) = F_4^c(\theta_1, \theta_2) + \phi_{12} F_2^c(\theta_1) + \phi_{21} F_2^c(\theta_2) \tag{11}$$

2.2 Finite Volume Form Factors in the BY Domain

In the BY domain we drop the exponentially suppressed $O(e^{-mL})$ terms and keep only the $O(L^{-1})$ polynomial volume dependence. The quantization of the momenta

is given by the BY equations

$$Q_j \equiv p(\theta_j)L - i \sum_{k:k \neq j} \log S(\theta_j - \theta_k) = 2\pi I_j \quad (12)$$

An n -particle state is labeled by the integers I_j , which can be traded for the momenta: $|I_1, \dots, I_n\rangle \equiv |\theta_1, \dots, \theta_n\rangle_L$. These states are normalized to Kronecker delta functions $\langle I'|I\rangle = \prod_j \delta_{I'_j I_j}$. Since two point functions in finite and infinite volume are equal up to exponentially small $O(e^{-mL})$ terms, the finite and infinite volume form factors differ only in the normalization of states [15]. In particular, this implies the non-diagonal finite volume form factor formula

$$\langle \theta'_1, \dots, \theta'_m | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle_L = \frac{F_{n+m}(\bar{\theta}'_1, \dots, \bar{\theta}'_m, \theta_n, \dots, \theta_1)}{\sqrt{\rho_n \rho'_m}} + O(e^{-mL}) \quad (13)$$

where the densities of states are defined through the Bethe Ansatz equation via

$$\rho_n = \det |Q_{ij}| \quad ; \quad Q_{ij} = \partial_i Q_j \equiv \frac{\partial Q_j}{\partial \theta_i} \quad (14)$$

The conjectured formula for diagonal form factors takes the form [18]:

$$\langle \theta_1, \dots, \theta_n | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle_L = \frac{\sum_{\alpha \cup \bar{\alpha}} F_\alpha^c \rho_{\bar{\alpha}}^c}{\rho_n} + O(e^{-mL}) \quad (15)$$

where the index set $I = \{1, \dots, n\}$ is split in all possible ways $I = \alpha \cup \bar{\alpha}$, $F_\alpha^c = F_{2k}^c(\theta_{\alpha_1}, \dots, \theta_{\alpha_k})$ with $|\alpha| = k$ and $\rho_{\bar{\alpha}}$ is the shorthand for $\rho_{n-k}(\theta_{\bar{\alpha}_1}, \dots, \theta_{\bar{\alpha}_{n-k}})$, which denotes the sub-determinant of the matrix, Q_{ij} , with indices only from $\bar{\alpha}$. There is an analogous expression in terms of the symmetric form factors [16]

$$\langle \theta_1, \dots, \theta_n | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle_L = \frac{\sum_{\alpha \cup \bar{\alpha}} F_\alpha^s \rho_{\bar{\alpha}}^s}{\rho_n} + O(e^{-mL}) \quad (16)$$

where now ρ_α^s is the density of states corresponding to the variables with labels in α . The equivalence of the two formulas was shown in [16]. Let us note that for $L = 0$ the sum reduces to one single term $\sum_{\alpha \cup \bar{\alpha}} F_\alpha^s \rho_{\bar{\alpha}}^s \rightarrow F_n^s$ as all other ρ^s factor vanish.

Let us spell out the details for two particles. The diagonal finite volume form factor up to exponential correction is

$$\langle \theta_1, \theta_2 | \mathcal{O} | \theta_2, \theta_1 \rangle_L = \frac{F_4^c(\theta_1, \theta_2) + \rho_1(\theta_1)F_2^c(\theta_2) + \rho_1(\theta_2)F_2^c(\theta_1) + \rho_2(\theta_1, \theta_2)F_0}{\rho_2(\theta_1, \theta_2)} \quad (17)$$

where

$$\rho_2(\theta_1, \theta_2) = \begin{vmatrix} E_1 L + \phi_{12} & -\phi_{12} \\ -\phi_{21} & E_2 L + \phi_{21} \end{vmatrix} ; \quad \rho_1(\theta_i) = E_i L + \phi_{i3-i}$$

where $E_i = \partial_i p(\theta_i)$. The analogous formula with the symmetric evaluation reads as

$$\langle \theta_1, \theta_2 | \mathcal{O} | \theta_2, \theta_1 \rangle_L = \frac{F_4^s(\theta_1, \theta_2) + \rho_1^s(\theta_1) F_2^s(\theta_2) + \rho_1^s(\theta_2) F_2^s(\theta_1) + \rho_2^s(\theta_1, \theta_2) F_0^s}{\rho_2(\theta_1, \theta_2)} \quad (18)$$

where

$$\rho_2^s(\theta_1, \theta_2) = \rho_2(\theta_1, \theta_2) \quad ; \quad \rho_1^s(\theta_i) = E_i L$$

3 The Proof for Diagonal Large Volume Form Factors

The idea of the proof follows from the large θ behaviour of the scattering matrix, namely $S(\theta) \rightarrow 1$, for $\theta \rightarrow \infty$. This also lies behind the cluster property of the form factors. Thus by taking the non-diagonal form factor $\langle \theta, \theta'_1, \dots, \theta'_n | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle_L$ and sending $\theta \rightarrow \infty$, the extra particle decouples and we can approach the diagonal form factor. This can be achieved by choosing the same quantization numbers for both the θ_j and θ'_j particles:

$$Q'_j \equiv p(\theta'_j) L - i \sum_{k:k \neq j} \log S(\theta'_j - \theta'_k) - i \log S(\theta'_j - \theta) = 2\pi I_j \quad (19)$$

Indeed, by sending (the quantization number of) θ to infinity the BY equations, Q'_j , reduce Q_j . This means that in the limit considered $\theta'_i \rightarrow \theta_i$ as $\epsilon_i \rightarrow 0$. In principle, ϵ_i depends on $\{\theta_i\}$ and on the way how θ goes to infinity.

For finite θ , the form factor is non-diagonal and we can use

$$\langle \theta, \theta'_1, \dots, \theta'_n | \mathcal{O} | \theta_n, \dots, \theta_1 \rangle_L = \frac{F_{2n+1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_n, \theta_n, \dots, \theta_1)}{\sqrt{\rho'_{n+1} \rho_n}} + O(e^{-mL}) \quad (20)$$

The numerator is a finite quantity for any θ and has a finite $\theta \rightarrow \infty$ limit accordingly. We can see in the limit that $\rho'_{n+1}(\theta, \theta'_1, \dots, \theta'_n)$ goes to $\rho_1(\theta) \rho_n(\theta_1, \dots, \theta_n)$. Similarly, for the form factors $F_{2n+1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_n, \theta_n, \dots, \theta_1)$ the cluster property guaranties the factorization $F_{2n}(\bar{\theta}'_1, \dots, \bar{\theta}'_n, \theta_n, \dots, \theta_1) F_1(\bar{\theta})$, where additionally

$\theta'_i \rightarrow \theta_i$. Actually the expression depends in the direction we take the limit in which all ϵ_i go to zero and our main task is to calculate this limit explicitly. Fortunately, the direction is dictated by the difference of the BY equations:

$$Q'_j - Q_j = E_j L \epsilon_j + \sum_{k:k \neq j} \phi_{jk} (\epsilon_j - \epsilon_k) - \delta_j = \sum_k Q_{jk} \epsilon_k - \delta_j = 0 \quad (21)$$

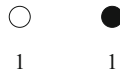
where we have used the notations $\delta_j = i \log S(\theta_j - \theta)$.

Clearly δ_j s are small and so are the ϵ_j s. In the following we analyze the ϵ and δ dependence of the form factor $F_{2n+1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_n, \theta_n, \dots, \theta_1)$. Similarly to the diagonal limit of form factors we can describe the δ and ϵ dependence by graphs. We claim that

$$F_{2n+1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_k, \theta_k, \dots, \theta_1) = \sum_{\text{allowed graphs, colorings}} F(\text{graph}) + O(\epsilon_i, \delta) \quad (22)$$

where, additionally to the previous graphs in (8), we should allow the coloring of those vertices, which do not have any outgoing edge, i.e. they can be either black or white. For each black dot with label i we associate a factor $\frac{\delta_i}{\epsilon_i}$. Note that in the $\theta \rightarrow \infty$ limit we will have an overall $F_1(\bar{\theta})$ factor, which we factor out.

Let us see how it works for $n = 1$: The single dot can be either black or white:



thus the two contributions are

$$F_3(\bar{\theta}, \bar{\theta}'_1, \theta_1) F_1(\bar{\theta})^{-1} = \frac{\delta_1}{\epsilon_1} + F_2^c(\theta_1) + \dots \quad (23)$$

where ellipsis represents terms vanishing in the $\delta, \epsilon \rightarrow 0$ limit. Let us show that $F_2^c(\theta_1)$ is not singular, i.e. the singularity of the lhs. is exactly $\frac{\delta_1}{\epsilon_1}$. The kinematical residue equation tells us that

$$F_3(\bar{\theta}, \bar{\theta}'_1, \theta_1) = \frac{i}{\epsilon_1} (1 - S(\theta'_1 - \theta + i\pi)) F_1(\bar{\theta}) + O(1) = \frac{\delta_1}{\epsilon_1} F_1(\bar{\theta}) + O(1) \quad (24)$$

Thus, once the singularity is subtracted, we can safely take the $\epsilon_1 \rightarrow 0$ and the $\delta \rightarrow 0$ limits leading to

$$\lim_{\delta, \epsilon_1 \rightarrow 0} (F_3(\bar{\theta}, \bar{\theta}'_1, \theta_1) - \frac{\delta_1}{\epsilon_1} F_1(\bar{\theta})) = F_2^c(\theta_1) F_1(\bar{\theta}) \tag{25}$$

where we used the cluster property of form factors and the fact that the two particle diagonal connected form factor is non-singular.

Now we adapt the proof in the induction step in (8) by noticing that the ϵ_n^{-1} singularity can come either from terms with only one outgoing edge or from being black. Thus the residue is

$$\frac{1}{\epsilon_n} (\delta_n + \epsilon_1 \phi_{1n} + \dots + \epsilon_{n-1} \phi_{n-1n}) F_{2n-1}(\bar{\theta}, \bar{\theta}'_1 + \epsilon_1, \dots, \bar{\theta}'_{n-1} + \epsilon_{n-1}, \theta_{n-1}, \dots, \theta_1) \tag{26}$$

Let us calculate the analogous term from the kinematical residue axiom:

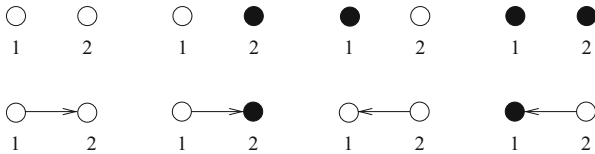
$$F_{2n+1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_n, \theta_n, \dots, \theta_1) \rightarrow \frac{i}{\epsilon_n} \left(1 - \frac{S(\theta'_n - \theta_{n-1}) \dots S(\theta'_n - \theta_1)}{S(\theta'_n - \theta'_{n-1}) \dots S(\theta'_n - \theta'_1)} \frac{1}{S(\theta'_n - \theta)} \right) \times F_{2n-1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_{n-1}, \theta_{n-1}, \dots, \theta_1) \tag{27}$$

The bracket can be expanded as

$$() = -i(\delta_n + \phi_{nn-1} \epsilon_{n-1} + \dots + \phi_{n1} \epsilon_1) \tag{28}$$

which completes the induction.

In particular, for two particles we have the following diagrams:



which lead to the formula

$$F_5(\bar{\theta}, \bar{\theta}'_1, \bar{\theta}'_2, \theta_2, \theta_1) F_1^{-1} = F_4^c(\theta_1, \theta_2) + \frac{\epsilon_2}{\epsilon_1} \phi_{21} \frac{\delta_2}{\epsilon_2} + \frac{\epsilon_1}{\epsilon_2} \phi_{12} \frac{\delta_1}{\epsilon_1} + \frac{\delta_1}{\epsilon_1} \frac{\delta_2}{\epsilon_2} + \frac{\epsilon_1}{\epsilon_2} \phi_{12} F_2^c(\theta_1) + \frac{\delta_2}{\epsilon_2} F_2^c(\theta_1) + \frac{\epsilon_2}{\epsilon_1} \phi_{21} F_2^c(\theta_2) + \frac{\delta_1}{\epsilon_1} F_2^c(\theta_2) \tag{29}$$

It is interesting to check the coefficient of $F_2^c(\theta_1)$:

$$\frac{\epsilon_1 \phi_{12} + \delta_2}{\epsilon_2} = E_2 L + \phi_{21} = \rho_1(\theta_2) \quad (30)$$

where we used the BY equations. Similarly

$$\frac{\delta_1}{\epsilon_1} \frac{\delta_2}{\epsilon_2} + \frac{\epsilon_1}{\epsilon_2} \phi_{12} \frac{\delta_1}{\epsilon_1} + \frac{\epsilon_2}{\epsilon_1} \phi_{21} \frac{\delta_2}{\epsilon_2} = \rho_2(\theta_1, \theta_2) \quad (31)$$

which leads to the sought for formula for $n = 2$:

$$F_5(\bar{\theta}, \bar{\theta}'_1, \bar{\theta}'_2, \theta_2, \theta_1) F_1^{-1} = F_4^c(\theta_1, \theta_2) + \rho_1(\theta_2) F_2^c(\theta_1) + \rho_1(\theta_1) F_2^c(\theta_2) + \rho_2(\theta_1, \theta_2) \quad (32)$$

In the following we prove the form of the diagonal form factors in the general case by induction. First we notice that once we use the BY equations to express δ_i in terms of ϵ_k then all denominators of ϵ s disappear. Focus on ϵ_n^{-1} and observe that

$$\delta_n + \epsilon_1 \phi_{1n} + \dots + \epsilon_{n-1} \phi_{n-1n} = \epsilon_n (E_n L + \phi_{n-1n} + \dots + \phi_{1n}) \quad (33)$$

This implies that the diagonal finite volume form factor is a polynomial in L and linear in each $E_k L$. We first check the $L = 0$ piece and then calculate the derivative wrt. $E_n L$ as the full expression is symmetric in all variables. Note that the naively singular term in ϵ_n at $L = 0$ takes the form:

$$\frac{1}{\epsilon_n} \epsilon_n (E_n L + \phi_{n-1n} + \dots + \phi_{1n}) |_{L=0} = \frac{1}{\epsilon} (\epsilon \phi_{n-1n} + \dots + \epsilon \phi_{1n}) \quad (34)$$

which is exactly the same we would obtain if we had calculated the diagonal limit of the form factor in the symmetric evaluation, i.e. for $L = 0$ we obtain the symmetric n -particle form factor. We now check the linear term in $E_n L$. In doing so we differentiate the expression (22) wrt. $E_n L$:

$$\partial_{E_n L} F_{2n+1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_n, \theta_n, \dots, \theta_1) = F_{2n-1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_{n-1}, \theta_{n-1}, \dots, \theta_1) \quad (35)$$

since the term $E_n L$ can come only through the singularity at $\epsilon_n = 0$. Note that on the rhs. θ_k satisfies the original BY equations and not the one where θ_n is missing. Let us now take a look at the expression we would like to prove:

$$F_{2n+1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_n, \theta_n, \dots, \theta_1) F_1^{-1} = \sum_{\alpha \cup \bar{\alpha} = I} F_\alpha^c \rho_{\bar{\alpha}} = \sum_{\alpha \cup \bar{\alpha} = I} F_\alpha^s \rho_{\bar{\alpha}}^s \quad (36)$$

where $I = \{1, \dots, n\}$. Clearly the rhs. is also a polynomial in L , which is linear in each $E_k L$. To finish the proof, we note that the $L = 0$ constant part of the rhs. is the symmetric form factor. Using that $\partial_{E_n L} \rho_\alpha = \rho_{\alpha \setminus \{n\}}$ if $n \in \alpha$ and 0 otherwise we can see that

$$\partial_{E_n L} \sum_{\alpha \cup \bar{\alpha} = I} F_\alpha^c \rho_{\bar{\alpha}} = \sum_{\beta \cup \bar{\beta} = I \setminus \{n\}} F_\beta^c \rho_{\bar{\beta}} = F_{2n-1}(\bar{\theta}, \bar{\theta}'_1, \dots, \bar{\theta}'_{n-1}, \theta_{n-1}, \dots, \theta_1) F_1^{-1} \quad (37)$$

by the induction hypothesis, which completes the proof.

4 Conclusion

In this paper we proved the large volume expression for the diagonal form factors by taking carefully the limit of a nondiagonal form factor. Our result completes the proof of the LM formula, which describes exactly the one-point function in finite volume.

Diagonal finite volume form factors are relevant in the AdS/CFT correspondence as they are conjectured to describe the Heavy-Heavy-Light (HHL) type three point functions of the maximally supersymmetric 4D gauge theory [3]. This conjecture was first proved at weak coupling [6] then at strong coupling [2], finally for all couplings in [7, 8]. We have profited from all of these proofs and used them in the present paper.

There is a natural extension of our results for diagonal form factors in non-diagonal theories. Clearly the same idea of adding one more particle and sending its rapidity to infinity can be applied there too and we have an ongoing research into this direction.

Acknowledgements We thank Yunfeng Jiang the enlightening discussions and ZB thanks the hospitality of the mathematical research institute MATRIX in Australia where the paper was finalized. The work was supported by a Lendület and by the NKFIH 116505 Grant.

References

1. Babujian, H., Karowski, M.: Towards the construction of Wightman functions of integrable quantum field theories. *Int. J. Mod. Phys. A* **19S2**, 34–49 (2004). <https://doi.org/10.1142/S0217751X04020294>
2. Bajnok, Z., Janik, R.A.: Classical limit of diagonal form factors and HHL correlators. *J. High Energy Phys.* **01**, 063 (2017). [https://doi.org/10.1007/JHEP01\(2017\)063](https://doi.org/10.1007/JHEP01(2017)063)
3. Bajnok, Z., Janik, R.A., Wereszczynski, A.: HHL correlators, orbit averaging and form factors. *J. High Energy Phys.* **09**, 050 (2014). [https://doi.org/10.1007/JHEP09\(2014\)050](https://doi.org/10.1007/JHEP09(2014)050)
4. Dorey, P.: Exact S matrices, pp. 85–125 (1996)

5. Dorey, P., Tateo, R.: Excited states by analytic continuation of TBA equations. *Nucl. Phys.* **B482**, 639–659 (1996). [https://doi.org/10.1016/S0550-3213\(96\)00516-0](https://doi.org/10.1016/S0550-3213(96)00516-0)
6. Hollo, L., Jiang, Y., Petrovskii, A.: Diagonal form factors and heavy-heavy-light three-point functions at weak coupling. *J. High Energy Phys.* **09**, 125 (2015). [https://doi.org/10.1007/JHEP09\(2015\)125](https://doi.org/10.1007/JHEP09(2015)125)
7. Jiang, Y.: Diagonal form factors and hexagon form factors II. Non-BPS light operator. *J. High Energy Phys.* **01**, 021 (2017). [https://doi.org/10.1007/JHEP01\(2017\)021](https://doi.org/10.1007/JHEP01(2017)021)
8. Jiang, Y., Petrovskii, A.: Diagonal form factors and hexagon form factors. *J. High Energy Phys.* **07**, 120 (2016). [https://doi.org/10.1007/JHEP07\(2016\)120](https://doi.org/10.1007/JHEP07(2016)120)
9. Leclair, A., Mussardo, G.: Finite temperature correlation functions in integrable QFT. *Nucl. Phys.* **B552**, 624–642 (1999). [https://doi.org/10.1016/S0550-3213\(99\)00280-1](https://doi.org/10.1016/S0550-3213(99)00280-1)
10. Luscher, M.: Volume dependence of the energy spectrum in massive quantum field theories. 1. Stable particle states. *Commun. Math. Phys.* **104**, 177 (1986). <https://doi.org/10.1007/BF01211589>
11. Luscher, M.: Volume dependence of the energy spectrum in massive quantum field theories. 2. Scattering states. *Commun. Math. Phys.* **105**, 153–188 (1986). <https://doi.org/10.1007/BF01211097>
12. Mussardo, G.: *Statistical Field Theory*. Oxford University Press, New York (2010)
13. Pozsgay, B.: Mean values of local operators in highly excited Bethe states. *J. Stat. Mech.* **1101**, P01011 (2011). <https://doi.org/10.1088/1742-5468/2011/01/P01011>
14. Pozsgay, B.: Form factor approach to diagonal finite volume matrix elements in integrable QFT. *J. High Energy Phys.* **07**, 157 (2013). [https://doi.org/10.1007/JHEP07\(2013\)157](https://doi.org/10.1007/JHEP07(2013)157)
15. Pozsgay, B., Takacs, G.: Form-factors in finite volume I: form-factor bootstrap and truncated conformal space. *Nucl. Phys.* **B788**, 167–208 (2008). <https://doi.org/10.1016/j.nuclphysb.2007.06.027>
16. Pozsgay, B., Takacs, G.: Form factors in finite volume. II. Disconnected terms and finite temperature correlators. *Nucl. Phys.* **B788**, 209–251 (2008). <https://doi.org/10.1016/j.nuclphysb.2007.07.008>
17. Pozsgay, B., Szecsenyi, I.M., Takacs, G.: Exact finite volume expectation values of local operators in excited states. *J. High Energy Phys.* **04**, 023 (2015). [https://doi.org/10.1007/JHEP04\(2015\)023](https://doi.org/10.1007/JHEP04(2015)023)
18. Saleur, H.: A Comment on finite temperature correlations in integrable QFT. *Nucl. Phys.* **B567**, 602–610 (2000). [https://doi.org/10.1016/S0550-3213\(99\)00665-3](https://doi.org/10.1016/S0550-3213(99)00665-3)
19. Smirnov, F.: Form-factors in completely integrable models of quantum field theory. *Adv. Ser. Math. Phys.* **14**, 1–208 (1992)
20. Zamolodchikov, A.B.: Thermodynamic Bethe ansatz in relativistic models. scaling three state Potts and Lee-yang models. *Nucl. Phys.* **B342**, 695–720 (1990). [https://doi.org/10.1016/0550-3213\(90\)90333-9](https://doi.org/10.1016/0550-3213(90)90333-9)
21. Zamolodchikov, A.B., Zamolodchikov, A.B.: Factorized S-matrices in two-dimensions as the exact solutions of certain relativistic quantum field models. *Ann. Phys.* **120**, 253–291 (1979). [https://doi.org/10.1016/0003-4916\(79\)90391-9](https://doi.org/10.1016/0003-4916(79)90391-9)

Narayana Number, Chebyshev Polynomial and Motzkin Path on RNA Abstract Shapes



Sang Kwan Choi, Chaiho Rim, and Hwajin Um

Abstract We consider a certain abstract of RNA secondary structures, which is closely related to so-called RNA shapes. The generating function counting the number of the abstract structures is obtained in three different ways, namely, by means of Narayana numbers, Chebyshev polynomials and Motzkin paths. We show that a combinatorial interpretation on 2-Motzkin paths explains a relation between Motzkin paths and RNA shapes and also provides an identity related to Narayana numbers and Motzkin polynomial coefficients.

1 Introduction

Ribonucleic acid (RNA) is a single stranded molecule with a backbone of nucleotides, each of which has one of the four bases, adenine (A), cytosine (C), guanine (G) and uracil (U). Base pairs are formed intra-molecularly between A-U, G-C or G-U, leading the sequence of bases to form helical regions. The primary structure of a RNA is merely the sequence of bases and its three-dimensional conformation by base pairs is called the tertiary structure. As an intermediate structure between the primary and the tertiary, the secondary structure is a planar structure allowing only nested base pairs. This is easy to see in its diagrammatic representation, see Fig. 1. A sequence of n bases is that of labeled vertices $(1, 2, \dots, n)$ in a horizontal line and base pairs are drawn as arcs in the upper half-plane. The condition of nested base pairs means non-crossing arcs: for two arcs (i, j) and (k, l) where $i < j, k < l$ and $i < k$, either $i < j < k < l$ or

S. K. Choi

Center for Theoretical Physics, College of Physical Science and Technology, Sichuan University, Chengdu, China

e-mail: hermit1231@sogang.ac.kr

C. Rim (✉) · H. Um

Department of Physics, Sogang University, Seoul, South Korea

e-mail: rimpine@sogang.ac.kr; um16@sogang.ac.kr

© Springer Nature Switzerland AG 2019

D. R. Wood et al. (eds.), 2017 *MATRIX Annals*, MATRIX Book Series 2,

https://doi.org/10.1007/978-3-030-04161-8_11

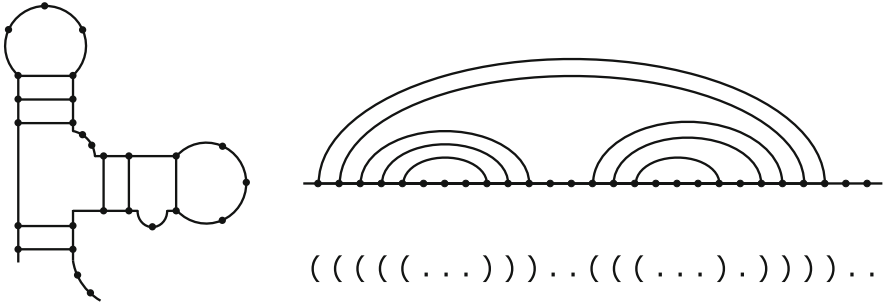


Fig. 1 Representations of secondary structures. The RNA structure on the left hand side is represented as the diagram (top right) and the dot-bracket string (bottom right)

$i < k < l < i$. Since the functional role of a RNA depends mainly on its 3D conformation, prediction of RNA folding from the primary structure has long been an important problem in molecular biology. The most common approach for the prediction is free energy minimization and many algorithms to compute the structures with minimum free energy has been developed (see for instance, [13, 17, 21, 22]).

On the other hand, RNA structures are often considered as combinatorial objects in terms of representations such as strings over finite alphabets, linear trees or the diagrams. Combinatorial approaches enumerate the number of possible structures under various kinds of constraints and observe its statistics to compare with experimental findings [1, 4, 9, 16, 18]. They also provide classifications of structures to advance prediction algorithms [8, 14, 15, 20].

In this paper, we consider a certain abstract of secondary structures under a pure combinatorial point of view regardless of primary structures. The abstract structure is, in fact, closely related to so-called RNA shapes [8, 10, 12], see Sect. 3. Although we will consider it apart from prediction algorithms, let us review briefly the background to RNA shapes in the context of prediction problem. In free energy minimization scheme, the lowest free energy structures are not necessarily native structures. One needs to search suboptimal foldings in a certain energy bandwidth and, in general, obtains a huge set of suboptimal foldings. RNA shapes classify the foldings according to their structural similarities and provide so-called shape representatives such that native structures can be found among those shape representatives. Consequently, it can greatly narrow down the huge set of suboptimal foldings to probe in order to find native structures.

In the following preliminary, we introduce our combinatorial object, what we call island diagrams and present basic definitions needed to describe the diagrams. In Sect. 2, we find the generating function counting the number of island diagrams in three different ways and through which, one may see the intertwining relations between Narayana numbers, Chebyshev polynomials and Motzkin paths. In particular, we find a combinatorial identity, see Eq. (15), which reproduces the following two identities that Coker provided [5] (see also [3] for a combinatorial

interpretation):

$$\sum_{k=1}^n \frac{1}{n} \binom{n}{k} \binom{n}{k-1} x^{k-1} = \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} C_k \binom{n-1}{2k} x^k (1+x)^{n-2k-1} \tag{1}$$

$$\sum_{k=1}^n \frac{1}{n} \binom{n}{k} \binom{n}{k-1} x^{2(k-1)} (1+x)^{2(n-k)} = \sum_{k=1}^n C_k \binom{n-1}{k-1} x^{k-1} (1+x)^{k-1} \tag{2}$$

where C_k is the Catalan number defined by $C_k = \frac{1}{k+1} \binom{2k}{k}$ for $k \geq 0$. We also provide a combinatorial interpretation on 2-Motzkin paths to explain the identity (15). The interpretation implies the bijection between π -shapes and Motzkin paths which was shown in [7, 11].

1.1 Preliminary

A formal definition of secondary structures is given as follows:

Definition 1 (Waterman [20]) A secondary structure is a vertex-labeled graph on n vertices with an adjacency matrix $A = (a_{ij})$ (whose element $a_{ij} = 1$ if i and j are adjacent, and $a_{ij} = 0$ otherwise with $a_{ii} = 0$) fulfilling the following three conditions:

1. $a_{i,i+1} = 1$ for $1 \leq i \leq n - 1$.
2. For each fixed i , there is at most one $a_{ij} = 1$ where $j \neq i \pm 1$
3. If $a_{ij} = a_{kl} = 1$, where $i < k < j$, then $i \leq l \leq j$.

An edge (i, j) with $|i - j| \neq 1$ is said to be a base pair and a vertex i connected only to $i - 1$ and $i + 1$ is called unpaired. We will call an edge $(i, i + 1)$, $1 \leq i \leq n - 1$, a backbone edge. Note that a base pair between adjacent two vertices is not allowed by definition and the second condition implies non-existence of base triples.

There are many other representations of secondary structures than the diagrammatic representation. In this paper, we often use the so-called dot-bracket representation, see Fig. 1. A secondary structure can be represented as a string S over the alphabet set $\{(,), .\}$ by the following rules [9]:

1. If vertex i is unpaired then $S_i = \text{"."}$.
2. If (i, j) is a base pair and $i < j$ then $S_i = \text{"("}$ and $S_j = \text{"}"}$.

In the following, we present the basic definitions of structure elements needed for our investigations.

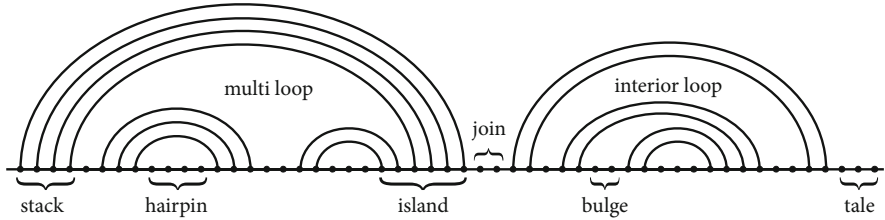


Fig. 2 Structure elements of secondary structures

Definition 2 A secondary structure on $(1, 2, \dots, n)$ consists of the following structure elements (cf. Fig. 2). By a base pair (i, j) , we always assume $i < j$.

1. The sequence of unpaired vertices $(i + 1, i + 2, \dots, j - 1)$ is a *hairpin* if (i, j) is a base pair. The pair (i, j) is said to be the *foundation of the hairpin*.
2. The sequence of unpaired vertices $(i + 1, i + 2, \dots, j - 1)$ is a *bulge* if either (k, j) , $(k + 1, i)$ or $(i, k + 1)$, (j, k) are base pairs.
3. The sequence of unpaired vertices $(i + 1, i + 2, \dots, j - 1)$ is a *join* if (k, i) and (j, l) are base pairs.
4. A *tail* is a sequence of unpaired vertices $(1, 2, \dots, i - 1)$, resp. $(j + 1, j + 2, \dots, n)$ such that i , resp. j is paired.
5. An *interior loop* is two sequences of unpaired vertices $(i + 1, i + 2, \dots, j - 1)$ and $(k + 1, k + 2, \dots, l - 1)$ such that (i, l) and (j, k) are pairs, where $i < j < k < l$.
6. For any $k \geq 3$ and $0 \leq l, m \leq k$ with $l + m = k$, a *multi loop* is l sequences of unpaired vertices and m empty sequences $(i_1 + 1, \dots, j_1 - 1)$, $(i_2 + 1, \dots, j_2 - 1)$, \dots , $(i_k + 1, \dots, j_k - 1)$ such that (i_1, j_k) , (j_1, i_2) , \dots , (j_{k-1}, i_k) are base pairs. Here, a sequence $(i + 1, \dots, j - 1)$ is an empty sequence if $i + 1 = j$.
7. A *stack (or stem)* consists of uninterrupted base pairs $(i + 1, j - 1)$, $(i + 2, j - 2)$, \dots , $(i + k, j - k)$ such that neither (i, j) nor $(i + k + 1, j - k - 1)$ is a base pair. Here the *length* of the stack is k .

Note that, while other structure elements consist of at least one vertex, a multiloop does not necessarily have a vertex. In the diagrammatic representation, a multiloop is a structure bounded by three or more base pairs and backbone edges.

Definition 3 An *island* is a sequence of paired vertices $(i, i + 1, \dots, j)$ such that

1. $i - 1$ and $j + 1$ are both unpaired, where $1 < i \leq j < n$.
2. $j + 1$ is unpaired, where $i = 1$ and $1 < j < n$.
3. $i - 1$ is unpaired, where $1 < i < n$ and $j = n$.

Now we introduce the abstract structures to consider throughout this paper. From here on, we will call the structures *island diagrams* for convenience. An island diagram (cf. Fig. 3) is obtained from secondary structures by

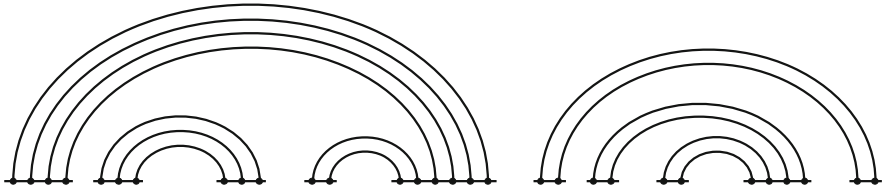


Fig. 3 An example of island diagrams. This island diagram is the abstract structure of the secondary structure given in Fig. 2

1. Removing tails.
2. Representing a sequence of consecutive unpaired vertices between two islands by a single blank.

Accordingly, we retain unpaired regions except for tails but do not account for the number of unpaired vertices. In terms of the dot-bracket representation, we shall use the underscore “_” for the blank: for example, the island diagram “(((_)_)” abstracts the secondary structure “((.)”. Since the abstraction preserves all the structure elements (except for tails) in definition 7, we will use them to describe island diagrams in such a way that, for instance, the blank is a hairpin if its left and right vertices are paired to each other.

2 Generating Function

We enumerate the number of island diagrams $g(h, I, \ell)$, filtered by the number of hairpins(h), islands(I) and basepairs(ℓ). Let $G(x, y, z) = \sum_{h,I,\ell} g(h, I, \ell)x^h y^I z^\ell$ denotes the corresponding generating function. We obtain the generating function in three different ways, by means of Narayana numbers, Chebyshev polynomials and Motzkin paths. In particular, we provide a bijection map between 2-Motzkin paths and sequences of matching brackets.

2.1 Narayana Number

The easiest way to obtain the generating function $G(x, y, z)$ is to use a combinatorial interpretation of the Narayana numbers, which are defined by

$$N(n, k) = \frac{1}{n} \binom{n}{k} \binom{n}{k-1}, \quad 1 \leq k \leq n. \tag{3}$$

The Narayana number $N(n, k)$ counts the number of ways arranging n pairs of brackets to be correctly matched and contain k pairs as “()”. For instance, the bracket

representations for $N(4, 2) = 6$ are given as follows:

$$((()())) \quad ((()())) \quad ((())()) \quad ()((())) \quad (())()() \quad ((()))()$$

It is easy to recover island diagrams from this representation.

Proposition 1 *The generating function has the form*

$$G(x, y, z) = \sum_{\ell, h} N(\ell, h) x^\ell y^{h+1} (1 + y)^{2\ell-1-h} z^\ell. \tag{4}$$

Its closed form is

$$G(x, y, z) = \left(\frac{y}{1 + y} \right) \frac{1 - A(1 + B) - \sqrt{1 - 2A(1 + B) + A^2(1 - B)^2}}{2A} \tag{5}$$

where $A = z(1 + y)^2$ and $B = xy/(1 + y)$.

Proof One may immediately associate bracket representations of the Narayana numbers with island diagrams. Without regard to underscores, the pair of brackets is associated with the basepair and the sub-pattern “()” corresponds to the foundation of the hairpin. It clearly explains the factor $N(\ell, h)x^\ell y^{h+1}$. Now we consider the insertions of underscores to recover the string representation of island diagrams. Recall that, in secondary structures, a hairpin consists of at least one unpaired vertices. Therefore, the foundation of the hairpin “()” must contain an underscore “()”. The number h of underscores are so inserted that we have the factor y^{h+1} . After the insertion of hairpin underscores, there are $(2\ell - 1 - h)$ places left to possibly insert underscores. The numbers of all possible insertions are summarized by the factor $(1 + y)^{2\ell-1-h}$. The generating function of the Narayana numbers is well-known (see for instance [2]) so that one writes the closed form.

2.2 Chebyshev Polynomial


One can also count the number of island diagrams by using the Chebyshev polynomials of the second kind, which are defined by the recurrence relation:

$$U_0(\xi) = 1, \quad U_1(\xi) = 2\xi, \quad U_{n+1}(\xi) = 2\xi U_n(\xi) - U_{n-1}(\xi). \tag{6}$$

The product of the polynomials expands as

$$U_m(\xi)U_n(\xi) = \sum_{k=0}^n U_{m-n+2k}(\xi) \quad \text{for } n \leq m. \tag{7}$$

The relation between island diagrams and Chebyshev polynomials are based on the Feynman diagram of the Hermitian matrix model, refer to [4]. One may have an insight from the simplest example:

$$U_2 \times U_2 = U_4 + U_2 + U_0$$


The polynomial U_k corresponds to the island with k vertices. The product $U_2 U_2$ expands to U_4 (no basepair), U_2 (one basepair) and U_0 (all vertices are paired). The island diagram is the one associated with U_0 in the expansion of the product. In general, we have the following theorem. See [4] for its proof.

Theorem 1 *Suppose that there exist the number I of islands such that each of which has $k_a \geq 1$ vertices for $a \in \{1, \dots, I\}$. The number of island diagrams one finds by making base pairs is given by*

$$\left\langle \prod_{a=1}^I U_{k_a}, U_0 \right\rangle := \frac{2}{\pi} \int_{-1}^1 \prod_{a=1}^I U_{k_a}(\xi) U_0(\xi) \sqrt{1 - \xi^2} d\xi. \tag{8}$$

where $U_k(\xi)$ is the second kind Chebyshev polynomial of degree k .

The Chebyshev polynomials of the second kind are orthogonal with respect to the weight $\sqrt{1 - \xi^2}$: $\langle U_m, U_n \rangle = \delta_{m,n}$. Thus, Theorem 1 means that the number of island diagrams is the coefficient of $U_0 = 1$ when the product $\prod_{a=1}^I U_{k_a}(\xi)$ expands to the linear combination of Chebyshev polynomials.

In order to reproduce the generating function given in (4), we need to take the number of hairpins into account as well. Let us first consider the case of island diagrams in which every blank(underscore) is a hairpin. A hairpin is accompanied with the foundation of the hairpin, that is, h basepairs are assigned as the foundations. Since those basepairs are the most nested ones, the number of the island diagrams is simply given by $\langle U_{k_1-1} \prod_{j=2}^h U_{k_j-2} U_{k_{h+1}-1}, U_0 \rangle$. The foundations of the hairpin take one vertex from the outermost islands and take two vertices from the others. In fact, the island diagrams having only hairpins are no different from strings of matching brackets which represents Narayana numbers as shown in the previous subsection. By just putting $(_)$ \rightarrow $()$, we recover the bracket representations. Thus, we have the following corollary:

Corollary 1.1 *For any $\ell \in \mathbb{N}$ and $1 \leq h \leq \ell$,*

$$N(\ell, h) = \sum_{k_1 + \dots + k_{h+1} = 2(\ell - h)} \left\langle \prod_{a=1}^{h+1} U_{k_a}, U_0 \right\rangle \tag{9}$$

where k_a for $a \in \{1, \dots, h + 1\}$ are non-negative integers.

Now we find the generating function $G(x, y, z)$. Note that a basepair must be made across at least one hairpin. Conversely, no basepair can be made amongst consecutive islands that do not have a hairpin inbetween. We regard a group of maximally consecutive islands with no hairpin inbetween as one effective island. Then, a backbone of island diagram can be seen as an alternate arrangement of effective island and hairpin. This is nothing but the case that every blank is a hairpin. One additional thing to consider is the number of ways to make an effective island having k_a vertices out of I_a islands, which is given by $\binom{k_a-1}{I_a-1}$. Therefore, we find

$$g(h, I, \ell) = \sum_{\{k_a, I_a\}} \prod_{a=1}^{h+1} \binom{k_a-1}{I_a-1} \left\langle U_{k_1-1} \prod_{j=2}^h U_{k_j-2} U_{k_{h+1}-1}, U_0 \right\rangle \tag{10}$$

where the summation runs over $k_1 + \dots + k_{h+1} = 2\ell$ and $I_1 + \dots + I_{h+1} = I$. By means of Corollary 1.1, one can obtain the generating function (4).

We mention that one may also find the generating function by direct calculation of the integral in (10). Using the generating function of the Chebyshev polynomial,

$$\sum_{k \geq 0} \sum_{i=0}^k \binom{k}{i} z^{k/2} y^i U_k(\xi) = \frac{1}{1 - 2\sqrt{z}(1+y)\xi + z(1+y)^2}, \tag{11}$$

the integral is calculated to give

$$G(x, y, z) = \sum_h x^h z^h y^{h+1} (1+y)^{h-1} {}_2F_1(h+1, h; 2; z(1+y)^2) \tag{12}$$

where ${}_2F_1(a, b; c; z)$ is the hypergeometric function. One may easily show that ${}_2F_1(h+1, h; 2; z) = \sum_{k \geq 0} N(h+k, h)z^k$ and therefore obtains the generating function (4).

2.3 Motzkin Path

The generating function $G(x, y, z)$ can also be written in terms of Motzkin polynomial coefficients. The Motzkin numbers M_n and the Motzkin polynomial coefficients $M(n, k)$ are defined as

$$M_n = \sum_{k=0}^{\lfloor n/2 \rfloor} M(n, k) \quad \text{where} \quad M(n, k) = \binom{n}{2k} C_k. \tag{13}$$

Let us consider the combinatorial identity in the following theorem. It is easy to prove using the generating function of the Motzkin polynomials:

$$\sum_{\ell \geq 1} \sum_{p=0}^{\lfloor (\ell-1)/2 \rfloor} M(\ell-1, p) A^{\ell-1} B^p = \frac{1 - A - \sqrt{(1-A)^2 - 4A^2B}}{2A^2B}. \tag{14}$$

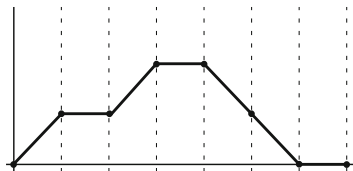
Theorem 2 For any integer $\ell \geq 1$, there holds

$$\begin{aligned} & \frac{y}{1+y} \sum_{h=1}^{\ell} N(\ell, h) (x y)^h (1+y)^{2\ell-h} \\ &= x y^2 \sum_{p=0}^{\lfloor \frac{\ell-1}{2} \rfloor} M(\ell-1, p) (x y (1+y)^3)^p ((1+y)(1+y+x y))^{\ell-2p-1}. \end{aligned} \tag{15}$$

Proof The left hand side is $[z^\ell]G(x, y, z)$ given in (4). Multiplying z^ℓ and taking the summation over ℓ at each side, one can check that the right hand side is indeed the generating function $G(x, y, z)$.

Note that the identity (15) reproduces the Coker’s two identities. When we substitute x/y for x and then put $y = 0$, we get the identity (1). Furthermore, the substitution $x \rightarrow y/(1+y)$ leads to the identity (2).¹

We will investigate how the right hand side in (15) represents island diagrams. In order to do that, we need a combinatorial interpretation of 2-Motzkin paths. Let us first introduce the Motzkin paths, that can also be called 1-Motzkin paths. A Motzkin path of size n is a lattice path starting at $(0, 0)$ and ending at $(n, 0)$ in the integer plane $\mathbb{Z} \times \mathbb{Z}$, which satisfies two conditions: (1) It never passes below the x -axis. (2) Its allowed steps are the up step $(1, 1)$, the down step $(1, -1)$ and the horizontal step $(1, 0)$. We denote by U , D and H an up step, a down step and a horizontal step, respectively. The Motzkin polynomial coefficient $M(n, k)$ is the number of Motzkin paths of size n with k up steps. Since the Motzkin number M_n is given by the sum of $M(n, k)$ over the number of up steps, M_n is the number of Motzkin paths of size n . See for instance, the following figure depicting a Motzkin path of $M(7, 2)$:



¹In order to deduce the identity, one may need the Touchard’s identity [19]: $C_n = \sum_k C_k \binom{n-1}{2k} 2^{n-2k-1}$, which can be also derived from (1) when $x = 1$.

On the other hand, 2-Motzkin paths allow two kinds of horizontal steps, which often distinguish one from another by a color, let us say, R and B denoting a red and a blue step, respectively. We provide a bijection map between 2-Motzkin paths and strings of matching brackets.² Suppose we have a 2-Motzkin path of size n given by a string $q_1 q_2 \cdots q_n$ over the set $\{U, D, R, B\}$. The corresponding string of brackets S_n can be obtained by the following rules:

1. We begin with “()”: Let $S_0 = ()$.
2. For any $1 \leq k \leq n$, suppose there exist a string of brackets S' and a string of matching brackets S'' which are possibly empty such that S_{k-1} has the form $S'(S'')$. Then S_k is given by

$$\begin{aligned}
 S'((S'')()) \text{ if } q_k = U, & \quad S'(S'') \text{ if } q_k = D, \\
 S'(S'')() \text{ if } q_k = R, & \quad S'((S'')) \text{ if } q_k = B.
 \end{aligned}$$

For example, the string of matching brackets corresponding to the 2-Motzkin path $UBURDD$ is obtained as follows:

$$\begin{aligned}
 () \xrightarrow{U} (()) \xrightarrow{B} (())() \xrightarrow{U} (())(()) \\
 \xrightarrow{R} (())(())() \xrightarrow{D} (())(())() \xrightarrow{D} (())(())(())
 \end{aligned}$$

We remark here that only blue steps can make a stack. In other words, directly nested structures such as “(())” never occur without blue steps. Therefore, a 1-Motzkin path can be translated into a string of matching brackets without directly nested brackets. This is one of the 14 interpretations of Motzkin numbers provided by Donaghey and Shapiro in [7]. Later, in [11], it was also shown using context-free grammars in the context of RNA shapes. We also remark that the Motzkin polynomial coefficient $M(\ell - 1, u)$ is the number of ways arranging ℓ pairs of brackets to be correctly matched and contain $\ell - u$ pairs as “(())” with no occurrence of directly nested bracket.

Now we go back to the generating function on the right hand side in (15) and rewrite it as

$$\begin{aligned}
 G(x, y, z) = \sum_{\ell, u} M(\ell - 1, u) (xy^2z) ((1 + y)\sqrt{z})^u (xy(1 + y)z)^u \\
 \times ((1 + y)\sqrt{z})^d ((1 + y)^2z + xy(1 + y)z)^s
 \end{aligned} \tag{16}$$

²Sequences of matching brackets are only Dyck paths. A bijection map between Dyck paths and 2-Motzkin paths was introduced by Delest and Viennot [6]. But here we present a different way of mapping than the well-known one.

where u, d and s stand for the number of up, down and horizontal steps, respectively ($u = d, u + d + s = \ell - 1$). Let us explain each factor in detail by means of the above rules. The term xy^2z is merely the starting hairpin “()” (recall that the exponent of x, y and z are the number of hairpins, islands and basepairs, resp.). At each up step, one has a left bracket and a hairpin to add. For a given non-empty string S of island diagrams, suppose that we add a left bracket then there are the two possibilities, “(S)” and “(S)” corresponding to \sqrt{z} and $y\sqrt{z}$, respectively. Thus, we get the factor $(1 + y)\sqrt{z}$ at every up step and, in the same manner, at every down step. Likewise, adding a hairpin introduces the factor $xy(1 + y)z$ since “S()” and “S()” corresponds to xyz and xy^2z , respectively. On the other hand, a horizontal step can be either R or B . A red step is to add a hairpin and corresponds to $xy(1 + y)z$. A blue step is to add one basepair nesting the string “(S)” and there are three possibilities: the stack “((S))” for z , the two bulges “(S)” and “((S))” for yz and the interior loop “(S)” for y^2z . Therefore, we get $((1 + y)^2z + xy(1 + y)z)$ at each horizontal step.

Note that the number of up steps is the number of multiloops since every up step opens a new multiloop. Thus the generating function written in terms of Motzkin polynomials can be said to classify island diagrams by the number of basepairs and multiloops while the one written in terms of Narayana numbers classify island diagrams by the number of basepairs and hairpins.

3 Single-Stack Diagrams and RNA Shapes

An island diagram is called a single-stack diagram if the length of each stack in the diagram is 1 so that each basepair is a stem by itself. Let $s(h, I, k)$ denotes the number of single-stack diagrams classified by the number of hairpins(h), islands(I) and stems(k) and let $S(x, y, z) = \sum_{h,I,k} s(h, I, k)x^h y^I z^k$ denotes its generating function. The island diagrams with k stems and ℓ basepairs build on the single-stack diagrams with k stems. The number of ways stacking $\ell - k$ basepairs on k stems is $\binom{\ell-1}{k-1}$ and we have

$$g(h, I, \ell) = \sum_{k=1}^{\ell} \binom{\ell - 1}{k - 1} s(h, I, k). \tag{17}$$

Multiplying $x^h y^I z^\ell$ at each side and summing over h, I, ℓ , one finds the relation

$$G(x, y, z) = S\left(x, y, \frac{z}{1 - z}\right) \tag{18}$$

and equivalently, $S(x, y, z) = G(x, y, z/(1+z))$. In terms of Motzkin polynomials, the generating function $S(x, y, z)$ expands to

$$S(x, y, z) = \sum_{k,u} M(\ell - 1, u) (xy^2z) ((1+y)\sqrt{z})^u (xy(1+y)z)^u \times ((1+y)\sqrt{z})^d ((2y+y^2)z + xy(1+y)z)^s \tag{19}$$

where u, d and s stand for the number of up, down and horizontal steps, respectively ($u = d, u + d + s = k - 1$). This is the same as (16) except for one thing. Recall that only blue steps make a directly nested bracket and from which we get three possibilities by putting underscores, i.e., a stack for z , two bulges for yz and an interior loop for y^2z . One obtains single-stack diagrams by getting rid of the possibility of stacking and hence the one different thing is the factor z such that one has $(2y + y^2)z$ instead of $(1 + y)^2z$.

We mention that the single-stack diagram is closely related to the π' -shape (or type 1), which is one of the five RNA abstract shapes provided in [8] classifying secondary structures according to their structural similarities. π' -shape is an abstraction of secondary structures preserving their loop configurations and unpaired regions. A stem is represented as one basepair and a sequence of maximally consecutive unpaired vertices is considered as an unpaired region regardless of the number of unpaired vertices in it. In terms of the dot-bracket representation, a length k stem “ $(^k \dots)^k$ ” is represented by a pair of squared brackets “[\dots]” and an unpaired region is depicted by an underscore. For instance, the π' -shape “ $_ [[[_] _ [_] _]]$ ” can abstract from the secondary structure “ $\dots (((\dots) \dots ((\dots))) \dots)$ ”. The only difference between single-stack diagrams and π' -shapes is whether or not to retain tails.

On the other hand, π -shape (or type 5) ignores unpaired regions such that, for example, the π' -shape “ $_ [[[_] _ [_] _]]$ ” results in the π -shape “[$[\] \]$ ”. Consequently, π -shapes retain only hairpin and multiloop configurations. One may immediately notice that the string representations of π -shapes are nothing but the sequences of matching brackets without directly nested brackets. Therefore, as was shown in the previous section, there is a bijection map between π -shapes and 1-Motzkin paths. Accordingly, one finds the theorem 3.1 in [11] that the number of π -shapes with ℓ pairs of squared brackets is the Motzkin number $M_{\ell-1}$. Furthermore, the Motzkin polynomial coefficient $M(\ell - 1, u)$ is the number of π -shapes with u multiloops and $\ell - u$ hairpins.

Acknowledgements The authors acknowledge the support of this work by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2017R1A2A2A05001164). S.K. Choi is partially supported by the National Natural Science Foundation of China under grant 11575119.

References

1. Barrett, C.L., Li, T.J., Reidys, C.M.: RNA secondary structures having a compatible sequence of certain nucleotide ratios. *J. Comput. Biol.* **23**, 857–873 (2016)
2. Barry, P., Hennessy, A.: A note on Narayana triangles and related polynomials, Riordan arrays, and MIMO capacity calculations. *J. Integer Seq.* **14**(3), 1–26 (2011)
3. Chen, W.Y., Yan, S.H., Yang, L.L.: Identities from weighted Motzkin paths. *Adv. Appl. Math.* **41**(3), 329–334 (2008). <https://doi.org/10.1016/j.aam.2004.11.007>. <http://www.sciencedirect.com/science/article/pii/S0196885808000158>
4. Choi, S.K., Rim, C., Um, H.: RNA substructure as a random matrix ensemble. arXiv:1612.07468 [q-bio.QM]
5. Coker, C.: Enumerating a class of lattice paths. *Discrete Math.* **271**(1), 13–28 (2003). [https://doi.org/10.1016/S0012-365X\(03\)00037-2](https://doi.org/10.1016/S0012-365X(03)00037-2). <http://www.sciencedirect.com/science/article/pii/S0012365X03000372>
6. Delest, M.P., Viennot, G.: Algebraic languages and polyominoes enumeration. *Theor. Comput. Sci.* **34**(1), 169–206 (1984). [https://doi.org/10.1016/0304-3975\(84\)90116-6](https://doi.org/10.1016/0304-3975(84)90116-6). <http://www.sciencedirect.com/science/article/pii/0304397584901166>
7. Donaghey, R., Shapiro, L.W.: Motzkin numbers. *J. Comb. Theory Ser. A* **23**(3), 291–301 (1977). [https://doi.org/10.1016/0097-3165\(77\)90020-6](https://doi.org/10.1016/0097-3165(77)90020-6). <http://www.sciencedirect.com/science/article/pii/0097316577900206>
8. Giegerich, R., Voss, B., Rehmsmeier, M.: Abstract shapes of RNA. *Nucleic Acids Res.* **32**(16), 4843–4851 (2004). <https://doi.org/10.1093/nar/gkh779>
9. Hofacker, I.L., Schuster, P., Stadler, P.F.: Combinatorics of RNA secondary structures. *Discrete Appl. Math.* **88**(1–3), 207–237 (1998)
10. Janssen, S., Reeder, J., Giegerich, R.: Shape based indexing for faster search of RNA family databases. *BMC Bioinformatics* **9**(1), 131 (2008). <https://doi.org/10.1186/1471-2105-9-131>
11. Lorenz, W.A., Ponty, Y., Clote, P.: Asymptotics of RNA shapes. *J. Comput. Biol.* **15**(1), 31–63 (2008)
12. Nebel, M.E., Scheid, A.: On quantitative effects of RNA shape abstraction. *Theory Biosci.* **128**(4), 211–225 (2009). <https://doi.org/10.1007/s12064-009-0074-z>
13. Nussinov, R., Jacobson, A.B.: Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA* **77**(11), 6309–13 (1980). <https://doi.org/10.1073/pnas.77.11.6309>
14. Orland, H., Zee, A.: RNA folding and large N matrix theory. *Nucl. Phys.* **B620**, 456–476 (2002)
15. Reidys, C.M., Huang, F.W.D., Andersen, J.E., Penner, R.C., Stadler, P.F., Nebel, M.E.: Topology and prediction of RNA pseudoknots. *Bioinformatics* **27**(8), 1076–1085 (2011). <https://doi.org/10.1093/bioinformatics/btr090>. <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr090>
16. Schmitt, W.R., Waterman, M.S.: Linear trees and RNA secondary structure. *Discrete Appl. Math.* **51**(3), 317–323 (1994). [https://doi.org/10.1016/0166-218X\(92\)00038-N](https://doi.org/10.1016/0166-218X(92)00038-N). <http://www.sciencedirect.com/science/article/pii/0166218X9200038N>
17. Schuster, P., Stadler, P.F., Renner, A.: RNA structures and folding: from conventional to new issues in structure predictions. *Curr. Opin. Struct. Biol.* **7**(2), 229–235 (1997). [https://doi.org/10.1016/S0959-440X\(97\)80030-9](https://doi.org/10.1016/S0959-440X(97)80030-9). <http://www.sciencedirect.com/science/article/pii/S0959440X97800309>
18. Stein, P., Waterman, M.: On some new sequences generalizing the Catalan and Motzkin numbers. *Discrete Math.* **26**(3), 261–272 (1979). [https://doi.org/10.1016/0012-365X\(79\)90033-5](https://doi.org/10.1016/0012-365X(79)90033-5). <http://www.sciencedirect.com/science/article/pii/0012365X79900335>
19. Touchard, J.: Sur certaines équations fonctionnelles. In: *Proceedings of the International Congress on Mathematics, Toronto, 1924*, vol. 1, pp. 465–472 (1928)
20. Waterman, M.S.: Secondary structure of single-stranded nucleic acids. *Adv. Math. Suppl. Stud.* **1**, 167–212 (1978)

21. Zuker, M., Sankoff, D.: RNA secondary structures and their prediction. *Bull. Math. Biol.* **46**(4), 591–621 (1984). [https://doi.org/10.1016/S0092-8240\(84\)80062-2](https://doi.org/10.1016/S0092-8240(84)80062-2). <http://www.sciencedirect.com/science/article/pii/S0092824084800622>
22. Zuker, M., Stiegler, P.: Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.* **9**(1), 133–148 (1981). <https://doi.org/10.1093/nar/9.1.133>

A Curious Mapping Between Supersymmetric Quantum Chains



Gyorgy Z. Feher, Alexandr Garbali, Jan de Gier, and Kareljan Schoutens

We dedicate this work to our friend and colleague Bernard Nienhuis, on the occasion of his 65-th birthday.

Abstract We present a unitary transformation relating two apparently different supersymmetric lattice models in one dimension. The first (Fendley and Schoutens, *J Stat Mech*, P02017, 2007) describes semionic particles on a 1D ladder, with supersymmetry moving particles between the two legs. The second (de Gier et al., *J Stat Mech*, 023104, 2016) is a fermionic model with particle-hole symmetry and with supersymmetry creating or annihilating pairs of domain walls. The mapping we display features non-trivial phase factors that generalise the sign factors occurring in the Jordan-Wigner transformation.

G. Z. Feher

BME “Momentum” Statistical Field Theory Research Group, Budapest, Hungary
e-mail: g.feher@eik.bme.hu

A. Garbali (✉)

Australian Research Council Centre of Excellence (ACEMS), School of Mathematics and Statistics, The University of Melbourne, VIC, Australia
e-mail: alexandr.garbali@unimelb.edu.au

J. de Gier

Australian Research Council Centre of Excellence (ACEMS), The University of Melbourne, VIC, Australia
e-mail: jdgier@unimelb.edu.au

K. Schoutens

Institute for Theoretical Physics Amsterdam and Delta Institute for Theoretical Physics, University of Amsterdam, Amsterdam, The Netherlands
e-mail: c.j.m.schoutens@uva.nl

1 Introduction

The concept of supersymmetry was conceived in the realm of (high-energy) particle physics, where it expresses a fundamental symmetry between bosonic and fermionic (elementary) particles or excitations in a quantum field theory or string theory. It would seem that in the context of (low-energy) condensed matter systems a similar concept is out of place as superpartners to, say, the electron, if such exist at all, are far out of sight. Nevertheless, we have learned that supersymmetry can be a useful ingredient in relatively simple model systems describing a condensed phase of matter. As soon as the relevant degrees of freedom are not all bosonic, the notion of a fermionic symmetry becomes feasible.

A particularly simple supersymmetric lattice model, commonly referred to as the M_1 model, was proposed in [1]. It features itinerant spin-less fermions on a lattice (or graph), with supersymmetry adding or taking out a single fermion. Denoting the supercharges as Q^\dagger and Q , the Hamiltonian of what is called $\mathcal{N} = 2$ supersymmetric quantum mechanics [2] is defined as

$$H = \{Q^\dagger, Q\}. \quad (1)$$

In the M_1 model the non-trivial nature of H is induced by stipulating that fermions are forbidden to occupy nearest neighbour sites on the lattice. These simple definitions lead to surprisingly rich and diverse phenomena. On a 1D lattice, the M_1 model was found to be critical, and described by the simplest unitary minimal model of $\mathcal{N} = 2$ superconformal field theory [1]. On 2D lattices, there is the remarkable phenomenon of *superfrustration*: an extensive (in the area, that is to say the number of sites) entropy for zero-energy supersymmetric ground states [3–5].

Additional features of the M_1 model in 1D are integrability by Bethe Ansatz and the existence of a mapping to the XXZ model at anisotropy $\Delta = -1/2$ [6]. These features were generalized to a class of models called M_k , where up to k fermions are allowed on consecutive lattice sites [6]. At critical behaviour of these models is captured by the k -th minimal model of $\mathcal{N} = 2$ superconformal field theory, while massive deformations give rise to integrable massive $\mathcal{N} = 2$ QFT's with superpotentials taking the form of Chebyshev polynomials [7].

This paper is concerned with two other, and seemingly different, incarnations of supersymmetry in one spatial dimension. The first is a model, proposed by Fendley and Schoutens (FS) [7], where the supercharges Q and Q^\dagger move particles between two legs of a zig-zag ladder. This would suggest that the particles on the two legs be viewed as bosonic and fermionic, respectively, but the situation in the FS model is different: the phases between the (fermionic) supercharges and the particles are such that the particles on the two legs are naturally viewed as anyons with statistical angle $\pm\pi/2$, that is, as semionic particles. Interestingly, pairs of semions on the two legs can form zero-energy ‘Cooper pairs’ and the model allows multiple supersymmetric groundstates that are entirely made up of such pairs. The FS model is integrable by Bethe Ansatz and has a close-to-free-fermion spectrum: all energies agree with those of free fermions on a single chain, but the degeneracies are different.

The second model we discuss was introduced by Feher, de Gier, Nienhuis and Ruzaczonek (FGNR) [8]. It can be viewed as a particle-hole symmetric version of the M_1 model, where the ‘exclusion’ constraint on the Hilbert space has been relaxed and where the supercharges are now symmetric between particles and holes. In this model fermion number conservation is violated as terms creating and annihilating pairs of fermions are included in the Hamiltonian. The FGNR can be conveniently described in terms of domain walls between particle and hole segments, as the number of such walls is conserved. Also this model allows a Bethe Ansatz solution and the spectrum of the periodic chain has been shown to have characteristic degeneracies. Just as in the FS model, the degeneracies in the FGNR model can be explained by the formation of zero-energy ‘Cooper pairs’.

The sole purpose of this contribution to the 2017 MATRIX Annals is to establish a unitary transformation between the FS and FGNR models on an open chain. Based on the similarity of the Bethe ansatz solutions and that of the physical properties the existence of such a map is not too surprising. Nevertheless, the details are quite intricate. This holds in particular for the phase factors involved in the mapping, which achieve the task of transforming domain walls in the FGNR formulation to particles which, in the FS formulation, are best interpreted as semions. The non-local patterns of the phase factors can be compared to the ‘strings’ of minus signs featuring in the Jordan-Wigner transformation from spins to (spin-less) fermions.

2 Models

In this section, we define the models [8, 9]. We refer to the model of [9] as FS model, and the model of [8] as FGNR model. Both are spinless fermion models on a chain of length L with some boundary conditions. The fermionic creation and annihilation operators c_i, c_i^\dagger ($i = 1, \dots, L$) satisfy the usual anticommutation relations

$$\{c_i^\dagger, c_j\} = \delta_{ij}, \quad \{c_i^\dagger, c_j^\dagger\} = \{c_i, c_j\} = 0. \tag{2}$$

Based on the fermionic creation operators, the on site fermion-number and hole-number operators are defined as

$$n_i = c_i^\dagger c_i \quad p_i = 1 - n_i. \tag{3}$$

These operators act in a fermionic Fock space spanned by ket vectors of the form

$$|\tau\rangle = \prod_{i=1}^L \left(c_i^\dagger\right)^{\tau_i} |\mathbf{0}\rangle, \tag{4}$$

where the product is ordered such that c_i^\dagger with higher i act first. The label $\tau = \{\tau_1, \dots, \tau_L\}$, with $\tau_i = 1$ if there is a fermion at site i and $\tau_i = 0$ if there is a hole.

The vacuum state is defined as usual $c_i |\mathbf{0}\rangle = 0$ for $i = 1, \dots, L$. Both models are supersymmetric chain models where the nilpotent supercharges Q, Q^\dagger are built as sums of local operators.

Originally, the FS model was considered with open boundary conditions [9] and the FGNR model with periodic boundary conditions [8]. In this section we give a short overview of [9] and [8]. In Sect. 3 we restrict ourselves to the open boundary conditions for both models with even L and discuss the mapping between them.

2.1 FS Model Definition

In this section we give a short overview of the FS model [9]. Consider the following supersymmetry generator

$$Q_{FS} = c_2^\dagger c_1 + \sum_{k=1}^{L/2-1} \left(e^{i\frac{\pi}{2}\alpha_{2k-2}} c_{2k-1}^\dagger + e^{i\frac{\pi}{2}\alpha_{2k}} c_{2k+1}^\dagger \right) c_{2k}, \quad (5)$$

where

$$\alpha_k = \sum_{j=1}^k (-1)^j n_j. \quad (6)$$

This supersymmetry generator is nilpotent

$$Q_{FS}^2 = \left(Q_{FS}^\dagger \right)^2 = 0. \quad (7)$$

The Hamiltonian is built up in the usual way

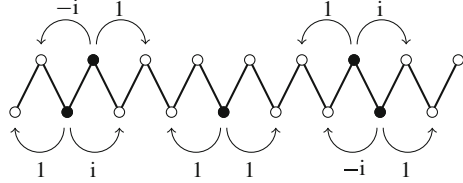
$$\begin{aligned} H_{FS} &= \{Q_{FS}, Q_{FS}^\dagger\} \\ &= \sum_{j=1}^{L-1} (c_{j+1}^\dagger p_j c_{j-1} + c_{j-1}^\dagger p_j c_{j+1} + i c_{j+1}^\dagger n_j c_{j-1} - i c_{j-1}^\dagger n_j c_{j-1}) \\ &\quad - 2 \sum_{j=1}^{L-1} n_j n_{j+1} + 2F_1 + 2F_2 + H_{bdry}, \end{aligned} \quad (8)$$

where

$$F_1 = \sum_{j=1}^{L/2} n_{2j-1}, \quad F_2 = \sum_{j=1}^{L/2} n_{2j}, \quad H_{bdry} = -n_1 - n_L. \quad (9)$$

This model describes an effective one-dimensional model where the fermions are hopping on two chains. Conveniently, these chains are denoted with odd and even

Fig. 1 Statistical interaction between the two chains: Filled dots represent fermions, empty dots empty sites. The hopping amplitudes depend on the occupation of the other chain



site indices and they are coupled in a zig-zag fashion. The F_1 and F_2 operators are counting the fermions on the two chains, and H_{FS} is block diagonal in these operators. The interaction between the chains is statistical: the hopping amplitude picks up an extra i or $-i$ factor, if the fermion “hops over” another fermion on the other chain, see Fig. 1. There is a further attractive interaction between the two chains.

The model is defined on an open chain, where the boundary interaction is encoded in H_{bdry} . The model can be shown to be solvable by nested coordinate Bethe Ansatz [9]. The spectrum of the model is of the same form as for the free model which is defined by

$$H_{free} = 2F_1 + 2F_2 + \sum_{j=2}^{L-1} \left(c_{j+1}^\dagger c_{j-1} + c_{j-1}^\dagger c_{j+1} \right) + H_{bdry}. \quad (10)$$

The eigenenergies of H_{FS} and H_{free} are

$$E = 2 \sum_{a=1}^{f_1+f_2} (1 + \cos(2p_a)), \quad (11)$$

$$p_a = m_a \frac{\pi}{L+1}, \quad m_a \in \{1, 2, \dots, L/2\}, \quad (12)$$

where p_a are called momenta, f_1 and f_2 are the number of fermions on the respective chains, i.e. the eigenvalues of F_1 and F_2 .

The difference between the free and the interacting model is the *degeneracy* of the energy levels. For the free model the Pauli principle is realized by fermions on the same chain not sharing the same momentum. The same momentum can be shared by two fermions on the two chains. Hence for an eigenenergy characterized by the set $\{m_a\}_{a=1}^{f_1+f_2}$ there are $\binom{L/2}{f_1} \binom{L/2}{f_2}$ possible choices, giving the degeneracy for the free model. For the interacting chain instead of thinking in terms of fermions it worth to consider *exclusons* and *Cooper pairs*. Exclusons are fermionic excitations satisfying quantization condition (12) with the further restriction that an exclusion prohibits any other particle to have the same momentum p_a . A pair of fermions located on different chains can form a Cooper pair. In this case, two of the momenta

(say p_1 and p_2) do not satisfy quantization condition (12) and instead they obey

$$\cos^2 p_1 = -\cos^2 p_2. \quad (13)$$

The net energy contribution of the Cooper pair to (11) is zero.

The spectrum of the FS model is built up as follows: there are f_1 and f_2 fermions on the respective chains. Out of these fermions $2C$ form Cooper pairs and $N_1 = f_1 - C$, $N_2 = f_2 - C$ are exclusions on the respective chains. An energy level E is characterized by the quantum numbers $\{m_a\}_{a=1}^{N_1+N_2}$ has the degeneracy

$$d_E = \binom{N_1 + N_2}{N_1} \binom{L/2 - N_1 - N_2}{C}. \quad (14)$$

The first term counts the possible distributions of the exclusions (with fixed $\{m_a\}_{a=1}^{N_1+N_2}$ quantum numbers) on the two chains. The second term counts the degeneracy of the Cooper pairs. The interpretation of the second piece is that the Cooper pairs can be thought of indistinguishable quasiparticles like the exclusions and there is one possible Cooper pair for each allowed momentum. Moreover, the presence of an exclusion with a given momentum prohibits a Cooper pair from occupying the corresponding level. This gives the spectrum and the degeneracy of the FS model. For further details we suggest the original publication [9].

2.2 FGNR Model Definition

In this section we define the FGNR model [8]. We consider a one-dimensional supersymmetric lattice model which is a fermion-hole symmetric extension of the M_1 model of [1]. For this purpose define the operators d_i^\dagger and e_i by

$$d_i^\dagger = p_{i-1} c_i^\dagger p_{i+1}, \quad e_i = n_{i-1} c_i n_{i+1}. \quad (15)$$

Hence d_i^\dagger creates a fermion at position i provided all three of positions $i - 1$, i and $i + 1$ are empty. Similarly, e_i annihilates a fermion at position i provided i and its neighbouring sites are occupied, i.e.

$$\begin{aligned} d_i^\dagger |\tau_1 \dots \tau_{i-2} 000 \tau_{i+2} \dots \tau_L\rangle &= (-1)^{\mathcal{N}_{i-1}} |\tau_1 \dots \tau_{i-2} 010 \tau_{i+2} \dots \tau_L\rangle, \\ e_i |\tau_1 \dots \tau_{i-2} 111 \tau_{i+2} \dots \tau_L\rangle &= (-1)^{\mathcal{N}_{i-1}} |\tau_1 \dots \tau_{i-2} 101 \tau_{i+2} \dots \tau_L\rangle, \end{aligned} \quad (16)$$

while these operators nullify all other states. Here \mathcal{N}_i is the number operator. It counts the number of fermions to the left of site i .

$$\mathcal{N}_i = \sum_{j=1}^i n_j, \quad \mathcal{N}_F = \mathcal{N}_L \quad (17)$$

where \mathcal{N}_F is the total fermion number operator.

We now define the nilpotent supersymmetric generators for the FGNR model

$$Q_{FGNR} = \sum_{i=1}^L (d_i^\dagger + e_i), \quad Q_{FGNR}^2 = 0. \quad (18)$$

The Hamiltonian is defined in the usual way

$$H_{FGNR} = \{Q_{FGNR}^\dagger, Q_{FGNR}\}. \quad (19)$$

The Hamiltonian splits up naturally as a sum of three terms. The first term consists solely of d -type operators, the second solely of e -type operators and the third contains mixed terms.

$$H_{FGNR} = H_I + H_{II} + H_{III}, \quad (20)$$

$$\begin{aligned} H_I &= \sum_i (d_i^\dagger d_i + d_i d_i^\dagger) + \sum_i (d_i^\dagger d_{i+1} + d_{i+1}^\dagger d_i) \\ H_{II} &= \sum_i (e_i e_i^\dagger + e_i^\dagger e_i) + \sum_i (e_i e_{i+1}^\dagger + e_{i+1} e_i^\dagger) \\ H_{III} &= \sum_i (e_i^\dagger d_{i+1}^\dagger + d_{i+1} e_i + e_{i+1}^\dagger d_i^\dagger + d_i e_{i+1}), \end{aligned} \quad (21)$$

where we use periodic boundary conditions

$$c_{i+L}^\dagger = c_i^\dagger. \quad (22)$$

Because the d_i 's and e_i 's are not simple fermion operators, they do not satisfy the canonical anticommutation relations. As a result this bilinear Hamiltonian can not be diagonalized by taking linear combinations of d , e , d^\dagger and e^\dagger .

The term H_I alone is the Hamiltonian of the M_1 model of [1]. The addition of the operator e_i introduces an obvious fermion-hole symmetry $d_i^\dagger \leftrightarrow e_i$ to the model. It turns out that this symmetry results in a surprisingly large degeneracy across the full spectrum of H_{FGNR} .

Note that the Hamiltonians H_I and H_{II} each contain only number operators and hopping terms and thus conserve the total number of fermions. The third Hamiltonian H_{III} breaks this conservation law. For example, the term $e_i^\dagger d_{i+1}^\dagger$ sends the state $|\dots 1000 \dots\rangle$ to $|\dots 1110 \dots\rangle$, thus creating two fermions. Hence the fermion number is not conserved and therefore is not a good quantum number. However, the number of interfaces or domain walls between fermions and holes is conserved and we shall therefore describe our states in terms of these.

2.2.1 Domain Walls

We call an interface between a string of 0's followed by a string of 1's a 01-domain wall and a string of 1's followed by a string of 0's, a 10-domain wall. For example, assuming periodic boundary conditions the configuration

$$000|11|000|1|0000|111|,$$

contains six domain walls, three of each type and starting with a 01-domain wall. Let us consider the effect of various terms appearing in (21). As already discussed in an example above, the terms in H_{III} correspond to hopping of domain walls and map between the following states

$$\left| \dots 1 | 000 \dots \right\rangle \leftrightarrow \left| \dots 111 | 0 \dots \right\rangle, \quad \left| \dots 0 | 111 \dots \right\rangle \leftrightarrow - \left| \dots 000 | 1 \dots \right\rangle, \quad (23)$$

where the minus sign in the second case arises because of the fermionic nature of the model. Hopping of a domain wall always takes place in steps of two hence the parity of the positions of the domain walls is conserved. Aside from their diagonal terms, H_I and H_{II} correspond to hopping of single fermions or holes and therefore to hopping of *pairs* of domain walls. They give rise to transitions between the states

$$\left| \dots 0 | 1 | 00 \dots \right\rangle \leftrightarrow \left| \dots 00 | 1 | 0 \dots \right\rangle, \quad \left| \dots 1 | 0 | 11 \dots \right\rangle \leftrightarrow - \left| \dots 11 | 0 | 1 \dots \right\rangle. \quad (24)$$

Note that in these processes the total parity of positions of interfaces is again conserved, i.e. all processes in H_{FGNR} conserve the number of domain walls at even and odd positions separately.

Finally, the diagonal term $\sum_i (d_i^\dagger d_i + d_i d_i^\dagger + e_i^\dagger e_i + e_i e_i^\dagger)$ in H_I and H_{II} counts the number of 010, 000, 111 and 101 configurations. In other words they count the number of pairs of second neighbour sites that are both empty or both occupied

$$\sum_i (d_i^\dagger d_i + d_i d_i^\dagger + e_i^\dagger e_i + e_i e_i^\dagger) = \sum_i (p_{i-1} p_{i+1} + n_{i-1} n_{i+1}). \quad (25)$$

This is equivalent to counting the total number of sites minus twice the number of domain walls that do not separate a single fermion or hole, i.e. twice the number of well separated domain walls.

Since the number of odd and even domain walls is conserved the Hilbert space naturally breaks into sectors labeled by (m, k) , where m is the total number of domain walls, and k is the number of odd domain walls. Due to periodic boundary conditions m is even.

2.2.2 Solution of the FGNR Model by Bethe ansatz

The FGNR model with periodic boundary conditions is solved by coordinate Bethe Ansatz [8]. There are two kinds of conserved particles (even and odd domain walls), and hence the model is solved by a nested version of the Bethe Ansatz. A solution in the (m, k) sector (with $m - k$ even and k odd domain walls) is characterized by the Bethe roots $\{z_1, \dots, z_m; u_1, \dots, u_k\}$. In other words, z type Bethe roots are associated with both kinds of domain walls, and u type of Bethe roots are associated with odd domain walls. The complex numbers $\{z_1, \dots, z_m; u_1, \dots, u_k\}$ satisfy the following Bethe equations,

$$z_j^L = \pm i^{-L/2} \prod_{l=1}^k \frac{u_l - (z_j - 1/z_j)^2}{u_l + (z_j - 1/z_j)^2}, \quad j = 1, \dots, m \quad (m \in 2\mathbb{N}), \quad (26)$$

$$1 = \prod_{j=1}^m \frac{u_l - (z_j - 1/z_j)^2}{u_l + (z_j - 1/z_j)^2}, \quad l = 1, \dots, k, \quad (27)$$

where the \pm is the same for all j . Solutions corresponding to a nonzero Bethe vector are such that

$$\begin{aligned} z_i^2 &\neq z_j^2, \quad \forall i, j \in \{1, \dots, m\}, \\ u_i &\neq u_j, \quad \forall i, j \in \{1, \dots, k\}. \end{aligned} \quad (28)$$

Two solutions $\{z_1, \dots, z_m; u_1, \dots, u_k\}$, $\{z'_1, \dots, z'_m; u'_1, \dots, u'_k\}$ lead to the same Bethe vector if there exist two permutations $\pi \in S_m$ and $\sigma \in S_k$ and a set of \mathbb{Z}_2 numbers $\{\epsilon_i\}_{i=1}^m$ (i.e. $\epsilon_j = \pm 1$) such that $z_j = \epsilon_j z'_{\pi(j)}$ and $u_l = u'_{\sigma(l)}$. The eigenenergy corresponding to the Bethe roots $\{z_1, \dots, z_m; u_1, \dots, u_k\}$ is

$$\Lambda = L + \sum_{i=1}^m (z_i^2 + z_i^{-2} - 2), \quad (29)$$

which in fact depends only on the non-nested z Bethe roots. The Bethe equations have free fermionic solutions in the following cases. When $k = 0$ there are no Bethe equations at the nested level and the first set of equations simplifies to the free fermionic case. When $k = 1$ the $u = 0$ solutions give the free fermionic part of the spectrum. It is worth to note that the spectrum of the FGNR model with periodic boundary conditions does have a non free fermionic part. This part will not transfer to the open boundary case.

2.2.3 Cooper Pairs

Consider a free fermionic solution

$$z_j = i^{-1/2} e^{2i\pi I_j/L}, \quad j = 1, \dots, m, \quad (30)$$

where I_j is a (half-)integer. This solves the Bethe equations for the $k = 0$ case, or the $k = 1$ case with $u = 0$. This same solution can be used to find a solution in the sector with two more odd domain walls (with $k = 2$ and 3 in the respective cases). Consider the $k = 2$ case. Bethe equations (26) with $k = 2$ are solved by (30) if the nested Bethe roots u_1, u_2 satisfy

$$u_2 = -u_1 \quad (31)$$

as in this case the two new terms in the first Bethe equations cancel each other

$$\begin{aligned} \frac{u_1 - (z_j - 1/z_j)^2}{u_1 + (z_j - 1/z_j)^2} \cdot \frac{u_2 - (z_j - 1/z_j)^2}{u_2 + (z_j - 1/z_j)^2} = \\ \frac{u_1 - (z_j - 1/z_j)^2}{u_1 + (z_j - 1/z_j)^2} \cdot \frac{u_1 + (z_j - 1/z_j)^2}{u_1 - (z_j - 1/z_j)^2} = 1. \end{aligned} \quad (32)$$

Hence the first Bethe equations (26) with solution (30) serve as a consistency condition for u_1, u_2

$$1 = \prod_{j=1}^m \frac{u_l - (z_j - 1/z_j)^2}{u_l + (z_j - 1/z_j)^2}, \quad l = 1, 2. \quad (33)$$

For a free fermionic solution there are always (purely imaginary) $u_2 = -u_1$ type nested roots. As this solution has the same z type roots, a solution with the Cooper pair has the same energy Λ as the original one. We can continue like this introducing new Cooper pairs. The creation of Cooper pairs is limited by the number of domain walls. For further details we suggest [8].

3 Mapping Between the Domain Wall and Particle Representations

The definition of the open boundary version of the FGNR model is straightforward. In the open boundary version with a system size L the operators d_i, d_i^\dagger and e_i^\dagger, e_i are defined for $i = 1, \dots, L - 1$. When $i = 1$ we need to introduce the extra 0-th site

and fix its state to be empty. With these definitions the open boundary supercharge

$$Q_{FGNR}^{(OBC)} = \sum_{i=1}^{L-1} (d_i^\dagger + e_i) \quad (34)$$

is well defined and nilpotent. In this section we would like to lighten the notation for the supercharges and work with their conjugated counterparts. We also need to introduce an intermediate model given in terms of the operators

$$g_i = p_{i+1} c_i n_{i-1}, \quad f_i = n_{i+1} c_i p_{i-1}.$$

Fix L to be even. We have the following supercharges

$$Q_L^\dagger = (Q_{FGNR}^{(OBC)})^\dagger = \sum_{i=1}^{L-1} (d_i + e_i^\dagger), \quad (35)$$

$$\tilde{Q}_L^\dagger = Q_{FS}^\dagger = c_1^\dagger c_2 + \sum_{i=1}^{L/2-1} c_{2i+1}^\dagger (c_{2i} e^{i\alpha_{2i-1}\pi/2} + c_{2i+2} e^{i\alpha_{2i}\pi/2}), \quad (36)$$

We also define two additional supercharges $Q_{e,L}^\dagger$ and $Q_{o,L}^\dagger$ which represent hopping of domain walls and are required as an intermediate step of the mapping,

$$Q_{e,L}^\dagger = \sum_{i=1}^{L/2-1} (g_{2i} + f_{2i}^\dagger), \quad (37)$$

$$Q_{o,L}^\dagger = g_1 + f_1^\dagger + \sum_{i=1}^{L/2-1} (g_{2i+1} + f_{2i+1}^\dagger), \quad (38)$$

Notice that the first terms in the summations in the charges Q_L^\dagger and $Q_{o,L}^\dagger$ contain n_0 . As mentioned above, we need to fix the 0-th site to be unoccupied. Hence the eigenvalue of n_0 is always 0.

We would like to find the map between Q_L^\dagger and \tilde{Q}_L^\dagger . Assume that it is given by a transformation T ,

$$\tilde{Q}_L^\dagger = T Q_L^\dagger T^\dagger, \quad T = P \Gamma M, \quad (39)$$

which itself consists of three terms. The first operator M turns creation and annihilation of domain walls into hopping of domain walls. The second operator Γ turns domain walls into particles and the third operator P fixes the phase factors such that they match (36).

Now we turn to the discussion of the transformations M , Γ and P .

3.1 Transformation M

The first term M translates to the dynamics of domain walls, i.e. it transforms linear combinations of the operators d, d^\dagger and e, e^\dagger into linear combinations of f, f^\dagger and g, g^\dagger (see [8] for more details). More precisely:

$$M = \prod_{i=0}^{\lfloor (L-1)/4 \rfloor} (c_{4i+1} - c_{4i+1}^\dagger)(c_{4i+2} - c_{4i+2}^\dagger), \quad (40)$$

and for all even i we have

$$M(d_i + e_i^\dagger)M^\dagger = f_i + g_i^\dagger, \quad M(d_{i+1} + e_{i+1}^\dagger)M^\dagger = f_{i+1}^\dagger + g_{i+1}. \quad (41)$$

In other words M turns Q^\dagger into a combination of Q_e and Q_o^\dagger

$$MQ_L^\dagger M^\dagger = Q_{e,L} + Q_{o,L}^\dagger. \quad (42)$$

Thus we get an intermediate model.

3.2 Transformation Γ

The next transformation Γ turns domain walls into particles. In fact, there is a family of such transformations. We select Γ to be of the following form

$$\Gamma_L = \prod_{i=1}^{L-1} \left(p_i + n_i(c_{i+1}^\dagger + c_{i+1}) \right). \quad (43)$$

This operator satisfies $\Gamma\Gamma^\dagger = 1$ and transforms the monomials in c_i and c_i^\dagger of (42) into those of \tilde{Q}^\dagger (36). More precisely, conjugation by Γ has the following effect

$$\Gamma_L(Q_{e,L} + Q_{o,L}^\dagger)\Gamma_L^\dagger = \hat{Q}_L^\dagger, \quad (44)$$

where the new supercharge \hat{Q}_L^\dagger differs from \tilde{Q}^\dagger only by phase factors. Let us take Γ and act termwise on $Q_e + Q_o^\dagger$, i.e. on the combination

$$f_i + g_i^\dagger + f_{i+1}^\dagger + g_{i+1},$$

for the labels $i > 0$ and separately on the first term $f_1^\dagger + g_1$. We find

$$\Gamma_L \left(f_1^\dagger + g_1 \right) \Gamma_L^\dagger = -c_1^\dagger c_2 e^{i\pi \sum_{j=1}^{L/2-1} n_{2j+1}}, \quad (45)$$

and

$$\begin{aligned} \Gamma_L \left(f_i + g_i^\dagger + f_{i+1}^\dagger + g_{i+1} \right) \Gamma_L^\dagger \\ = \left(c_{2i+1}^\dagger c_{2i} e^{i\pi \sum_{j=1}^{2i-1} n_j} - c_{2i+1}^\dagger c_{2i+2} \right) e^{i\pi \sum_{j=i+1}^{L/2-1} n_{2j+1}}. \end{aligned} \quad (46)$$

Therefore \hat{Q}_L^\dagger as defined in (44) becomes

$$\begin{aligned} \hat{Q}_L^\dagger = -c_1^\dagger c_2 e^{i\pi \sum_{j=1}^{L/2-1} n_{2j+1}} \\ + \sum_{i=1}^{L/2-1} \left(c_{2i+1}^\dagger c_{2i} e^{i\pi \sum_{j=1}^{2i-1} n_j} - c_{2i+1}^\dagger c_{2i+2} \right) e^{i\pi \sum_{j=i+1}^{L/2-1} n_{2j+1}}. \end{aligned} \quad (47)$$

This agrees with (36) up to the phase factors.

3.3 Transformation P

Let $p(v_1, v_2, \dots, v_L)$ be an unknown function of a binary string, $v_i = 0, 1$. Write a generic phase factor transformation P_L

$$P_L = e^{i\frac{\pi}{2} \hat{p}(n_1, n_2, \dots, n_L)}, \quad (48)$$

and the function p introduced above denotes the eigenvalue of \hat{p} on the state $|v_1, \dots, v_L\rangle$. The commutation relations

$$\hat{p}(n_1, \dots, n_i, \dots, n_L) c_i = c_i \hat{p}(n_1, \dots, n_i - 1, \dots, n_L), \quad (49)$$

$$\hat{p}(n_1, \dots, n_i, \dots, n_L) c_i^\dagger = c_i^\dagger \hat{p}(n_1, \dots, n_i + 1, \dots, n_L), \quad (50)$$

hold on all states of the Hilbert space. Let us find P such that

$$\tilde{Q}_L^\dagger = P_L \hat{Q}_L^\dagger P_L^{-1}. \quad (51)$$

Commuting P_L through each monomial of \hat{Q}_L^\dagger and comparing it with the corresponding monomial of \tilde{Q}_L^\dagger we find $L - 1$ conditions. Commuting P_L with the first monomial in (47) and acting on the state $|v_1, \dots, v_L\rangle$ leads to the first condition

$$p(v_1 + 1, v_2 - 1, \dots, v_L) - p(v_1, v_2, \dots, v_L) + 2 \sum_{j=1}^{L/2-1} v_{2j+1} + 2 = 0. \quad (52)$$

Commuting P_L with the two bulk terms in (47) leads to

$$p(v_1, \dots, v_{2i} - 1, v_{2i+1} + 1, \dots, v_L) - p(v_1, \dots, v_L) + 2 \sum_{j=1}^{2i-1} v_j + 2 \sum_{j=i+1}^{L/2-1} v_{2j+1} = \sum_{j=1}^{2i-1} (-1)^j v_j, \quad (53)$$

and

$$p(v_1, \dots, v_{2i+1} + 1, v_{2i+2} - 1, \dots, v_L) - p(v_1, \dots, v_L) + 2 \sum_{j=i+1}^{L/2-1} v_{2j+1} + 2 = \sum_{j=1}^{2i} (-1)^j v_j. \quad (54)$$

The second equation here with $i = 0$ reproduces (52). In the second equation we can replace v_{2i+1} with $v_{2i+1} - 1$ and v_{2i+2} with $v_{2i+2} + 1$. As a result it becomes of the same form as the first one. Hence (53) and (54) together define L equations for $k = 0, \dots, L - 1$, where for even k one uses (54) and for odd k one uses (53). These equations are valid modulo 4 and can be further simplified

$$p(v_1, \dots, v_{2i} - 1, v_{2i+1} + 1, \dots, v_L) - p(v_1, \dots, v_L) = \sum_{j=1}^{2i-1} (-1)^{j+1} v_j + 2 \sum_{j=i+1}^{L/2-1} v_{2j+1}, \quad (55)$$

and

$$p(v_1, \dots, v_{2i+1} - 1, v_{2i+2} + 1, \dots, v_L) - p(v_1, \dots, v_L) = \sum_{j=1}^{2i} (-1)^{j+1} v_j + 2 \sum_{j=i+1}^{L/2-1} v_{2j+1} + 2. \quad (56)$$

These two equations can be united into one equation using one index k which can be odd or even

$$p(v_1, \dots, v_k, v_{k+1}, \dots, v_L) - p(v_1, \dots, v_k + 1, v_{k+1} - 1, \dots, v_L) = w_k(v_1, \dots, v_{2N}), \quad (57)$$

where the right hand side is given by

$$w_k(v_1, \dots, v_L) = (1 - (-1)^k) + \sum_{j=1}^{k-1} (-1)^{j+1} v_j + \sum_{j=k+2}^{L-1} (1 - (-1)^j) v_j. \quad (58)$$

Therefore we find a set of recurrence relations for the functions p . Note that for a given configuration with the binary string (v_1, \dots, v_L) these equations are assumed to hold for those values of k for which $v_k = 0$ and $v_{k+1} = 1$. Hence the number of such equations is equal to the number of domain walls of type 01.

3.4 Particle Position Coordinates

It is more natural to solve Eq. (57) in a basis where the vectors are labelled using particle positions x_k . The Hilbert spaces in both models are given by vectors labelled by strings of L numbers $\tau_i = 0, 1$

$$|\tau_1, \dots, \tau_L\rangle_n = |\boldsymbol{\tau}\rangle = \prod_{i=1}^L \left(c_i^\dagger \right)^{\tau_i} |\mathbf{0}\rangle. \quad (59)$$

We attached the subscript n in the above notation in order to distinguish it from another labelling of the vectors in the same Hilbert space. Let m be the number of particles in the system. Let us introduce a basis labelled by the positions of the particles and let ρ denote the mapping between the two labellings

$$\rho : |v_1, \dots, v_L\rangle_n \mapsto |x_1, \dots, x_m\rangle_x. \quad (60)$$

The numbers x_k are the eigenvalues of the operators \hat{x}_k which coincide with the eigenvalues of the operators jn_j with $j = x_k$.

Fix m to be the total number of particles in the system and define two functions \tilde{p} and \tilde{w} using the mapping ρ

$$\begin{aligned} \rho(\hat{p}(n_1, \dots, n_L) |v_1, \dots, v_L\rangle_n) &= \tilde{p}(x_1, \dots, x_m) |x_1, \dots, x_m\rangle_x, \\ \rho(\hat{w}_k(n_1, \dots, n_L) |v_1, \dots, v_L\rangle_n) &= \tilde{w}_k(x_1, \dots, x_m) |x_1, \dots, x_m\rangle_x, \end{aligned}$$

with \hat{w}_k being the diagonal operator with the eigenvalues (58). We can now rewrite (57) and (58) in the particle position basis

$$\tilde{p}(x_1, \dots, x_j, \dots, x_m) - \tilde{p}(x_1, \dots, x_j - 1, \dots, x_m) = \tilde{w}_j(x_1, \dots, x_m), \quad (61)$$

with

$$\tilde{w}_k(x_1, \dots, x_m) = (1 + (-1)^{x_k}) - \sum_{i=1}^{k-1} (-1)^{x_i} + \sum_{i=k+1}^m (1 - (-1)^{x_i}). \quad (62)$$

Once again this equation is considered to hold for j such that $x_j - x_{j-1} > 1$. The generic solution of (61) is

$$\tilde{p}(x_1, \dots, x_m) = \tilde{p}(1, 2, \dots, m) + \sum_{k=1}^m \sum_{i=k+1}^{x_k} \tilde{w}_k(1, 2, \dots, k-1, i, x_{k+1}, \dots, x_m), \quad (63)$$

which can be checked by a direct calculation. Here $\tilde{p}(1, 2, \dots, m)$ is the initial condition and can be chosen to be 0. Inserting \tilde{w}_k we get

$$\tilde{p}(x_1, \dots, x_m) = \sum_{k=1}^m \sum_{i=k+1}^{x_k} \left(1 + (-1)^i - \sum_{j=1}^k (-1)^j + \sum_{j=k+1}^m (1 - (-1)^{x_j}) \right). \quad (64)$$

The required phase transformation takes the form

$$P_L = e^{i\frac{\pi}{2}(\tilde{p}(\hat{x}_1, \dots, \hat{x}_m) \circ \rho)}. \quad (65)$$

3.5 Examples

To illustrate the mapping, we show some examples of corresponding states between the FS model (zig-zag ladder) and FGNR model states (up to phase factors $\pm 1, \pm i$),

empty ladder :	$ 0000\ 0000\ 0000\rangle_{FS}$	\leftrightarrow	$0 1100\ 1100\ 1100\rangle_{FGNR}$
single FS semion :	$ 0000\ 1000\ 0000\rangle_{FS}$	\leftrightarrow	$0 1100\ 0011\ 0011\rangle_{FGNR}$
single FS pair :	$ 0000\ 1100\ 0000\rangle_{FS}$	\leftrightarrow	$0 1100\ 0100\ 1100\rangle_{FGNR}$
lower leg filled :	$ 1010\ 1010\ 1010\rangle_{FS}$	\leftrightarrow	$0 0000\ 0000\ 0000\rangle_{FGNR}$
single FGNR particle :	$ 1010\ 0110\ 1010\rangle_{FS}$	\leftrightarrow	$0 0000\ 1000\ 0000\rangle_{FGNR}$
upper leg filled :	$ 0101\ 0101\ 0101\rangle_{FS}$	\leftrightarrow	$0 1010\ 1010\ 1010\rangle_{FGNR}$
upper leg plus semion :	$ 1101\ 0101\ 0101\rangle_{FS}$	\leftrightarrow	$0 0101\ 0101\ 0101\rangle_{FGNR}$
filled ladder :	$ 1111\ 1111\ 1111\rangle_{FS}$	\leftrightarrow	$0 0110\ 0110\ 0110\rangle_{FGNR}$

For $L = 4$ the phase factors take the explicit values

$$\begin{aligned}
 \tilde{p}(1) &= 0, \quad \tilde{p}(2) = 2, \quad \tilde{p}(3) = 0, \quad \tilde{p}(4) = 0 \\
 \tilde{p}(1, 2) &= 0, \quad \tilde{p}(1, 3) = 1, \quad \tilde{p}(1, 4) = 0, \quad \tilde{p}(2, 3) = 1, \quad \tilde{p}(2, 4) = 2, \quad \tilde{p}(3, 4) = 2 \\
 \tilde{p}(1, 2, 3) &= 0, \quad \tilde{p}(1, 2, 4) = 2, \quad \tilde{p}(1, 3, 4) = 3, \quad \tilde{p}(2, 3, 4) = 3 \\
 \tilde{p}(1, 2, 3, 4) &= 0
 \end{aligned} \tag{66}$$

Finally we provide an explicit example of all the steps in the mapping. Let us act with both sides of (39) on the state $|010101\rangle$

$$\tilde{Q}_6^\dagger |010101\rangle = P\Gamma M Q_6^\dagger M^\dagger \Gamma^\dagger P^{-1} |010101\rangle,$$

The action of \tilde{Q}_6^\dagger results in

$$\tilde{Q}_6^\dagger |010101\rangle = |001101\rangle + i|010011\rangle - |010110\rangle + i|011001\rangle + |100101\rangle,$$

and the right hand side is computed as follows

$$\begin{aligned}
 P\Gamma M Q_6^\dagger M^\dagger \Gamma^\dagger P^{-1} |010101\rangle &= -P\Gamma M Q_6^\dagger M^\dagger \Gamma^\dagger |010101\rangle \\
 &= P\Gamma M Q_6^\dagger M^\dagger |011001\rangle = P\Gamma M Q_6^\dagger |101010\rangle \\
 &= P\Gamma M (|001010\rangle - |100010\rangle + |101000\rangle + |101110\rangle - |111010\rangle) \\
 &= P\Gamma (|001001\rangle - |010001\rangle + |011011\rangle + |011101\rangle + |111001\rangle) \\
 &= P (|001001\rangle - |010001\rangle + |011011\rangle + |011101\rangle + |111001\rangle) \\
 &= |001101\rangle + i|010011\rangle - |010110\rangle + i|011001\rangle + |100101\rangle.
 \end{aligned}$$

4 Conclusion

We have established a unitary transformation between the FS and FGNR models with open boundary conditions. We are confident that this map will be helpful for unraveling the properties of these highly intriguing models. For example, in the FS formulation it was not clear how to impose periodic boundary conditions without losing the supersymmetry—this issue is now resolved.

It is a pleasure to dedicate this work to Bernard Nienhuis on the occasion of his 65th birthday. Bernard has always had a special eye for ingenious maps relating apparently different models of statistical physics to one another. We can only hope that dedicating this work to him finds some justification in our following a similar strategy for one of the many integrable models that he has pioneered.

Acknowledgements We thank the hospitality of the international mathematical research institute MATRIX where a large part of this work was performed. JdG and AG gratefully thank financial support of the ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS). GZF was supported by the BME-Nanotechnology FIKP grant of EMMI (BME FIKP-NAT) and by the National Research Development and Innovation Office (NKFIH) (KH-17 grant no. 125567).

References

1. Fendley, P., Schoutens, K., de Boer, J.: Lattice models with $\mathcal{N} = 2$ supersymmetry. *Phys. Rev. Lett.* **90**, 120402 (2003). [arXiv/hep-th/0210161](https://arxiv.org/abs/hep-th/0210161)
2. Witten, E.: Constraints on supersymmetry breaking. *Nucl. Phys.* **B 202**, 253–316 (1982)
3. Fendley, P., Schoutens, K.: Exact results for strongly-correlated fermions in 2+1 dimensions. *Phys. Rev. Lett.* **95**, 046403 (2005). [arXiv/cond-mat/0504595](https://arxiv.org/abs/cond-mat/0504595)
4. van Eerten, H.: Extensive ground state entropy in supersymmetric lattice models. *J. Math. Phys.* **46**, 123302 (2005). [arXiv/cond-mat/0509581](https://arxiv.org/abs/cond-mat/0509581)
5. Huijse, L., Schoutens, K.: Supersymmetry, lattice fermions, independence complexes and cohomology theory. *Adv. Theor. Math. Phys.* **14**, 643–694 (2010). [arXiv/0903.0784](https://arxiv.org/abs/0903.0784)
6. Fendley, P., Nienhuis, B., Schoutens, K.: Lattice fermion models with supersymmetry. *J. Phys. A.* **36**, 12399–12424 (2003). [arXiv/cond-mat/0307338](https://arxiv.org/abs/cond-mat/0307338)
7. Fokkema, T., Schoutens, K.: M_k models: the field theory connection. *SciPost Phys.* **3**, 004 (2017). [arXiv/1703.10079](https://arxiv.org/abs/1703.10079)
8. de Gier, J., Feher, G., Nienhuis, B., Rusaczonek, M.: Integrable supersymmetric chain without particle conservation. *J. Stat. Mech.* 023104 (2016). [arXiv/math-ph/1510.02520](https://arxiv.org/abs/math-ph/1510.02520)
9. Fendley, P., Schoutens, K.: Cooper pairs and exclusion statistics from coupled free-fermion chains. *J. Stat. Mech.* P02017 (2007). [arXiv/cond-mat/0612270](https://arxiv.org/abs/cond-mat/0612270)

Remarks on $A_n^{(1)}$ Face Weights



Atsuo Kuniba

Abstract Elementary proofs are presented for the factorization of the elliptic Boltzmann weights of the $A_n^{(1)}$ face model, and for the sum-to-1 property in the trigonometric limit, at a special point of the spectral parameter. They generalize recent results obtained in the context of the corresponding trigonometric vertex model.

1 Introduction

In the recent work [8], the quantum R matrix for the symmetric tensor representation of the Drinfeld-Jimbo quantum affine algebra $U_q(A_n^{(1)})$ was revisited. A new factorized formula at a special value of the spectral parameter and a certain sum rule called sum-to-1 were established. These properties have led to vertex models that can be interpreted as integrable Markov processes on one-dimensional lattice including several examples studied earlier [7, Fig. 1,2]. In this note we report analogous properties of the Boltzmann weights for yet another class of solvable lattice models known as IRF (interaction round face) models [2] or face models for short. More specifically, we consider the elliptic fusion $A_n^{(1)}$ face model corresponding to the symmetric tensor representation [5, 6]. For $n = 1$, it reduces to [1] and [4] when the fusion degree is 1 and general, respectively. There are restricted and unrestricted versions of the model. The trigonometric case of the latter reduces to the $U_q(A_n^{(1)})$ vertex model when the site variables tend to infinity. See Proposition 1. In this sense Theorems 1 and 2 given below, which are concerned with the unrestricted version, provide generalizations of [8, Th. 2] and [8, eq. (30)] so as to include finite site variables (and also to the elliptic case in the former). In Sect. 3 we will also comment on the restricted version and difficulties to associate integrable stochastic models.

A. Kuniba (✉)

Institute of Physics, University of Tokyo, Komaba, Tokyo, Japan

© Springer Nature Switzerland AG 2019

D. R. Wood et al. (eds.), 2017 *MATRIX Annals*, MATRIX Book Series 2,

https://doi.org/10.1007/978-3-030-04161-8_13

185

2 Results

Let $\theta_1(u) = \theta_1(u, p) = 2p^{\frac{1}{4}} \sin \pi u \prod_{k=1}^{\infty} (1 - 2p^{2k} \cos 2\pi u + p^{4k})(1 - p^{2k})$ be one of the Jacobi theta function ($|p| < 1$) enjoying the quasi-periodicity

$$\theta_1(u + 1; e^{\pi i \tau}) = -\theta_1(u; e^{\pi i \tau}), \quad \theta_1(u + \tau; e^{\pi i \tau}) = -e^{-\pi i \tau - 2\pi i u} \theta_1(u; e^{\pi i \tau}), \tag{1}$$

where $\text{Im} \tau > 0$. We set

$$[u] = \theta_1\left(\frac{u}{L}, p\right), \quad [u]_k = [u][u - 1] \cdots [u - k + 1], \quad \begin{bmatrix} u \\ k \end{bmatrix} = \frac{[u]_k}{[k]_k} \quad (k \in \mathbb{Z}_{\geq 0}), \tag{2}$$

with a nonzero parameter L . These are elliptic analogue of the q -factorial and the q -binomial:

$$(z)_m = (z; q)_m = \prod_{i=0}^{m-1} (1 - zq^i), \quad \binom{m}{l}_q = \frac{(q)_m}{(q)_l (q)_{m-l}}.$$

For $\alpha = (\alpha_1, \dots, \alpha_k)$ with any k we write $|\alpha| = \alpha_1 + \dots + \alpha_k$. The relation $\beta \geq \gamma$ or equivalently $\gamma \leq \beta$ means $\beta_i \geq \gamma_i$ for all i .

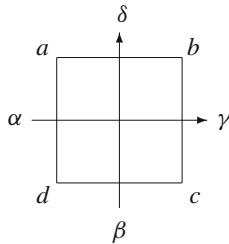
We take the set of local states as $\mathcal{P} = \eta + \mathbb{Z}^{n+1}$ with a generic $\eta \in \mathbb{C}^{n+1}$. Given positive integers l and m , let $a, b, c, d \in \mathcal{P}$ be the elements such that

$$\alpha = d - a \in B_l, \quad \beta = c - d \in B_m, \quad \gamma = c - b \in B_l, \quad \delta = b - a \in B_m, \tag{3}$$

where B_m is defined by

$$B_m = \{\alpha = (\alpha_1, \dots, \alpha_{n+1}) \in \mathbb{Z}_{\geq 0}^{n+1} \mid |\alpha| = m\}. \tag{4}$$

The relations (3) imply $\alpha + \beta = \gamma + \delta$. The situation is summarized as



To the above configuration round a face we assign a function of the spectral parameter u called Boltzmann weight. Its unnormalized version, denoted by $\overline{W}_{l,m} \left(\begin{smallmatrix} a & b \\ d & c \end{smallmatrix} \middle| u \right)$, is constructed from the $l = 1$ case as follows:

$$\overline{W}_{l,m} \left(\begin{smallmatrix} a & b \\ d & c \end{smallmatrix} \middle| u \right) = \sum \prod_{i=0}^{l-1} \overline{W}_{1,m} \left(\begin{smallmatrix} a^{(i)} & b^{(i)} \\ a^{(i+1)} & b^{(i+1)} \end{smallmatrix} \middle| u - i \right), \tag{5}$$

$$\overline{W}_{1,m} \left(\begin{smallmatrix} a & b \\ d & c \end{smallmatrix} \middle| u \right) = \frac{[u + b_v - a_\mu] \prod_{j=1}^{n+1} (j \neq \mu) [b_v - a_j + 1]}{\prod_{j=1}^{n+1} [c_v - b_j]} \quad (d = a + \mathbf{e}_\mu, \quad c = b + \mathbf{e}_v),$$

where $\mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)$. In (5), $a^{(0)}, \dots, a^{(l)} \in \tilde{\mathcal{P}}$ is a path from $a^{(0)} = a$ to $a^{(l)} = d$ such that $a^{(i+1)} - a^{(i)} \in B_1$ ($0 \leq i < l$). The sum is taken over $b^{(1)}, \dots, b^{(l-1)} \in \tilde{\mathcal{P}}$ satisfying the conditions $b^{(i+1)} - b^{(i)} \in B_1$ ($0 \leq i < l$) with $b^{(0)} = b$ and $b^{(l)} = c$. It is independent of the choice of $a^{(1)}, \dots, a^{(l-1)}$ (cf. [4, Fig. 2.4]). We understand that $\overline{W}_{l,m} \left(\begin{smallmatrix} a & b \\ d & c \end{smallmatrix} \middle| u \right) = 0$ unless (3) is satisfied for some $\alpha, \beta, \gamma, \delta$.

The normalized weight is defined by

$$W_{l,m} \left(\begin{smallmatrix} a & b \\ d & c \end{smallmatrix} \middle| u \right) = \overline{W}_{l,m} \left(\begin{smallmatrix} a & b \\ d & c \end{smallmatrix} \middle| u \right) \frac{[1]_l^m}{[l]_l^m}. \tag{6}$$

It satisfies [6] the (unrestricted) star-triangle relation (or dynamical Yang-Baxter equation) [2]:

$$\begin{aligned} & \sum_g W_{k,m} \left(\begin{smallmatrix} a & b \\ f & g \end{smallmatrix} \middle| u \right) W_{l,m} \left(\begin{smallmatrix} f & g \\ e & d \end{smallmatrix} \middle| v \right) W_{k,l} \left(\begin{smallmatrix} b & c \\ g & d \end{smallmatrix} \middle| u - v \right) \\ &= \sum_g W_{k,l} \left(\begin{smallmatrix} a & g \\ f & e \end{smallmatrix} \middle| u - v \right) W_{l,m} \left(\begin{smallmatrix} a & b \\ g & c \end{smallmatrix} \middle| v \right) W_{k,m} \left(\begin{smallmatrix} g & c \\ e & d \end{smallmatrix} \middle| u \right), \end{aligned} \tag{7}$$

where the sum extends over $g \in \tilde{\mathcal{P}}$ giving nonzero weights. Under the same setting (3) as in (6), we introduce the product

$$S_{l,m} \left(\begin{smallmatrix} a & b \\ d & c \end{smallmatrix} \right) = [m]_l^{-1} \prod_{1 \leq i, j \leq n+1} \frac{[c_i - d_j]_{c_i - b_i}}{[c_i - b_j]_{c_i - b_i}}. \tag{8}$$

Note that $S_{l,m} \left(\begin{smallmatrix} a & b \\ d & c \end{smallmatrix} \right) = 0$ unless $d \leq b$ because of the factor $\prod_{i=1}^{n+1} [c_i - d_i]_{c_i - b_i}$. The following result giving an explicit factorized formula of the weight $W_{l,m}$ at special value of the spectral parameter is the elliptic face model analogue of [8, Th. 2].

Theorem 1 *If $l \leq m$, the following equality is valid:*

$$W_{l,m} \left(\begin{matrix} a & b \\ d & c \end{matrix} \middle| u = 0 \right) = S_{l,m} \left(\begin{matrix} a & b \\ d & c \end{matrix} \right). \tag{9}$$

Proof We are to show

$$\bar{W}_{l,m} \left(\begin{matrix} a & b \\ d & c \end{matrix} \middle| 0 \right) = \frac{[l]_l}{[1]^l} \prod_{i,j} \frac{[c_i - d_j]_{c_i - b_i}}{[c_i - b_j]_{c_i - b_i}}. \tag{10}$$

Here and in what follows unless otherwise stated, the sums and products are taken always over $1, \dots, n + 1$ under the condition (if any) written explicitly. We invoke the induction on l . It is straightforward to check (10) for $l = 1$. By the definition (5) the $l + 1$ case is expressed as

$$\bar{W}_{l+1,m} \left(\begin{matrix} a & b \\ d & c \end{matrix} \middle| 0 \right) = \sum_v \bar{W}_{l,m} \left(\begin{matrix} a & b \\ d' & c' \end{matrix} \middle| 0 \right) \bar{W}_{1,m} \left(\begin{matrix} d' & c' \\ d & c \end{matrix} \middle| -l \right) \quad (d' = d - \mathbf{e}_\mu, c' = c - \mathbf{e}_\nu)$$

for some fixed $\mu \in [1, n + 1]$. Due to the induction hypothesis on $\bar{W}_{l,m}$, the equality to be shown becomes

$$\begin{aligned} & \sum_v \frac{[l]_l}{[1]^l} \left(\prod_{i,j} \frac{[c'_i - d'_j]_{c'_i - b_i}}{[c'_i - b_j]_{c'_i - b_i}} \right) \frac{[-l + c'_\nu - d'_\mu] \prod_{k \neq \mu} [c'_\nu - d'_k + 1]}{\prod_k [c_\nu - c'_k]} \\ &= \frac{[l + 1]_{l+1}}{[1]^{l+1}} \prod_{i,j} \frac{[c_i - d_j]_{c_i - b_i}}{[c_i - b_j]_{c_i - b_i}}. \end{aligned} \tag{11}$$

After removing common factors using $c'_i = c_i - \delta_{i\nu}, d'_i = d_i - \delta_{i\mu}$, one finds that (11) is equivalent to

$$\sum_v [c_\nu - d_\mu - l] \prod_{i \neq \nu} \frac{[c_i - d_\mu + 1]}{[c_\nu - c_i]} \prod_j [c_\nu - b_j] = [l + 1] \prod_i [b_i - d_\mu + 1]$$

with l determined by $l + 1 = \sum_j (c_j - b_j)$. One can eliminate d_μ and rescale the variables by $(b_j, c_j) \rightarrow (Lb_j + d_\mu, Lc_j + d_\mu)$ for all j . The resulting equality follows from Lemma 1.

Lemma 1 *Let $b_1, \dots, b_n, c_1, \dots, c_n \in \mathbb{C}$ be generic and set $s = \sum_{i=1}^n (c_i - b_i)$. Then for any $n \in \mathbb{Z}_{\geq 1}$ the following identity holds:*

$$\sum_{i=1}^n \theta_1(z + c_i - s) \prod_{j=1 (j \neq i)}^n \frac{\theta_1(z + c_j)}{\theta_1(c_i - c_j)} \prod_{j=1}^n \theta_1(c_i - b_j) = \theta_1(s) \prod_{i=1}^n \theta_1(z + b_i).$$

Proof Denote the LHS–RHS by $f(z)$. From (1) we see that $f(z)$ satisfies (12) with $B = \frac{n}{2}$, $A_1 = \frac{n(1+\tau)}{2} + \sum_{j=1}^n b_j$ and $A_2 = n$. Moreover it is easily checked that $f(z)$ possesses zeros at $z = -c_1, \dots, -c_n$. Therefore Lemma 2 claims $-(c_1 + \dots + c_n) - (B\tau + \frac{1}{2}A_2 - A_1) \equiv 0 \pmod{\mathbb{Z} + \mathbb{Z}\tau}$. But this gives $s \equiv 0$ which is a contradiction since b_j, c_j can be arbitrary. Therefore $f(z)$ must vanish identically.

Lemma 2 *Let $\text{Im}\tau > 0$. Suppose an entire function $f(z) \not\equiv 0$ satisfies the quasi-periodicity*

$$f(z + 1) = e^{-2\pi i B} f(z), \quad f(z + \tau) = e^{-2\pi i(A_1 + A_2 z)} f(z). \tag{12}$$

Then $A_2 \in \mathbb{Z}_{\geq 0}$ holds and $f(z)$ has exactly A_2 zeros $z_1, \dots, z_{A_2} \pmod{\mathbb{Z} + \mathbb{Z}\tau}$. Moreover $z_1 + \dots + z_{A_2} \equiv B\tau + \frac{1}{2}A_2 - A_1 \pmod{\mathbb{Z} + \mathbb{Z}\tau}$ holds.

Proof Let C be a period rectangle $(\xi, \xi + 1, \xi + 1 + \tau, \xi + \tau)$ on which there is no zero of $f(z)$. From the Cauchy theorem the number of zeros of $f(z)$ in C is equal to $\int_C \frac{f'(z)}{f(z)} \frac{dz}{2\pi i}$. Calculating the integral by using (12) one gets A_2 . The latter assertion can be shown similarly by considering the integral $\int_C \frac{zf'(z)}{f(z)} \frac{dz}{2\pi i}$.

From Theorem 1 and (7) it follows that $S_{l,m} \begin{pmatrix} a & b \\ d & c \end{pmatrix}$ also satisfies the (unrestricted) star-triangle relation (7) without spectral parameter. The discrepancy of the factorizing points $u = 0$ in (9) and “ $u = l - m$ ” in [8, Th. 2] is merely due to a conventional difference in defining the face and the vertex weights.

Since (6) and (8) are homogeneous of degree 0 in the symbol $[\dots]$, the trigonometric limit $p \rightarrow 0$ may be understood as replacing (2) by $[u] = q^{u/2} - q^{-u/2}$ with generic $q = \exp \frac{2\pi i}{L}$. Under this prescription the elliptic binomial $\begin{bmatrix} m \\ l \end{bmatrix}$ from (2) is replaced by $q^{l(l-m)/2} \begin{pmatrix} m \\ l \end{pmatrix}_q$, therefore the trigonometric limit of (8) becomes

$$S_{l,m} \begin{pmatrix} a & b \\ d & c \end{pmatrix}_{\text{trig}} = \begin{pmatrix} m \\ l \end{pmatrix}_q^{-1} \prod_{1 \leq i, j \leq n+1} \frac{(q^{b_i - d_j + 1})_{c_i - b_i}}{(q^{b_i - b_j + 1})_{c_i - b_i}}. \tag{13}$$

The following result is a trigonometric face model analogue of [8, Th. 6].

Theorem 2 *Suppose $l \leq m$. Then the sum-to-1 holds in the trigonometric case:*

$$\sum_b S_{l,m} \begin{pmatrix} a & b \\ d & c \end{pmatrix}_{\text{trig}} = 1, \tag{14}$$

where the sum runs over those b satisfying $c - d \in B_m$ and $d - a \in B_l$.

Proof The relation (14) is equivalent to

$$\begin{pmatrix} m \\ l \end{pmatrix}_q = \sum_{\gamma \in B_l, \gamma \leq \beta} \prod_{1 \leq i, j \leq n+1} \frac{(q^{c_{ij} - \gamma_i + \beta_j + 1})_{\gamma_i}}{(q^{c_{ij} - \gamma_i + \gamma_j + 1})_{\gamma_i}} \quad (c_{ij} = c_i - c_j) \tag{15}$$

for any fixed $\beta = (\beta_1, \dots, \beta_{n+1}) \in B_m, l \leq m$ and the parameters c_1, \dots, c_{n+1} , where the sum is taken over $\gamma \in B_l$ (4) under the constraint $\gamma \leq \beta$. In fact we are going to show

$$\frac{(w_1^{-1} \dots w_n^{-1} q^{-l+1})_l}{(q)_l} = \sum_{|\gamma|=l} \prod_{1 \leq i, j \leq n} \frac{(q^{-\gamma_i+1} z_i / (z_j w_j))_{\gamma_i}}{(q^{\gamma_j-\gamma_i+1} z_i / z_j)_{\gamma_i}} \quad (l \in \mathbb{Z}_{\geq 0}), \tag{16}$$

where the sum is over $\gamma \in \mathbb{Z}_{\geq 0}^n$ such that $|\gamma| = l$, and $w_1, \dots, w_n, z_1, \dots, z_n$ are arbitrary parameters. The relation (15) is deduced from (16) $_{|n \rightarrow n+1}$ by setting $z_i = q^{c_i}, w_i = q^{-\beta_i}$ and specializing β_i 's to nonnegative integers. In particular, the constraint $\gamma \leq \beta$ automatically arises from the $i = j$ factor $\prod_{i=1}^n (q^{-\gamma_i+1+\beta_i})_{\gamma_i}$ in the numerator. To show (16) we rewrite it slightly as

$$q^{\frac{l^2}{2}} \frac{(w_1 \dots w_n)_l}{(q)_l} = \sum_{|\gamma|=l} \prod_{i=1}^n q^{\frac{\gamma_i^2}{2}} \frac{(w_i)_{\gamma_i}}{(q)_{\gamma_i}} \prod_{1 \leq i \neq j \leq n} \frac{(z_j w_j / z_i)_{\gamma_i}}{(q^{-\gamma_j} z_j / z_i)_{\gamma_i}}. \tag{17}$$

Denote the RHS by $F_n(w_1, \dots, w_n | z_1, \dots, z_n)$. We will suppress a part of the arguments when they are kept unchanged in the formulas. It is easy to see

$$F_n(w_1, w_2 | z_1, z_2) = F_n(w_2, w_1 | z_2, z_1) = F_n\left(\frac{z_2 w_2}{z_1}, \frac{z_1 w_1}{z_2} | z_1, z_2\right).$$

Thus the coefficients in the expansion $F_n(w_1, w_2 | z_1, z_2) = \sum_{0 \leq i, j \leq l} C_{i,j}(z_1, z_2) w_1^i w_2^j$ are rational functions in z_1, \dots, z_n obeying $C_{i,j}(z_1, z_2) = C_{j,i}(z_2, z_1) = (\frac{z_1}{z_2})^{i-j} C_{j,i}(z_1, z_2)$. On the other hand from the explicit formula (17), one also finds that any $C_{i,j}(z_1, z_2)$ remains finite in the either limit $\frac{z_1}{z_2}, \frac{z_2}{z_1} \rightarrow \infty$ or $\frac{z_1}{z_2}, \frac{z_2}{z_1} \rightarrow 0$ for $i \geq 3$. It follows that $C_{i,j}(z_1, z_2) = 0$ unless $i = j$, hence

$$F_n(w_1, w_2, \dots, w_n | z_1, \dots, z_n) = F_n(1, w_1 w_2, w_3, \dots, w_n | z_1, \dots, z_n).$$

Moreover it is easily seen

$$F_n(1, w_1 w_2, w_3, \dots, w_n | z_1, z_2, \dots, z_n) = F_{n-1}(w_1 w_2, w_3, \dots, w_n | z_2, \dots, z_n).$$

Repeating this we reach $F_1(w_1 \dots w_n | z_n)$ giving the LHS of (17).

We note that the sum-to-1 (14) does not hold in the elliptic case. Remember that our local states are taken from $\tilde{\mathcal{S}} = \eta + \mathbb{Z}^{n+1}$ with a generic $\eta \in \mathbb{C}^{n+1}$. So we set $a = \eta + \tilde{a}$ with $\tilde{a} \in \mathbb{Z}^{n+1}$ etc. in (4), and assume that it is valid also for $\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}$. It is easy to check

Proposition 1 *Assume $l \leq m$ and $|q| < 1$. Then the following equality holds:*

$$\lim_{\eta \rightarrow \infty} S_{l,m} \begin{pmatrix} \eta + \tilde{a} & \eta + \tilde{b} \\ \eta + \tilde{d} & \eta + \tilde{c} \end{pmatrix}_{\text{trig}} = q^{\sum_{i < j} (\beta_i - \gamma_i) \gamma_j} \binom{m}{l}_q^{-1} \prod_{i=1}^{n+1} \begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix}_q, \quad (18)$$

where the limit means $\eta_i - \eta_{i+1} \rightarrow \infty$ for all $1 \leq i \leq n$, and the RHS is zero unless $0 \leq \gamma_i \leq \beta_i, \forall i$.

The limit reduces the unrestricted trigonometric $A_n^{(1)}$ face model to the vertex model at a special value of the spectral parameter in the sense that the RHS of (18)| $_{q \rightarrow q^2}$ reproduces [8, eq. (23)] that was obtained as the special value of the quantum R matrix associated with the symmetric tensor representation of $U_q(A_n^{(1)})$.

3 Discussion

Since the weights $W_{l,m} \begin{pmatrix} a & b \\ c & d \end{pmatrix} |u\rangle$ remain unchanged by shifting $a, b, c, d \in \tilde{\mathcal{P}}$ by $\text{const} \cdot (1, \dots, 1)$, we regard them as elements from $\mathcal{P} := \tilde{\mathcal{P}}/\mathbb{C}(1, \dots, 1)$ in the sequel. Given $l, m_1, \dots, m_M \in \mathbb{Z}_{\geq 1}$ and $u, w_1, \dots, w_M \in \mathbb{C}$, the transfer matrix $T_l(u) = T_l \left(u \begin{matrix} m_1, \dots, m_M \\ w_1, \dots, w_M \end{matrix} \right)$ of the unrestricted $A_n^{(1)}$ face model with periodic boundary condition is a linear map on the space of independent row configurations on length M row $\bigoplus \mathbb{C} |a^{(1)}, \dots, a^{(M)}\rangle$ where the sum is taken over $a^{(1)}, \dots, a^{(M)} \in \mathcal{P}$ such that $a^{(i+1)} - a^{(i)} \in B_{m_i}$ ($a^{(M+1)} = a^{(1)}$). Its action is specified as $T_l(u) |b^{(1)}, \dots, b^{(M)}\rangle = \sum_{a^{(1)}, \dots, a^{(M)}} T_l(u)_{b^{(1)}, \dots, b^{(M)}}^{a^{(1)}, \dots, a^{(M)}} |a^{(1)}, \dots, a^{(M)}\rangle$ in terms of the matrix elements

$$T_l(u)_{b^{(1)}, \dots, b^{(M)}}^{a^{(1)}, \dots, a^{(M)}} = \prod_{i=1}^M W_{l, m_i} \begin{pmatrix} a^{(i)} & a^{(i+1)} \\ b^{(i)} & b^{(i+1)} \end{pmatrix} |u - w_i\rangle \quad (a^{(M+1)} = a^{(1)}, b^{(M+1)} = b^{(1)}). \quad (19)$$

Theorem 1 tells that $S_l := T_l(u)_{u=w_1=\dots=w_M}$ has a simple factorized matrix elements. We write its elements as $S_{l, b^{(1)}, \dots, b^{(M)}}^{a^{(1)}, \dots, a^{(M)}}$. The star-triangle relation (7) implies the commutativity $[T_l(u), T_{l'}(u')] = [S_l, S_{l'}] = 0$.

Let us consider whether $X = T_l(u)$ or S_l admits an interpretation as a Markov matrix of a discrete time stochastic process. The related issue was treated in [3] for $n = 1$ and mainly when $\min(l, m_1, \dots, m_M) = 1$. One needs (i) sum-to-1 property $\sum_{a^{(1)}, \dots, a^{(M)}} X_{b^{(1)}, \dots, b^{(M)}}^{a^{(1)}, \dots, a^{(M)}} = 1$ and (ii) nonnegativity $\forall X_{b^{(1)}, \dots, b^{(M)}}^{a^{(1)}, \dots, a^{(M)}} \geq 0$. We concentrate on the trigonometric case in what follows. From Theorem 1 and the fact that $S_{l,m} \begin{pmatrix} a & b \\ c & d \end{pmatrix}_{\text{trig}}$ in (13) is independent of a , (i) indeed holds for S_l . On the other hand (13) also indicates that (ii) is not valid in general without confining the site

variables in a certain range. A typical such prescription is *restriction* [4–6], where one takes $L = \ell + n + 1$ in (2) with some $\ell \in \mathbb{Z}_{\geq 1}$ and lets the site variables range over the finite set of level ℓ dominant integral weights $\{(L + a_{n+1} - a_1 - 1)\Lambda_0 + \sum_{i=1}^n (a_i - a_{i+1} - 1)\Lambda_i \mid L + a_{n+1} > a_1 > \cdots > a_{n+1}, a_i - a_j \in \mathbb{Z}\}$. They are to obey a stronger adjacency condition [6, p. 546, (c-2)] than (3) which is actually the fusion rule of the WZW conformal field theory. (The formal limit $\ell \rightarrow \infty$ still works to restrict the site variables to the positive Weyl chamber and is called “classically restricted”.) Then the star-triangle relation remains valid by virtue of nontrivial cancellation of unwanted terms. However, discarding the contribution to the sum (14) from those b not satisfying the adjacency condition spoils the sum-to-1 property. For example when $(n, l, m) = (2, 1, 2)$, $a = (2, 1, 0)$, $c = (4, 2, 0)$, $d = (3, 1, 0)$ and ℓ is sufficiently large, the unrestricted sum (14) consists of two terms $S_{l,m} \left(\begin{smallmatrix} a & b \\ d & c \end{smallmatrix} \right)_{\text{trig}} = \binom{2}{1}_q^{-1} \frac{(q^{-1}; q)_1}{(q^{-2}; q)_1}$ for $b = (4, 1, 0)$ and $S_{l,m} \left(\begin{smallmatrix} a & b' \\ d & c \end{smallmatrix} \right)_{\text{trig}} = \binom{2}{1}_q^{-1} \frac{(q^3; q)_1}{(q^2; q)_1}$ for $b' = (3, 2, 0)$ summing up to 1, but b' must be discarded in the restricted case since $a \stackrel{m=2}{\Rightarrow} b'$ [6, (c-2)] does not hold. Thus we see that in order to satisfy (i) and (ii) simultaneously one needs to resort to a construction different from the restriction.

Acknowledgements The author thanks Masato Okado for discussion. This work is supported by Grants-in-Aid for Scientific Research No. 15K13429 from JSPS.

References

1. Andrews, G.E., Baxter, R.J., Forrester, P.J.: Eight vertex SOS model and generalized Rogers-Ramanujan-type identities. *J. Stat. Phys.* **35**, 193–266 (1984)
2. Baxter, R.J.: *Exactly Solved Models in Statistical Mechanics*. Academic, London (1982)
3. Borodin, A.: Symmetric elliptic functions, IRF models, and dynamic exclusion processes (2017). arXiv:1701.05239
4. Date, E., Jimbo, M., Kuniba, A., Miwa, T., Okado, M.: Exactly solvable SOS models II: proof of the star-triangle relation and combinatorial identities. *Adv. Stud. Pure Math.* **16**, 17–122 (1988)
5. Jimbo, M., Miwa, T., Okado, M.: Symmetric tensors of the $A_{n-1}^{(1)}$ family. *Algebr. Anal.* **1**, 253–266 (1988)
6. Jimbo, M., Kuniba, A., Miwa, T., Okado, M.: The $A_n^{(1)}$ face models. *Commun. Math. Phys.* **119**, 543–565 (1989)
7. Kuan, J.: An algebraic construction of duality functions for the stochastic $U_q(A_n^{(1)})$ vertex model and its degenerations (2017). arXiv:1701.04468
8. Kuniba, A., Mangazeev, V.V., Maruyama, S., Okado, M.: Stochastic R matrix for $U_q(A_n^{(1)})$. *Nucl. Phys. B* **913**, 248–277 (2016)

Boundary Regularity of Mass-Minimizing Integral Currents and a Question of Almgren



Camillo De Lellis, Guido De Philippis, Jonas Hirsch, and Annalisa Massaccesi

Abstract This short note is the announcement of a forthcoming work in which we prove a first general boundary regularity result for area-minimizing currents in higher codimension, without any geometric assumption on the boundary, except that it is an embedded submanifold of a Riemannian manifold, with a mild amount of smoothness (C^{3,a_0} for a positive a_0 suffices). Our theorem allows to answer a question posed by Almgren at the end of his Big Regularity Paper. In this note we discuss the ideas of the proof and we also announce a theorem which shows that the boundary regularity is in general weaker than the interior regularity. Moreover we remark an interesting elementary byproduct on boundary monotonicity formulae.

1 Introduction

Consider a smooth complete Riemannian manifold Σ of dimension $m + \bar{n}$ and a smooth closed oriented submanifold $\Gamma \subset \Sigma$ of dimension $m - 1$ which is a boundary in integral homology. Since the pioneering work of Federer and Fleming (cf. [20]) we know that Γ bounds an integer rectifiable current T in Σ which minimizes the mass among all integer rectifiable currents bounded by Γ .

C. De Lellis (✉)
Institute for Advanced Study, Princeton, NJ, USA
e-mail: camillo.delellis@math.uzh.ch

G. De Philippis · J. Hirsch
SISSA, Trieste, Italy
e-mail: gdephili@sissa.it; jonas.hirsch@sissa.it

A. Massaccesi
Dipartimento di Informatica, Università di Verona, Verona, Italy
e-mail: annalisa.massaccesi@univr.it

In general, consider an open $U \subset \Sigma$ and a submanifold $\Gamma \subset \Sigma$ which has no boundary in U . If T is an integral current in U with $\partial T \llcorner U = \llbracket \Gamma \rrbracket \llcorner U$ we say that T is mass-minimizing if

$$\mathbf{M}(T + \partial S) \geq \mathbf{M}(T)$$

for every integral current S in U .

Starting with the pioneering work of De Giorgi (see [7]) and thanks to the efforts of several mathematicians in the sixties and the seventies (see [3, 8, 21, 29]), it is known that, if Σ is of class C^{2,a_0} for some $a_0 > 0$, in codimension 1 (i.e., when $\bar{n} = 1$) and away from the boundary Γ , T is a smooth submanifold except for a relatively closed set of Hausdorff dimension at most $m - 7$. Such set, which from now on we will call *interior singular set*, is indeed $(m - 7)$ -rectifiable (cf. [28]) and it has been recently proved that it must have locally finite Hausdorff $(m - 7)$ -dimensional measure (see [27]). In higher codimension, namely when $\bar{n} = 2$, Almgren proved in a monumental work (known as Almgren's Big Regularity Paper [4]) that, if Σ is of class C^5 , then the interior singular set of T has Hausdorff dimension at most $m - 2$. In a series of papers (cf. [9–13]) the first author and Emanuele Spadaro have revisited Almgren's theory introducing several new ideas which simplify his proof considerably. Furthermore, the first author together with Spadaro and Spolaor, in [14–17] applied these sets of ideas to establish a complete proof of Chang's interior regularity results for 2 dimensional mass-minimizing currents [6], showing that in this case interior singular points are isolated.

Both in codimension one and in higher codimension the interior regularity theory described above is, in terms of dimensional bounds for the singular set, optimal (cf. [5] and [19]). In the case of boundary points the situation is instead much less satisfactory. The first boundary regularity result is due to Allard who, in his Ph.D. thesis (cf. [1]), proved that, if $\Sigma = \mathbb{R}^{m+\bar{n}}$ and Γ is lying on the boundary of a uniformly convex set, then for every point $p \in \Gamma$ there is a neighborhood W such that $T \llcorner W$ is a classical oriented submanifold (counted with multiplicity 1) whose boundary (in the usual sense of differential topology) is $\Gamma \cap W$. In his later paper [2] Allard developed a more general boundary regularity theory from which he concluded the above result as a simpler corollary.

When we drop the “convexity assumption” described above, the same conclusion cannot be reached. Let for instance Γ be the union of two concentric circles γ_1 and γ_2 which are contained in a given 2-dimensional plane $\pi_0 \subset \mathbb{R}^{2+\bar{n}}$ and have the same orientation. Then the area-minimizing current T in $\mathbb{R}^{2+\bar{n}}$ which bounds Γ is unique and it is the sum of the two disks bounded by γ_1 and γ_2 in π_0 , respectively. At every point p which belongs to the inner circle the current T is “passing” through the circle while the multiplicity jumps from 2 to 1. However it is natural to consider such points as “regular”, motivating therefore the following definition.

Definition 1 A point $x \in \Gamma$ is a regular point for T if there exist a neighborhood $W \ni x$ and a regular m -dimensional connected submanifold $\Sigma_0 \subset W \cap \Sigma$ (without

boundary in W) such that $\text{spt}(T) \cap W \subset \Sigma_0$. The set of such points will be denoted by $\text{Reg}_b(T)$ and its complement in Γ will be denoted by $\text{Sing}_b(T)$.

By the Constancy Lemma, if $x \in \Gamma$ is a regular point, if Σ_0 is as in Definition 1 and if the neighborhood W is sufficiently small, then the following holds:

1. $\Gamma \cap W$ is necessarily contained in Σ_0 and divides it in two disjoint regular submanifolds Σ_0^+ and Σ_0^- of W with boundaries $\pm\Gamma$;
2. there is a positive $Q \in \mathbb{N}$ such that $T \llcorner W = Q \llbracket \Sigma_0^+ \rrbracket + (Q - 1) \llbracket \Sigma_0^- \rrbracket$.

We define the density of such points p as $Q - \frac{1}{2}$ and we denote it by $\Theta(T, p) = Q - \frac{1}{2}$.

If the density is $\frac{1}{2}$ then the point fulfills the conclusions of Allard’s boundary regularity theorem and Σ_0 is not uniquely determined: the interesting geometrical object is Σ_0^+ and any smooth “extension” of it across Γ can be taken as Σ_0 . On the other hand for $Q \geq 2$ the local behavior of the current is similar to the example of the two circles above: it is easy to see that Σ_0 is uniquely determined and that it has mean curvature zero.

When the codimension of the area-minimizing current is 1, Hardt and Simon proved in [23] that the set of boundary singular points is empty, hence solving completely the boundary regularity problem when $\bar{n} = 1$ (although the paper [23] deals only with the case $\Sigma = \mathbb{R}^{m+\bar{n}}$, its extension to a general Riemannian ambient manifold should not cause real issues). In the case of general codimension and general Γ , Allard’s theory implies the existence of (relatively few) boundary regular points only in special ambient manifolds Σ : for instance when $\Sigma = \mathbb{R}^{m+\bar{n}}$ we can recover the regularity of the “outermost” boundary points $q \in \Gamma$ (i.e., those points q where Γ touches the smallest closed ball which contains it, cf. [24]). According to the existing literature, however, we cannot even exclude that the set of regular points is empty when Σ is a closed Riemannian manifold. In the last remark of the last section of his Big Regularity Paper, cf. [4, Section 5.23, p. 835], Almgren states the following open problem, which is closely related to the discussion carried above.

Question 1 (Almgren) “I do not know if it is possible that the set of density $\frac{1}{2}$ points is empty when $U = \Sigma$ and Γ is connected.”

The interest of Almgren in Question 1 is motivated by an important geometric conclusion: in [4, Section 5.23] he shows that, if there is at least one density $\frac{1}{2}$ point and Γ is connected, then $\text{spt}(T)$ is as well connected and the current T has (therefore) multiplicity 1 almost everywhere. In other words the mass of T coincides with the Hausdorff m -dimensional measure of its interior regular set.

In the forthcoming paper [18] we show the first general boundary regularity result in any codimension, which guarantees the density of boundary regular points without any restriction (except for a mild regularity assumption on Γ and Σ : both are assumed to be of class C^{3,a_0} for some positive a_0 ; note that such regularity assumption for the ambient manifold coincides with the one of the interior regularity theory as developed in the papers [9–13], whereas Almgren’s Big Regularity Paper [4] assumes C^5). As a corollary we answer Almgren’s question in full generality

showing: when $U = \Sigma$ and Γ is connected, then there is always at least one point of density $\frac{1}{2}$ and the support of any minimizer is connected. In the next section we will state the main results of [18], whereas in Sect. 3 we will give an account of their (quite long) proofs. Finally, in Sect. 4 we outline an interesting side remark sparked by one of the key computations in [18]. The latter yields an alternative proof of Allard’s boundary monotonicity formula under slightly different assumptions: in particular it covers, at the same time, the Grüter-Jost monotonicity formula for free boundary stationary varifolds.

2 Main Theorems

Our main result in [18] is the following

Theorem 1 *Consider a C^{3,a_0} complete Riemannian submanifold $\Sigma \subset \mathbb{R}^{m+n}$ of dimension $m + \bar{n}$ and an open set $W \subset \mathbb{R}^{m+n}$. Let $\Gamma \subset \Sigma \cap W$ be a C^{3,a_0} oriented submanifold without boundary in $W \cap \Sigma$ and let T be an integral m -dimensional mass-minimizing current in $W \cap \Sigma$ with boundary $\partial T \llcorner W = \llbracket \Gamma \rrbracket$. Then $\text{Reg}_b(T)$ is dense in Γ .*

As a simple corollary of the theorem above, we conclude that Almgren’s Question 1 has a positive answer.

Corollary 1 *Let $W = \mathbb{R}^{m+n}$ and assume Σ, Γ and T are as in Theorem 1. If Γ is connected, then*

1. *Every point in $\text{Reg}_b(T)$ has density $\frac{1}{2}$;*
2. *The support $\text{spt}(T)$ of the current T is connected;*
3. *The multiplicity of the current is 1 at \mathcal{H}^m -a.e. interior point, and so the mass of the current coincides with $\mathcal{H}^m(\text{spt}(T))$.*

In fact the above corollary is just a case of a more general “structural” result, which is also a consequence of Theorem 1.

Theorem 2 *Let $W = \mathbb{R}^{m+n}$ and assume Σ, Γ and T are as in Theorem 1 and that Γ is in addition compact. Denote by $\Gamma_1, \dots, \Gamma_N$ the connected components of Γ . Then*

$$T = \sum_{j=1}^{\bar{N}} Q_j T_j, \tag{1}$$

where:

- (a) *For every $j = 1, \dots, \bar{N}$, T_j is an integral current with $\partial T_j = \sum_{i=1}^N \sigma_{ij} \llbracket \Gamma_i \rrbracket$ and $\sigma_{ij} \in \{-1, 0, 1\}$.*
- (b) *For every $j = 1, \dots, \bar{N}$, T_j is an area-minimizing current and $T_j = \mathcal{H}^m \llcorner \Lambda_j$, where $\Lambda_1, \dots, \Lambda_{\bar{N}}$ are the connected components of $\text{of } \text{Reg}_i(T)$, the interior regular set.*

(c) Each Γ_i is

- a. either one-sided, which means that all coefficients $\sigma_{ij} = 0$ except for one $j = o(i)$ for which $\sigma_{io(i)} = 1$;
- b. or two-sided, which means that:
 - i. there is one $j = p(i)$ such that $\sigma_{ip(i)} = 1$,
 - ii. there is one $j = n(i)$ such that $\sigma_{in(i)} = -1$,
 - iii. all other $\sigma_{ij} = 0$.

(d) If Γ_i is one-sided, then $Q_{o(i)} = 1$ and all points in $\Gamma_i \cap \text{Reg}_b T$ have multiplicity $\frac{1}{2}$.

(e) If Γ_i is two-sided, then $Q_{n(i)} = Q_{p(i)} - 1$, all points in $\Gamma_i \cap \text{Reg}_b T$ have multiplicity $Q_{p(i)} - \frac{1}{2}$ and $T_{p(i)} + T_{n(i)}$ is area minimizing.

Note that, as a simple consequence of Theorem 2 and the interior regularity theory, we conclude that in every two-sided component Γ_i of the boundary Γ the boundary singular points have dimension at most $m - 2$.

In view of the interior regularity results, one might be tempted to conjecture that Theorem 1 is very suboptimal and that the Hausdorff dimension of $\text{Sing}_b(T)$ is at most $m - 2$. Though currently we do not have an answer to this question, let us stress that at the boundary some new phenomena arise. Indeed, in [18], we can prove the following:

Theorem 3 *There are a smooth closed simple curve $\Gamma \subset \mathbb{R}^4$ and a mass minimizing current T in \mathbb{R}^4 such that $\partial T = \llbracket \Gamma \rrbracket$ and $\text{Sing}_b(T)$ has an accumulation point.*

In particular Chang’s result, namely the discreteness of interior singular points for two dimensional mass minimizing currents, does not hold at the boundary. Actually the example can be modified in order to obtain also a sequence of interior singular points accumulating towards the boundary, see [18].

3 The Main Steps to Theorem 1

In this section we outline the long road which is taken in [18] to prove Theorem 1. We fix therefore Σ , Γ and T as in Theorem 1.

3.1 Reduction to Collapsed Points

Recalling Allard’s monotonicity formula, we introduce at each boundary point $p \in \Gamma$ the density $\Theta(T, p)$, namely the limit, as $r \downarrow 0$, of the normalized mass ratio in the ball $\mathbf{B}_r(p) \subset \mathbb{R}^{m+n}$ (in particular the normalization is chosen so that at regular boundary points the density coincides with the one defined in the previous

section). Using a suitable variant of Almgren’s stratification theorem, we conclude first that, except for a set of Hausdorff dimension at most $m - 2$, at any boundary point p there is a tangent cone which is *flat*, namely which is contained in an m -dimensional plane $\pi \supset T_0\Gamma$. Secondly, using a classical Baire category argument, we show that, for a dense subset of boundary points p , additionally to the existence of a flat tangent cone, there is a sufficiently small neighborhood U where the density $\Theta(T, q)$ is bounded below, at any $q \in \Gamma \cap U$, by $\Theta(T, p)$. In particular the proof of Theorem 1 is reduced to the claim that any such point, which we call *collapsed*, is in fact regular.

3.2 The “Linear” Theory

Assume next that $0 \in \Gamma$ is a collapsed point and let $Q - \frac{1}{2}$ be its density. By Allard’s boundary regularity theory for stationary varifolds, we know a priori that 0 is a regular point if $Q = 1$ and thus we can assume, without loss of generality, that $Q \geq 2$. Fix a flat tangent cone to 0 and assume, up to rotations, that it is the plane $\pi_0 = \mathbb{R}^m \times \{0\}$ and that $T_0\Gamma = \{x_m = 0\} \cap \pi_0$. Denote by π_0^\pm the two half-planes $\pi_0^\pm = \{\pm x_m > 0\} \cap \pi_0$. Assume for the moment that, at suitably chosen small scales, the current T is formed by Q sheets over π_0^+ and $Q - 1$ sheets over π_0^- . By a simple linearization argument such sheets must then be almost harmonic (in a suitable sense).

Having this picture in mind, it is natural to develop a theory of $(Q - \frac{1}{2})$ -valued functions minimizing the Dirichlet energy. In order to explain the latter object consider the projection γ of Γ onto π_0 . On a sufficiently small disk $\mathbf{B}_r(0) \cap \pi_0$, γ divides π_0 into two regions. A Lipschitz $(Q - \frac{1}{2})$ -valued map consists of:

1. a Lipschitz Q -valued map (in the sense of Almgren, cf. [9]) u^+ on one side of γ
2. and a Lipschitz $(Q - 1)$ -valued map u^- on the other side,

satisfying the compatibility condition that the union of their graphs forms a current whose boundary is the submanifold Γ itself. A $(Q - \frac{1}{2})$ -map will then be called Dir-minimizing if it minimizes the sum of the Dirichlet energies of the two “portions” u^+ and u^- under the constraint that Γ and the boundary values on $\partial(\mathbf{B}_r(0) \cap \pi_0)$ are both fixed.

The right counterpart of the “collapsed point situation” described above is the assumption that all the $2Q - 1$ sheets meet at their common boundary Γ ; under such assumption we say that the $(Q - \frac{1}{2})$ Dir-minimizer has collapsed interface. We then develop a suitable regularity theory for minimizers with collapsed interface. First of all their Hölder continuity follows directly from the Ph.D. thesis of the third author, cf. [25]. Secondly, the most important conclusion of our analysis is that a minimizer can have collapsed interface only if it consists of a single harmonic sheet

“passing through” the boundary data, counted therefore with multiplicity Q on one side and with multiplicity $Q - 1$ on the other side.

The latter theorem is ultimately the *deus ex machina* of the entire argument leading to Theorem 1. The underlying reason for its validity is that a monotonicity formula for a suitable variant of Almgren’s frequency function holds. Given the discussion of [26], such monotonicity can only be hoped in the collapsed situation and, remarkably, this suffices to carry on our program.

The validity of the monotonicity formula is clear when the collapsed interface is flat. However, when we have a curved boundary, a subtle yet important point becomes crucial: we cannot hope in general for the exact first variation identities which led Almgren to his monotonicity formula, but we must replace them with suitable inequalities. Moreover the latter can be achieved only if we adapt the frequency function by integrating a suitable weight. We illustrate this idea in a simpler setting in the next section.

3.3 First Lipschitz Approximation

A first use of the linear theory is approximating the current with the graph of a Lipschitz $(Q - \frac{1}{2})$ -valued map around collapsed points. The approximation is then shown to be almost Dir-minimizing. Our approximation algorithm is a suitable adaptation of the one developed in [10] for interior points. In particular, after adding an “artificial sheet”, we can directly use the Jerrard-Soner modified BV estimates of [10] to give a rather accurate Lipschitz approximation: the subtle point is to engineer the approximation so that it has collapsed interface.

3.4 Height Bound and Excess Decay

The previous Lipschitz approximation, together with the linear regularity theory, is used to establish a power-law decay of the excess *à la* De Giorgi in a neighborhood of a collapsed point. The effect of such theorem is that the tangent cone is flat and unique at every point $p \in \Gamma$ in a sufficiently small neighborhood of the collapsed point $0 \in \Gamma$. Correspondingly, the plane $\pi(p)$ which contains such tangent cone is Hölder continuous in the variable $p \in \Gamma$ and the current is contained in a suitable horned neighborhood of the union of such $\pi(p)$.

An essential ingredient of our argument is an accurate height bound in a neighborhood of any collapsed point in terms of the spherical excess. The argument follows an important idea of Hardt and Simon in [23] and takes advantage of an appropriate variant of Moser’s iteration on varifolds, due to Allard, combined with a crucial use of the remainder in the monotonicity formula. The same argument has been also used by Spolaor in a similar context in [30], where he combines it with the decay of the energy for Dir-minimizers, cf. [30, Proposition 5.1 & Lemma 5.2].

3.5 *Second Lipschitz Approximation*

The decay of the excess proved in the previous step is used then to improve the accuracy of the Lipschitz approximation. In particular, by suitably decomposing the domain of the approximating map in a Whitney-type cubical decomposition which refines towards the boundary, we can take advantage of the interior approximation theorem of [10] on each cube and then patch the corresponding graphs together.

3.6 *Left and Right Center Manifolds*

The previous approximation result is combined with a careful smoothing and patching argument to construct a “left” and a “right” center manifold \mathcal{M}^+ and \mathcal{M}^- . The \mathcal{M}^\pm are $C^{3,\kappa}$ submanifolds of Σ with boundary Γ and they provide a good approximation of the “average of the sheets” on both sides of Γ in a neighborhood of the collapsed point $0 \in \Gamma$. They can be glued together to form a $C^{1,1}$ submanifold \mathcal{M} which “passes through Γ ”. Each portion has $C^{3,\kappa}$ estimates *up to the boundary*, but we only know that the tangent spaces at the boundary coincide: we have a priori no information on the higher derivatives. The construction algorithm follows closely that of [12] for the interior, but some estimates must be carefully adapted in order to ensure the needed boundary regularity.

The center manifolds are coupled with two suitable approximating maps N^\pm . The latter take values on the normal bundles of \mathcal{M}^\pm and provide an accurate approximation of the current T . Their construction is a minor variant of the one in [12].

3.7 *Monotonicity of the Frequency Function and Final Blow-Up Argument*

After constructing the center manifolds and the corresponding approximations we use a suitable Taylor expansion of the area functional to show that the monotonicity of the frequency function holds for the approximating maps N^\pm as well.

We then complete the proof of Theorem 1: in particular we show that, if 0 were a singular collapsed point, suitable rescalings of the approximating maps N^\pm would produce, in the limit, a $(Q - \frac{1}{2})$ Dir-minimizer violating the linear regularity theory. On the one hand the estimate on the frequency function plays a primary role in showing that the limiting map is nontrivial. On the other hand the properties of the center manifolds \mathcal{M}^\pm enter in a fundamental way in showing that the average of the sheets of the limiting $(Q - \frac{1}{2})$ map is zero on both sides.

4 Weighted Monotonicity Formulae

In this section we want to illustrate in a simple situation an idea which, in spite of being elementary, plays a fundamental role in our proof of Theorem 1: boundary monotonicity formulae can be derived from the arguments of their interior counterparts provided we introduce a suitable weight.

Let Γ be an $(m - 1)$ -dimensional submanifold of \mathbb{R}^{m+n} . We consider an m -dimensional varifold V in $\mathbb{R}^{m+n} \setminus \Gamma$ and assume it is stationary in $\mathbb{R}^{m+n} \setminus \Gamma$. Allard in [2] derived his famous monotonicity formula at the boundary under the additional assumption that the density of V has a uniform positive lower bound. His proof consists of two steps: he first derives a suitable representation for the first variation δV of V along general vector fields of \mathbb{R}^{m+n} , i.e., vector fields which might be nonzero on Γ . He then follows the derivation of the interior monotonicity formula, i.e., he tests the first variation along suitable radial vector fields. His proof needs the lower density assumption in the first part and although the latter can be removed (cf. Allard, W.K., Personal communication), the resulting argument is rather laborious.

We introduce here varifolds which are stationary along “tangent fields”:

Definition 2 Consider an m -dimensional varifold V in an open set $U \subset \mathbb{R}^{m+n}$ and let Γ be a k -dimensional C^1 submanifold of U . We say that V is stationary with respect to vector fields tangent to Γ if

$$\delta V(\chi) = 0 \quad \text{for all } \chi \in C_c^1(U, \mathbb{R}^{m+n}) \text{ which are tangent to } \Gamma. \tag{2}$$

Clearly, when $k = m - 1$, the condition above is stronger than that used by Allard in [2], where χ is assumed to *vanish* on Γ . On the other hand our condition is the natural one satisfied by classical minimal surfaces with boundary Γ , since the one-parameter family of isotopies generated by χ maps Γ onto itself. When $k > m - 1$, the condition is the one satisfied by classical “free-boundary” minimal surfaces, namely minimal surfaces with boundary contained in Γ and meeting it orthogonally. In the context of varifolds, the latter have been considered by Grüter and Jost in [22], where the two authors derived also an analog of Allard’s monotonicity formula. In this section we show how one can take advantage of a suitable distortion of the Euclidean balls to give a (rather elementary) unified approach to monotonicity formulae in both contexts.

Definition 3 Assume that $0 \in \Gamma$. We say that the function $d : \mathbb{R}^{m+n} \rightarrow \mathbb{R}$ is a distortion of the distance function adapted to Γ if the following two conditions hold:

- (a) d is of class C^2 on $\mathbb{R}^{m+n} \setminus \{0\}$ and $D^j d(x) = D^j |x| + O(|x|^{1-j+\alpha})$ for some fixed $\alpha \in (0, 1]$ and for $j = 0, 1, 2$;
- (b) ∇d is tangent to Γ .

The following lemma is a simple consequence of the Tubular Neighborhood Theorem and it is left to the reader.

Lemma 1 *If Γ is of class C^3 then there is a distortion of the distance function adapted to Γ where the exponent α of Definition 3(a) can be taken to be 1.*

The main point of our discussion is then the argument given below for the following

Theorem 4 *Consider Γ and V as in Definition 2, assume that $0 \in \Gamma$ and that d is a distorted distance function adapted to Γ . Let $\varphi \in C_c^1([0, 1])$ be a nonincreasing function which is constant in a neighborhood of the origin. If α is the exponent of Definition 3(a), then there are positive constants C and ρ such that the following inequality holds for every positive $s < \rho$*

$$\begin{aligned} & \frac{d}{ds} \left[e^{Cs^\alpha} s^{-m} \int \varphi \left(\frac{d(x)}{s} \right) d\|V\|(x) \right] \\ & \geq -e^{Cs^\alpha} s^{-m-1} \int \varphi' \left(\frac{d(x)}{s} \right) \frac{d(x)}{s} \left| P_{\pi^\perp} \left(\frac{\nabla d(x)}{|\nabla d(x)|} \right) \right|^2 dV(x, \pi) \end{aligned} \tag{3}$$

(where P_τ denotes the orthogonal projection on the subspace τ).

Note that if we let φ converge to the indicator function of the interval $[0, 1)$ we easily conclude that

$$s \mapsto \Phi(s) := e^{Cs^\alpha} \frac{\|V\|(\{d < s\})}{s^m}$$

is monotone nondecreasing: indeed, for $\rho > s > r > 0$, the difference $\Phi(s) - \Phi(r)$ controls the integral of a suitable nonnegative expression involving d and the projection of $\nabla d/|\nabla d|$ over π^\perp . When $d(x) = |x|$, namely when Γ is flat, the exponential weight disappears (i.e., the constant C might be taken to be 0), the inequality becomes an equality and (in the limit of $\varphi \uparrow \mathbf{1}_{[0,1)}$) we recover Allard’s identity

$$\frac{\|V\|(\mathbf{B}_s(0))}{\omega_m s^m} - \frac{\|V\|(\mathbf{B}_r(0))}{\omega_m r^m} = \int_{\mathbf{B}_s(0) \setminus \mathbf{B}_r(0)} \frac{|P_{\pi^\perp}(x)|^2}{|x|^{m+2}} d\|V\|(x).$$

In particular, since d is asymptotic to $|x|$, all the conclusions which are usually derived from Allard’s theorem (existence of the density and its upper semicontinuity, conicity of the tangent varifolds, Federer’s reduction argument and Almgren’s stratification) can be derived from Theorem 4 as well. Moreover, the argument given below can be easily extended to cover the more general situation of varifolds with mean curvature satisfying a suitable integrability condition.

Proof (of Theorem 4) Consider the vector field

$$X_s(x) = \varphi \left(\frac{d(x)}{s} \right) d(x) \frac{\nabla d(x)}{|\nabla d(x)|^2}.$$

X_s is obviously C^1 on $\mathbb{R}^{m+n} \setminus \{0\}$ and moreover we have

$$DX_s = \varphi \left(\frac{d}{s} \right) \left[\frac{\nabla d \otimes \nabla d}{|\nabla d|^2} + \frac{dD^2d}{|\nabla d|^2} - 2d \frac{\nabla d}{|\nabla d|^4} \otimes (D^2d \cdot \nabla d) \right] + \varphi' \left(\frac{d}{s} \right) \frac{d}{s} \frac{\nabla d \otimes \nabla d}{|\nabla d|^2}.$$

From the above formula, using that φ is constant in a neighborhood of the origin and Definition 3(a), we easily infer that (for every fixed s)

$$DX_s(x) = \varphi \left(\frac{d(x)}{s} \right) \text{Id} + O(|x|^\alpha).$$

In particular X_s is C^1 , compactly supported in U (provided s is sufficiently small), and tangent to Γ . Thus

$$0 = \delta V(X_s) = \int \text{div}_\pi X_s(p) dV(p, \pi).$$

Fix next an orthonormal basis e_1, \dots, e_m of π and use Definition 3(a) to compute

$$\begin{aligned} \text{div}_\pi X_s &= \sum_{i=1}^m e_i^T \cdot DX \cdot e_i = (m + O(s^\alpha))\varphi \left(\frac{d}{s} \right) + \varphi' \left(\frac{d}{s} \right) \frac{d}{s} \sum_i \frac{|\nabla d \cdot e_i|^2}{|\nabla d|^2} \\ &= (m + O(s^\alpha))\varphi \left(\frac{d}{s} \right) + \varphi' \left(\frac{d}{s} \right) \frac{d}{s} \left(1 - \left| P_{\pi^\perp} \left(\frac{\nabla d}{|\nabla d|} \right) \right|^2 \right). \end{aligned}$$

Plugging the latter identity in the first variation condition we achieve the following inequality for a sufficiently large constant C :

$$\begin{aligned} &\int \left(-m\varphi \left(\frac{d(x)}{s} \right) - \varphi' \left(\frac{d(x)}{s} \right) \frac{d(x)}{s} \right) d\|V\|(x) + C\alpha s^\alpha \int \varphi \left(\frac{d(x)}{s} \right) d\|V\|(x) \\ &\geq - \int \varphi' \left(\frac{d(x)}{s} \right) \frac{d(x)}{s} \left| P_{\pi^\perp} \left(\frac{\nabla d(x)}{|\nabla d(x)|} \right) \right|^2 dV(x, \pi). \end{aligned}$$

Multiplying both sides of the inequality by $e^{Cs^\alpha} s^{-m-1}$ we then conclude (3).

References

1. Allard, W.K.: On boundary regularity for Plateau’s problem. Bull. Am. Math. Soc. **75**, 522–523 (1969). <https://doi.org/10.1090/S0002-9904-1969-12229-9>
2. Allard, W.K.: On the first variation of a varifold: boundary behavior. Ann. Math. (2) **101**, 418–446 (1975)

3. Almgren, F.J.J.: Some interior regularity theorems for minimal surfaces and an extension of Bernstein's theorem. *Ann. Math. (2)* **84**, 277–292 (1966)
4. Almgren, J.F.J.: Almgren's big regularity paper. In: *World Scientific Monograph Series in Mathematics*, vol. 1. World Scientific Publishing, River Edge (2000)
5. Bombieri, E., De Giorgi, E., Giusti, E.: Minimal cones and the Bernstein problem. *Invent. Math.* **7**, 243–268 (1969)
6. Chang, S.X.: Two-dimensional area minimizing integral currents are classical minimal surfaces. *J. Am. Math. Soc.* **1**(4), 699–778 (1988). <http://dx.doi.org/10.2307/1990991>
7. De Giorgi, E.: *Frontiere orientate di misura minima*. In: *Seminario di Matematica della Scuola Normale Superiore di Pisa, 1960–61*. Editrice Tecnico Scientifica, Pisa (1961)
8. De Giorgi, E.: Una estensione del teorema di Bernstein. *Ann. Scuola Norm. Sup. Pisa* **19**(3), 79–85 (1965)
9. De Lellis, C., Spadaro, E.: Q -valued functions revisited. *Mem. Am. Math. Soc.* **211**(991), vi+79 (2011). <http://dx.doi.org/10.1090/S0065-9266-10-00607-1>
10. De Lellis, C., Spadaro, E.: Regularity of area minimizing currents I: gradient L^p estimates. *Geom. Funct. Anal.* **24**(6), 1831–1884 (2014). <https://doi.org/10.1007/s00039-014-0306-3>
11. De Lellis, C., Spadaro, E.: Multiple valued functions and integral currents. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (5)* **14**(4), 1239–1269 (2015)
12. De Lellis, C., Spadaro, E.: Regularity of area minimizing currents II: center manifold. *Ann. Math. (2)* **183**(2), 499–575 (2016). <https://doi.org/10.4007/annals.2016.183.2.2>
13. De Lellis, C., Spadaro, E.: Regularity of area minimizing currents III: blow-up. *Ann. Math. (2)* **183**(2), 577–617 (2016). <https://doi.org/10.4007/annals.2016.183.2.3>
14. De Lellis, C., Spadaro, E., Spolaor, L.: Regularity theory for 2 dimensional almost minimal currents: blow-up. *J. Differ. Geom.* (2015). Available at arXiv:1508.05510
15. De Lellis, C., Spadaro, E., Spolaor, L.: Regularity theory for 2-dimensional almost minimal currents II: Branched center manifold. *Ann. PDE* **3**(2), Art. 18, 85 (2017). <https://doi.org/10.1007/s40818-017-0035-7>
16. De Lellis, C., Spadaro, E., Spolaor, L.: Uniqueness of tangent cones for two-dimensional almost-minimizing currents. *Comm. Pure Appl. Math.* **70**(7), 1402–1421 (2017). <https://doi.org/10.1002/cpa.21690>
17. De Lellis, C., Spadaro, E., Spolaor, L.: Regularity theory for 2-dimensional almost minimal currents I: Lipschitz approximation. *Trans. Am. Math. Soc.* **370**(3), 1783–1801 (2018). <https://doi.org/10.1090/tran/6995>
18. De Lellis, C., De Philippis, G., Hirsch, J., Massaccesi, A.: On the boundary behavior of mass-minimizing integral currents (2018). Available at arXiv:1809.09457
19. Federer, H.: *Geometric measure theory*. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer, New York (1969)
20. Federer, H., Fleming, W.H.: Normal and integral currents. *Ann. Math. (2)* **72**, 458–520 (1960)
21. Fleming, W.H.: On the oriented Plateau problem. *Rend. Circ. Mat. Palermo (2)* **11**, 69–90 (1962)
22. Grüter, M., Jost, J.: Allard type regularity results for varifolds with free boundaries. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* **13**(1), 129–169 (1986)
23. Hardt, R., Simon, L.: Boundary regularity and embedded solutions for the oriented Plateau problem. *Ann. Math. (2)* **110**(3), 439–486 (1979). <http://dx.doi.org/10.2307/1971233>
24. Hardt, R.M.: On boundary regularity for integral currents or flat chains modulo two minimizing the integral of an elliptic integrand. *Comm. Partial Differential Equations* **2**(12), 1163–1232 (1977). <https://doi.org/10.1080/03605307708820058>
25. Hirsch, J.: Boundary regularity of Dirichlet minimizing Q -valued functions. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (5)* **16**(4), 1353–1407 (2016)
26. Hirsch, J.: Examples of holomorphic functions vanishing to infinite order at the boundary. *Trans. Am. Math. Soc.* **370**, 4249–4271 (2018)
27. Naber, A., Valtorta, D.: The singular structure and regularity of stationary and minimizing varifolds. *ArXiv e-prints* (2015)

28. Simon, L.: Rectifiability of the singular sets of multiplicity 1 minimal surfaces and energy minimizing maps. In: *Surveys in differential geometry, vol. II* (Cambridge, MA, 1993), pp. 246–305. Int. Press, Cambridge (1995)
29. Simons, J.: Minimal varieties in riemannian manifolds. *Ann. Math. (2)* **88**, 62–105 (1968)
30. Spolaor, L.: Almgren’s type regularity for Semicalibrated Currents. ArXiv e-prints (2015)

Optimal Transport with Discrete Mean Field Interaction



Jiakun Liu and Grégoire Loeper

Abstract In this note, we summarise some regularity results recently obtained for an optimal transport problem where the matter transported is either accelerated by an external force field, or self-interacting, at a given intermediate time.

1 Background

This note is a summary of an ongoing work [5]. The motivation comes from a previous work by the second author [6], where he studies the motion of a self-gravitating matter, classically described by the Euler-Poisson system. Letting ρ be the density of the matter, the gravitational field generated by a continuum of matter with density ρ is the gradient of a potential p linked to ρ by a Poisson coupling. The system is thus the following

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho v) = 0, \\ \partial_t (\rho v) + \nabla \cdot (\rho v \otimes v) = -\rho \nabla p, \\ \Delta p = \rho. \end{cases} \quad (1)$$

A well known problem in cosmology, named the reconstruction problem, is to find a solution to (1) satisfying

$$\rho|_{t=0} = \rho_0, \quad \rho|_{t=T} = \rho_T.$$

J. Liu (✉)

School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, NSW, Australia

e-mail: jiakunl@uow.edu.au

G. Loeper

School of Mathematical Sciences, Monash University, Melbourne, VIC, Australia

e-mail: gregoire.loeper@monash.edu

In [6], the reconstruction problem was formulated into a minimisation problem, minimising the action of the Lagrangian which is a convex functional. Through this variational formulation, the reconstruction problem becomes very similar to the time continuous formulation of the optimal transportation problem of Benamou and Brenier [1], and the existence, uniqueness of the minimiser was obtained by use of the Monge-Kantorovich duality. In the context of optimal transport as in [6], there holds $v = \nabla\phi$ for some potential ϕ , and the author obtained partial regularity results for ϕ and ρ , as well as the consistency of the minimiser with the solution of the Euler-Poisson system.

The optimal transport problem of [6] was formulated as finding minimisers of the action of the Lagrangian

$$I(\rho, v, p) = \frac{1}{2} \int_0^T \int_{\mathbb{T}^d} \rho(t, x) |v(t, x)|^2 + |\nabla p(t, x)|^2 dx dt, \quad (2)$$

over all ρ, p, v satisfying

$$\begin{aligned} \partial_t \rho + \nabla \cdot (\rho v) &= 0, \\ \rho(0) &= \rho_0, \quad \rho(T) = \rho_T, \\ \Delta p &= \rho, \end{aligned}$$

where \mathbb{T}^d denotes the d -dimensional torus, as the study in [6] was performed in the space-periodic case.

In the work [5] we address the more general problem of finding minimisers for the action

$$I(\rho, v, p) = \frac{1}{2} \int_0^T \int_{\mathbb{T}^d} \rho(t, x) |v(t, x)|^2 + \mathcal{F}(\rho(t, x)) dx dt, \quad (3)$$

for more general \mathcal{F} . The problem (2) falls in this class. In [3], Lee and McCann address the case where

$$\mathcal{F}(\rho) = - \int \rho(t, x) V(t, x) dx.$$

(Note that in the context of classical mechanics \mathcal{F} would be the potential energy.) This Lagrangian corresponds to the case of a continuum of matter evolving in an external force field given by $\nabla V(t, x)$. We call this the non-interacting case for obvious reasons. This can be recast as a classic optimal transport problem, where the cost functional is given by

$$c(x, y) = \inf_{\substack{\gamma(0)=x, \gamma(T)=y \\ \gamma \in C^1([0, T], \mathbb{R}^d)}} \int_0^T \frac{1}{2} |\dot{\gamma}(t)|^2 - V(t, \gamma(t)) dt. \quad (4)$$

For a small V satisfying some structure condition, they obtain that c satisfies the conditions found in [8] to ensure the regularity of the optimal map.

2 Time Discretisation

In [5] we restrict ourselves to the case where the force field only acts at a single discrete time between 0 and T :

$$V(t, x) = \delta_{t=T/2} V(x).$$

We will call this case the “discrete” case. The minimisation problem therefore becomes

$$I(\rho, v) = \frac{1}{2} \int_0^T \int_{\mathbb{R}^d} \rho(t, x) |v(t, x)|^2 dx dt + \int_{\mathbb{R}^d} \rho(T/2, x) Q(x) dx, \quad (5)$$

for some potential Q . This will allow to remove the smallness condition on V . Moreover, we will be able to extend our result to the mean-field case, where the force field is given by

$$\nabla V(x) = \int \rho(t, y) \nabla \kappa(x - y) dy. \quad (6)$$

This corresponds to the case where a particle located at x attracts or repels another particle located at y with a force equal to $\nabla \kappa(x - y)$. We will give a sufficient condition on κ to ensure a smooth transport map and intermediate density. Especially, we consider the gravitational case, which corresponds to the Coulomb kernel

$$\kappa(x - y) = \frac{c_d}{|x - y|^{d-2}},$$

that corresponds to the potential energy

$$\begin{aligned} \mathcal{E}(t) &= -\mathcal{F}(\rho(t)) = -\frac{1}{2} \int \rho(t, x) \kappa(x - y) \rho(t, y) dx dy \\ &= -\frac{1}{2} \int \|\nabla p\|^2, \end{aligned}$$

where $\Delta p = \rho$.

One sees straight away that between time 0 and $T/2$ we are solving the usual optimal transport problem in its “Benamou-Brenier” formulation [1], as well as between $T/2$ and T . More generally, as done in [6], one can consider multiple-steps

time discretisation, where the potential energy term contributes only at time

$$t_i = \frac{iT}{N}, \quad i = 1, \dots, N-1.$$

Between two time steps, the problem will be an optimal transport problem as in [1, 2] and [9]. Then at each time step, the gravitational effect will be taken into account, and the velocity will be discontinuous. From a Lagrangian point of view, the velocity of each particle will therefore be a piecewise constant function with respect to time. Then letting the time step go to 0, one will eventually recover the time continuous problem.

3 Main Results

Let us consider a two-step time discretisation in the interval $[0, T]$: At $t = T/2$, the velocity is changed by an amount equal to ∇Q , the gradient of a potential Q . The initial density ρ_0 is supported on a bounded domain $\Omega_0 \subset \mathbb{R}^d$, and the final density ρ_T is supported on a bounded domain $\Omega_T \subset \mathbb{R}^d$, satisfying the balance condition

$$\int_{\Omega_0} \rho_0(x) dx = \int_{\Omega_T} \rho_T(y) dy. \quad (7)$$

As is always the case in solving problems of the form (3), the velocity v is the gradient of a potential, and we let ϕ be the velocity potential at time 0, i.e. $v(0, x) = \nabla\phi(x)$. At time $t = T/2$, v will be changed into $v + \nabla Q$ and one can see that for an initial point $x \in \Omega_0$, the final point $y = \mathbf{m}(x) \in \Omega_T$ is given by

$$\mathbf{m}(x) = x + T\nabla\phi + \frac{T}{2}\nabla Q \left(x + \frac{T}{2}\nabla\phi \right).$$

By computing the determinant of the Jacobian $D\mathbf{m}$ and noting that \mathbf{m} pushes forward ρ_0 to ρ_T , one can derive the equation for ϕ . To be specific, define a modified potential

$$\tilde{\phi}(x) := \frac{T}{2}\phi(x) + \frac{1}{2}|x|^2, \quad \text{for } x \in \Omega_0. \quad (8)$$

It is readily seen [1, 2, 9] that the modified potential $\tilde{\phi}$ is a convex function. Since $\mathbf{m}_\# \rho_0 = \rho_T$, we obtain that $\tilde{\phi}$ satisfies a Monge-Ampère type equation

$$\det \left[D^2\tilde{\phi} - \left(D^2\tilde{Q}(\nabla\tilde{\phi}) \right)^{-1} \right] = \left(\frac{1}{\det D^2\tilde{Q}(\nabla\tilde{\phi})} \right) \frac{\rho_0}{\rho_T \circ \mathbf{m}}, \quad (9)$$

where \tilde{Q} is a modified potential given by

$$\tilde{Q}(z) := \frac{T}{2}Q(z) + |z|^2, \tag{10}$$

with an associated natural boundary condition

$$\mathbf{m}(\Omega_0) = \Omega_T. \tag{11}$$

For regularity of the solution $\tilde{\phi}$ to the boundary value problem (9) and (11) (equivalently that of ϕ), it is necessary to impose certain conditions on the potential energy function \tilde{Q} (equivalently on Q) and the domains Ω_0, Ω_T . In [5] we assume that \tilde{Q} satisfies the following conditions:

- (H0) The function \tilde{Q} is smooth enough, say at least C^4 ,
- (H1) The function \tilde{Q} is uniformly convex, namely $D^2\tilde{Q} \geq \varepsilon_0 I$ for some $\varepsilon_0 > 0$,
- (H2) The function \tilde{Q} satisfies that for all $\xi, \eta \in \mathbb{R}^d$ with $\xi \perp \eta$,

$$\sum_{i,j,k,l,p,q,r,s} \left(D_{ijrs}^4 \tilde{Q} - 2\tilde{Q}^{pq} D_{ijp}^3 \tilde{Q} D_{qrs}^3 \tilde{Q} \right) \tilde{Q}^{rk} \tilde{Q}^{sl} \xi_k \xi_l \eta_i \eta_j \leq -\delta_0 |\xi|^2 |\eta|^2, \tag{12}$$

where $\{\tilde{Q}^{ij}\}$ is the inverse of $\{\tilde{Q}_{ij}\}$, and δ_0 is a positive constant. When $\delta_0 = 0$, we call it (H2w), a weak version of (H2).

Note that conditions (H0) and (H1) imply that the inverse matrix $(D^2\tilde{Q})^{-1}$ exists, and ensure that Eq. (9) well defined. Condition (H2) is an analogue of the Ma-Trudinger-Wang condition [8] in optimal transportation, which is necessary for regularity results. We also use the notion of Q -convexity of domains as in [8].

Our first main result is the following

Theorem 1 *Let ϕ be the velocity potential in the reconstruction problem. Assume the gravitational function \tilde{Q} satisfies conditions (H0), (H1) and (H2), Ω_T is Q -convex with respect to Ω_0 . Assume that $\rho_T \geq c_0$ for some positive constant c_0 , $\rho_0 \in L^p(\Omega_0)$ for some $p > \frac{d+1}{2}$, and the balance condition (7) is satisfied. Then, the velocity potential ϕ is $C^{1,\alpha}(\overline{\Omega_0})$ for some $\alpha \in (0, 1)$.*

If furthermore, Ω_0, Ω_T are C^4 smooth and uniformly Q -convex with respect to each other, $\rho_0 \in C^2(\overline{\Omega_0})$, $\rho_T \in C^2(\overline{\Omega_T})$, then $\phi \in C^3(\overline{\Omega_0})$, and higher regularity follows from the theory of linear elliptic equations. In particular, if $\tilde{Q}, \Omega_0, \Omega_T, \rho_0, \rho_T$ are C^∞ , then the velocity potential $\phi \in C^\infty(\overline{\Omega_0})$.

The proof of Theorem 1 is done by linking the time discretisation problem to a transport problem, where the key observation is that the cost function $c(x, y)$ is given by $\tilde{Q}^*(x + y)$, where \tilde{Q}^* is the Legendre transform of the gravitational function \tilde{Q} . Under this formulation, the regularity then follows from the established theory of optimal transportation, see for example [4, 7, 8, 10] and references therein.

Our second main result is the following:

Theorem 2 *Assume that Q is given by*

$$Q(x) = \frac{1}{2} \int_{\Omega_{T/2}} \rho(T/2, y) \kappa(x - y) dy, \tag{13}$$

where $\Omega_{T/2} = (\text{Id} + \frac{T}{2} \nabla \phi)(\Omega_0)$ is the intermediate domain at $t = \frac{T}{2}$, and that κ satisfies conditions (H0), (H1) and

(H2C) for any $\xi, \eta \in \mathbb{R}^d, x, y \in \Omega_{T/2}$,

$$\sum_{i,j,k,l,p,q,r,s} \left(D_{ijrs}^4 \kappa(x - y) \right) \tilde{\kappa}^{rk} \tilde{\kappa}^{sl} \xi_k \xi_l \eta_i \eta_j \leq 0,$$

where $\{\tilde{\kappa}^{ij}\}$ is the inverse of $\{\kappa_{ij} + \frac{2}{T} I\}$,

We also assume some geometric conditions on the domains. Then the results of Theorem 1 remain true.

The proof of Theorem 2 relies on the observation that (H2c) implies (H2), and is preserved under convex combinations, and therefore by convolution with the density $\rho(T/2)$, and on some a priori C^1 estimates on the potential. Full details and further remarks are contained in our work [5].

References

1. Benamou, J.-D., Brenier, Y.: A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.* **84**, 375–393 (2000)
2. Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. *Commun. Pure Appl. Math.* **44**, 375–417 (1991)
3. Lee, P., McCann, R.J.: The Ma-Trudinger-Wang curvature for natural mechanical actions. *Calc. Var. PDEs* **41**, 285–299 (2011)
4. Liu, J.: Hölder regularity of optimal mappings in optimal transportation. *Calc. Var. PDEs* **34**, 435–451 (2009)
5. Liu, J., Loeper, G.: Optimal transport with discrete long range mean field interactions (submitted). Available at arXiv:1809.07432
6. Loeper, G.: The reconstruction problem for the Euler-Poisson system in cosmology. *Arch. Ration. Mech. Anal.* **179**, 153–216 (2006)
7. Loeper, G.: On the regularity of solutions of optimal transportation problems. *Acta Math.* **202**, 241–283 (2009)
8. Ma, X.-N., Trudinger, N.S., Wang, X.-J.: Regularity of potential functions of the optimal transportation problems. *Arch. Ration. Mech. Anal.* **177**, 151–183 (2005)
9. McCann, R.J.: A convexity principle for interacting gases. *Adv. Math.* **128**, 153–179 (1997)
10. Trudinger, N.S., Wang, X.-J.: On the second boundary value problem for Monge-Ampère type equations and optimal transportation. *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* **8**, 143–174 (2009)

A Sixth Order Curvature Flow of Plane Curves with Boundary Conditions



James McCoy, Glen Wheeler, and Yuhan Wu

Abstract We show that small energy curves under a particular sixth order curvature flow with generalised Neumann boundary conditions between parallel lines converge exponentially in the C^∞ topology in infinite time to straight line segments.

1 Introduction

Higher order geometric evolution problems have received increasing attention in the last few years. Particular geometric fourth order equations occur in physical problems and enjoy some interesting applications in mathematics. We mention in particular for curves the curve diffusion flow and L^2 -gradient flow of the elastic energy, and for surfaces the surface diffusion and Willmore flows. Flows of higher even order than four have been less thoroughly investigated, but motivation for them and their elliptic counterparts comes for example from computer design, where higher order equations are desirable as they allow more flexibility in terms of prescribing boundary conditions [8]. Such equations have also found applications in medical imaging [10].

In this article we are interested in curves γ meeting two parallel lines with Neumann (together with other) boundary conditions evolving under the L^2 gradient flow for the energy

$$\int_{\gamma} k_s^2 ds.$$

J. McCoy (✉)
University of Newcastle, Callaghan, NSW, Australia
e-mail: James.McCoy@newcastle.edu.au

G. Wheeler · Y. Wu
University of Wollongong, Wollongong, NSW, Australia
e-mail: glenw@uow.edu.au; yw120@uowmail.edu.au

Here k_s denotes the first derivative of curvature with respect to the arc length parameter s . Particularly relevant to us is the corresponding consideration of the curve diffusion and elastic flow in this setting in [12]. Other relevant works on fourth order flow of curves with boundary conditions are [1, 2, 7]. Of course if one instead considers closed curves without boundary evolving by higher order equations, these have been more thoroughly studied; we mention in particular [3–5, 9, 11].

The remainder of this article is organised as follows. In Sect. 2 we describe the set-up of our problem, the normal variation of the energy and the boundary conditions. We define our corresponding gradient flow, discuss local existence and give the relevant evolution equations of various geometric quantities. We also state our main theorem in this part, Theorem 2.2. In Sect. 3 we state the relevant tools from analysis to be used including an interpolation inequality valid in our setting. Under the small energy condition (7) below, we show that the winding number of curves under our flow is constant and remains equal to zero. We show further that under this condition the length of the curve does not increase and the curvature and all curvature derivatives in L^2 are bounded under the flow. That these bounds are independent of time implies solutions exist for all time. In Sect. 4 we show under a smaller energy assumption that in fact the L^2 norm of the second derivative of curvature decays exponentially under the flow. As a corollary we obtain uniform pointwise exponential decay of curvature and all curvature derivatives to zero. A stability argument shows that the solution converges to a unique horizontal line segment. The exponential convergence of the flow speed allows us to describe the bounded region in which the solution remains under the flow.

2 The Set-Up

Let $\gamma_0 : [-1, 1] \rightarrow \mathbb{R}^2$ be a (suitably) smooth embedded (or immersed) regular curve. Denote by ds the arc length element and k the (scalar) curvature. We consider the energy functional

$$E[\gamma] = \frac{1}{2} \int_{\gamma} k_s^2 ds$$

where k_s is the derivative of curvature with respect to arc length. We are interested in the L^2 gradient flow for curves of small initial energy with Neumann boundary conditions.

Under the normal variation $\tilde{\gamma} = \gamma + \varepsilon F \nu$ a straightforward calculation yields

$$\begin{aligned} \frac{d}{d\varepsilon} E[\tilde{\gamma}] \Big|_{\varepsilon=0} &= -2 \int_{\gamma} \left(k_{s^4} + k^2 k_{ss} - \frac{1}{2} k k_s^2 \right) F ds \\ &\quad + 2 \left[k_s F_{ss} + k_{ss} F_s + \left(k_{sss} + k^2 k_s \right) F \right]_{\partial\gamma}, \end{aligned} \tag{1}$$

where $k_{s^4} = k_{ssss}$.

‘Natural boundary conditions’ for the corresponding L^2 -gradient flow would ensure that the above boundary term is equal to zero. However, this term is rather complicated. In view of the first term in (1), we wish to take

$$F = k_{s^4} + k^2 k_{ss} - \frac{1}{2} k k_s^2 \quad (2)$$

and the corresponding gradient flow

$$\frac{\partial \gamma}{\partial t} = F \nu. \quad (3)$$

Differentiating the Neumann boundary condition (see also [12, Lemma 2.5] for example) implies

$$0 = -F_s(\pm 1, t) = -k_{s^5} - k k_s k_{ss} - k^2 k_{sss} + \frac{1}{2} k_s^3. \quad (4)$$

As in previous work, we will assume the ‘no curvature flux condition’ at the boundary,

$$k_s(\pm 1, t) = 0. \quad (5)$$

The boundary terms in (1) then disappear if we choose, for example,

$$k_{sss}(\pm 1, t) = 0. \quad (6)$$

This is in a way a natural choice because Eq. (4) then implies $k_{s^5}(\pm 1, t) = 0$. In fact by an induction argument we have

Lemma 2.1 *With Neumann boundary conditions and also (5) and (6) satisfied, a solution to the flow (3) satisfies $k_{s^{2\ell+1}} = 0$ on the boundary for $\ell \in \mathbb{N}$.*

Let us now state precisely the flow problem.

Let $\eta_{\pm}(\mathbb{R})$ denote two parallel vertical lines in \mathbb{R}^2 , with distance between them $|e|$. We consider a family of plane curves $\gamma : [-1, 1] \times [0, T) \rightarrow \mathbb{R}^2$ satisfying the evolution Eq. (3) with normal speed given by (2), boundary conditions

$$\gamma(\pm 1, t) \in \eta_{\pm}(\mathbb{R})$$

$$\langle \nu, \nu_{\eta_{\pm}} \rangle(\pm 1, t) = 0$$

$$k_s(\pm 1, t) = k_{sss}(\pm 1, t) = 0$$

and initial condition

$$\gamma(\cdot, 0) = \gamma_0(\cdot)$$

for initial smooth regular curve γ_0 .

Theorem 2.2 *There exists a universal constant $C > 0$ such that the following holds. For the flow problem described above, if the initial curve γ_0 satisfies $\omega = 0$ and*

$$\delta = \left(\frac{\sqrt{5129} - 67}{80} \right) \pi^3 - \|k_s\|_2^2 L_0^3 > 0, \tag{7}$$

where L_0 is the length of γ_0 , then the solution exists for all time $T = \infty$ and converges exponentially to a horizontal line segment γ_∞ with $\text{dist}(\gamma_\infty, \gamma_0) < C/\delta$.

In the above statement and throughout the article we use ω to denote the winding number, defined here as

$$\omega := \frac{1}{2\pi} \int_{\gamma} k \, ds.$$

Remarks

- The condition (7) is not optimal. By a standard argument it can be weakened for example to the requirement of Lemma 3.4, namely

$$\frac{\pi^3}{7} - \|k_s\|_2^2 L_0^3 > 0.$$

Details of this argument will appear in a future article. It is an open question whether the requirement can be further weakened.

- The exponential decay facilitates an explicit estimate on the distance of γ_∞ to γ_0 .

Local existence of a smooth regular curve solution $\gamma : [-1, 1] \times [0, T) \rightarrow \mathbb{R}^2$ to the flow problem $\gamma : [-1, 1] \times [0, T) \rightarrow \mathbb{R}^2$ is standard. If γ_0 also satisfies compatibility conditions, then the solution is smooth on $[0, T)$. In this article we focus on the case of smooth initial γ_0 . However, γ_0 may be much less smooth; Eq. (3) is smoothing. We do not pursue this here.

Similarly as in [12] and elsewhere we may derive the following:

Lemma 2.3 *Under the flow (3) we have the following evolution equations*

- (i) $\frac{\partial}{\partial t} ds = -kF \, ds;$
- (ii) $\frac{\partial}{\partial t} k = F_{ss} + k^2 F;$
- (iii) $\frac{\partial}{\partial t} k_s = F_{sss} + k^2 F_s + 3kk_s F;$

(iv)

$$\begin{aligned} \partial_t k_{s^l} = & k_{s^{l+6}} + \sum_{q+r+u=l} (c_{qru}^1 k_{s^{q+4}} k_{s^r} k_{s^u} + c_{qru}^2 k_{s^{q+3}} k_{s^{r+1}} k_{s^u} \\ & + c_{qru}^3 k_{s^{q+2}} k_{s^{r+2}} k_{s^u} + c_{qru}^4 k_{s^{q+2}} k_{s^{r+1}} k_{s^{u+1}}) \\ & + \sum_{a+b+c+d+e=l} c_{abcde} k_{s^a} k_{s^b} k_{s^c} k_{s^d} k_{s^e} \end{aligned}$$

for constants $c_{qru}^1, c_{qru}^2, c_{qru}^3, c_{qru}^4, c_{abcde} \in \mathbb{R}$ with $a, b, c, d, e, q, r, u \geq 0$.
In particular,

(v)

$$\begin{aligned} \frac{\partial}{\partial t} k_{ss} = & k_{s^8} + 10k_s k_{s^3} k_{ss} + \frac{21}{2} k_s^2 k_{s^4} + 12k k_{s^5} k_s + 14k k_{s^4} k_{ss} + 5k k_{s^3}^2 + 2k^2 k_{s^6} \\ & + \frac{11}{2} k^2 k_s^2 k_{ss} + 8k^3 k_{s^3} k_s + 5k^3 k_{ss}^2 + k^4 k_{s^4} - 4k k_s^4. \end{aligned}$$

3 Controlling the Geometry of the Flow

We begin with the following standard result for functions of one variable.

Lemma 3.1 (Poincaré-Sobolev-Wirtinger (PSW) Inequalities) *Suppose $f : [0, L] \rightarrow \mathbb{R}$, $L > 0$ is absolutely continuous.*

- If $\int_0^L f ds = 0$ then

$$\int_0^L f^2 ds \leq \frac{L^2}{\pi^2} \int_0^L f_s^2 ds \text{ and } \|f\|_\infty^2 \leq \frac{2L}{\pi} \int_0^L f_s^2 ds.$$

- Alternatively, if $f(0) = f(L) = 0$ then

$$\int_0^L f^2 ds \leq \frac{L^2}{\pi^2} \int_0^L f_s^2 ds \text{ and } \|f\|_\infty^2 \leq \frac{L}{\pi} \int_0^L f_s^2 ds.$$

To state the interpolation inequality we will use, we first need to set up some notation. For normal tensor fields S and T we denote by $S \star T$ any linear combination of S and T . In our setting, S and T will be simply curvature k or its arc length derivatives. Denote by $P_n^m(k)$ any linear combination of terms of type $\partial_s^{i_1} k \star \partial_s^{i_2} k \star \dots \star \partial_s^{i_n} k$ where $m = i_1 + \dots + i_n$ is the total number of derivatives.

The following interpolation inequality for closed curves appears in [3], for our setting with boundary we refer to [2].

Proposition 3.2 *Let $\gamma : I \rightarrow \mathbb{R}^2$ be a smooth curve. Then for any term $P_n^m(k)$ with $n \geq 2$ that contains derivatives of k of order at most $\ell - 1$,*

$$\int_I |P_n^m(k)| ds \leq c L^{1-m-n} \|k\|_2^{n-m} \|k\|_{\ell,2}^p$$

where $p = \frac{1}{\ell} \left(m + \frac{1}{2}n - 1 \right)$ and $c = c(\ell, m, n)$. Moreover, if $m + \frac{1}{2} < 2\ell + 1$ then $p < 2$ and for any $\varepsilon > 0$,

$$\int_I |P_n^m(k)| ds \leq \varepsilon \int_I |\partial_s^\ell k|^2 ds + c \varepsilon^{\frac{-p}{2-p}} \left(\int_I |k|^2 ds \right)^{\frac{n-p}{2-p}} + c \left(\int_I |k|^2 ds \right)^{m+n-1}.$$

Our first result concerns the winding number of the evolving curve γ . In view of the Neumann boundary condition, in our setting the winding number must be a multiple of $\frac{1}{2}$.

Lemma 3.3 *Under the flow (3), $\omega(t) = \omega(0)$.*

Proof We compute using Lemma 2.3 (i)

$$\frac{d}{dt} \int k ds = \int F_{ss} ds + \int k^2 F ds - \int k^2 F ds = 0,$$

so ω is constant under the flow. □

Remarks

- It follows immediately that the average curvature \bar{k} satisfies

$$\bar{k} := \frac{1}{L} \int_\gamma k ds \equiv 0$$

under the flow (3). This is important for applying the inequalities of Lemma 3.1.

- Unlike the situation in [12], here small energy does not automatically imply that the winding number is close to zero. Indeed, one may add loops (or half-loops) of circles that contribute an arbitrarily small amount of the energy $L^3 \|k_s\|_2^2$. Note that such loops must all be similarly oriented, as a change in contribution from positive to negative winding will necessitate a quantum of energy (for example a figure-8 style configuration with $\omega = 0$ can not have small energy despite comprising essentially only mollified arcs of circles).

Next we give an estimate on the length of the evolving curve in the case of small initial energy. Of course, this result does not require the energy as small as (7).

Lemma 3.4 *Under the flow (3) with $\omega(0) = 0$,*

$$\frac{d}{dt} L[\gamma(t)] \leq 0.$$

Proof We compute using integration by parts

$$\frac{d}{dt}L[\gamma(t)] = - \int kF ds = - \int k_{ss}^2 ds + \frac{7}{2} \int k^2 k_s^2 ds \leq - \left[1 - \frac{7L^3}{\pi^3} \|k_s\|_2^2 \right] \int k_{ss}^2 ds$$

where we have used Lemma 3.1. The result follows by the small energy assumption. \square

Thus under the small energy assumption we have the length of the evolving curve bounded above and below:

$$|e| \leq L[\gamma] \leq L_0.$$

We are now ready to show that the L^2 -norm of curvature remains bounded, independent of time.

Proposition 3.5 *Under the flow (3) with $\omega(0) = 0$, there exists a universal $C > 0$ such that*

$$\|k\|_2^2 \leq \|k\|_2^2|_{t=0} + C.$$

Proof Using integration by parts, Lemma 2.3 and the interpolation inequality Proposition 3.2

$$\begin{aligned} \frac{d}{dt} \int k^2 ds &= -2 \int k_{s^3}^2 ds + 5 \int k_{ss}^2 k^2 ds + 5 \int k_{ss} k_s^2 k ds + \int k_{ss} k^5 ds - \frac{1}{2} \int k_s^2 k^4 ds \\ &\leq (-2 + 3\varepsilon) \int k_{s^3}^2 ds + C \|k\|_2^{14} \leq -\frac{\pi^6}{L_0^6} \int k^2 ds + \frac{C\pi^7}{|e|}, \end{aligned}$$

where we have also used Lemma 3.1 and the length bounds. The result follows. \square

Moreover, we may show similarly using the evolution equation for k_{s^ℓ} that all derivatives of curvature are bounded in L^2 independent of time.

Proposition 3.6 *Under the flow (3) with $\omega(0) = 0$, there exists a universal $C > 0$ such that, for all $\ell \in \mathbb{N}$,*

$$\|k_{s^\ell}\|_2^2 \leq \|k_{s^\ell}\|_2^2|_{t=0} + C.$$

Pointwise bounds on all derivatives of curvature follow from Lemma 3.1. It follows that the solution of the flow remains smooth up to and including the final time, from which we may (if $T < \infty$) apply again local existence. This shows that the flow exists for all time, that is, $T = \infty$.

4 Exponential Convergence

Using Lemma 2.3 (i) and (v) and integrating by parts to reduce the order of the derivatives we obtain

Lemma 4.1 *Under the flow (3),*

$$\begin{aligned} \frac{d}{dt} \int k_{ss}^2 ds &= -2 \int k_{ss}^2 ds + 4 \int k^2 k_{s4}^2 ds - \int k_s^2 k_{s3}^2 ds - 8 \int k k_{ss} k_{s3}^2 ds - 2 \int k^4 k_{s3}^2 ds \\ &\quad + \frac{1}{3} \int k_{ss}^4 ds - \frac{1}{2} \int k^2 k_s^2 k_{ss}^2 ds + 5 \int k^3 k_{ss}^3 ds + \frac{8}{5} \int k_s^6 ds. \end{aligned}$$

Further integration by parts, use of Lemma 3.1 and throwing away some negative terms gives

Corollary 4.2 *Under the flow (3) with $\omega(0) = 0$,*

$$\frac{d}{dt} \int k_{ss}^2 ds \leq \left[-2 + \frac{67L^3}{2\pi^3} \|k_s\|_2^2 + \frac{20L^6}{\pi^6} \|k_s\|_2^4 \right] \|k_{s5}\|_2^2.$$

Under the small energy condition (7), the coefficient of $\|k_{s5}\|_2^2$ of Corollary 4.2 is bounded above by $-\delta$. Using also Lemma 3.1 we obtain

Corollary 4.3 *There exists a $\delta > 0$ such that, under the flow,*

$$\frac{d}{dt} \|k_{s5}\|_2^2 \leq -\delta \|k_{s5}\|_2^2.$$

It follows that $\|k_{s5}\|_2^2$ decays exponentially to zero.

Proof (Completion of the Proof of Theorem 2.2) Exponential decay of $\|k_{s5}\|_2^2$ implies exponential decay of $\|k\|_2^2, \|k_s\|_2^2, \|k\|_\infty, \|k_s\|_\infty$ via Lemma 3.1. Exponential decay of $\|k_{s\ell}\|_2$ and $\|k_{s\ell}\|_\infty$ then follows by a standard induction argument involving integration by parts and the curvature bounds of Propositions 3.5 and 3.6. That $\|k_s\|_2^2 \rightarrow 0$ implies subsequential convergence to straight line segments (horizontal, in view of boundary conditions). A stability argument (see [12] for the details of a similar argument) gives that in fact the limiting straight line is unique; all eigenvalues of the linearised operator

$$\mathcal{L}u = u_{x6}$$

are negative apart from the first zero eigenvalue, which corresponds precisely to vertical translations. By Hale-Raugel’s convergence theorem [6] uniqueness of the

limit follows. Although we don't know the precise height of the limiting straight line segment, we can estimate a-priori its distance from the initial curve, since

$$|\gamma(x, t) - \gamma(x, 0)| = \left| \int_0^t \frac{\partial \gamma}{\partial t}(x, \tau) d\tau \right| \leq \int_0^t |F| d\tau \leq \frac{C}{\delta} (1 - e^{-\delta t}).$$

□

Acknowledgements The research of the first and second authors was supported by Discovery Project grant DP150100375 of the Australian Research Council. The research was conducted while the first author was a staff member of the University of Wollongong. He completed part of this research at the MATRIX facilities in Cresswick and gratefully acknowledges their support. The research of the third author was supported by a University of Wollongong Faculty of Engineering and Information Sciences Postgraduate research scholarship.

References

1. Dall'Acqua, A., Lin, C.C., Pozzi, P.: Evolution of open elastic curves in \mathbb{R}^n subject to fixed length and natural boundary conditions. *Analysis* **34**(2), 209–222 (2014)
2. Dall'Acqua, A., Pozzi, P.: A Willmore-Helfrich L^2 -flow of curves with natural boundary conditions. *Comm. Anal. Geom.* **22**(4), 1485–1508 (2014)
3. Dziuk, G., Kuwert, E., Schätzle, R.: Evolution of elastic curves in \mathbb{R}^n : existence and computation. *SIAM J. Math. Anal.* **33**(5), 1228–1245 (2002)
4. Edwards, M., Gerhardt-Bourke, A., McCoy, J., Wheeler, G., Wheeler, V.-M.: The shrinking figure eight and other solitons for the curve diffusion flow. *J. Elast.* **119**(1–2), 191–211 (2014)
5. Giga, Y., Ito, K.: Loss of convexity of simple closed curves moved by surface diffusion. In: Escher, J., Simonett, G. (eds.) *Topics in Nonlinear Analysis, the Herbert Amann Anniversary Volume, Progress in Nonlinear Differential Equations and Their Applications*, vol. 35, pp. 305–320. Birkhäuser, Basel (1999)
6. Hale, J., Raugel, G.: Convergence in gradient-like systems with applications to PDE. *Z. Angew. Math. Phys.* **43**, 63–124 (1992)
7. Lin, C.C.: L^2 -flow of elastic curves with clamped boundary conditions. *J. Differ. Equ.* **252**(12), 6414–6428 (2012)
8. Liu, D., Xu, G.: A general sixth order geometric partial differential equation and its application in surface modeling. *J. Inf. Comp. Sci.* **4**, 1–12 (2007)
9. Parkins, S., Wheeler, G.: The polyharmonic heat flow of closed plane curves. *J. Math. Anal. Appl.* **439**, 608–633 (2016)
10. Ugail, H., Wilson, M.: Modeling of oedemous limbs and venous ulcers using partial differential equations. *Theor. Biol. Med. Model.* **2**(28) (2005)
11. Wheeler, G.: On the curve diffusion flow of closed plane curves. *Anali di Matematica* **192**, 931–950 (2013)
12. Wheeler, G., Wheeler, V.-M.: Curve diffusion and straightening flows on parallel lines. Preprint

Quasilinear Parabolic and Elliptic Equations with Singular Potentials



Maria Michaela Porzio

Abstract In this paper we describe the asymptotic behavior of the solutions to quasilinear parabolic equations with a Hardy potential. We prove that all the solutions have the same asymptotic behavior: they all tend to the solution of the original problem which satisfies a zero initial condition. Moreover, we derive estimates on the “distance” between the solutions of the evolution problem and the solutions of elliptic problems showing that in many cases (as for example the autonomous case) these last solutions are “good approximations” of the solutions of the original parabolic PDE.

1 Introduction

Let us consider the following nonlinear parabolic problem

$$\begin{cases} u_t - \operatorname{div}(a(x, t, \nabla u)) = \lambda \frac{u}{|x|^2} + f(x, t) & \text{in } \Omega_T \equiv \Omega \times (0, T), \\ u(x, t) = 0 & \text{on } \partial\Omega \times (0, T), \\ u(x, 0) = u_0(x) & \text{in } \Omega, \end{cases} \quad (1)$$

where $\Omega \subset \mathbb{R}^N$ ($N \geq 3$) is a bounded domain containing the origin and λ and T are positive constants.

Here $a(x, t, \xi) : \Omega \times \mathbb{R}^+ \times \mathbb{R}^N \rightarrow \mathbb{R}^N$ is a Caratheodory function¹ satisfying

$$a(x, t, \xi)\xi \geq \alpha|\xi|^2, \quad \alpha > 0, \quad (2)$$

¹That is, it is continuous with respect to ξ for almost every $(x, t) \in \Omega_T$, and measurable with respect to (x, t) for every $\xi \in \mathbb{R}^N$.

M. M. Porzio (✉)

Dipartimento di Matematica “Guido Castelnuovo”, Sapienza Università di Roma, Roma, Italy

e-mail: porzio@mat.uniroma1.it

$$|a(x, t, \xi)| \leq \beta[|\xi| + \mu(x, t)], \quad \beta > 0, \quad \mu \in L^2(\Omega_T), \tag{3}$$

$$(a(x, t, \xi) - a(x, t, \xi')) \cdot (\xi - \xi') \geq \alpha|\xi - \xi'|^2, \tag{4}$$

and the data satisfy (for example)

$$u_0 \in L^2(\Omega) \quad f \in L^2(\Omega_T). \tag{5}$$

The model problem we have in mind is the following

$$\begin{cases} u_t - \Delta u = \lambda \frac{u}{|x|^2} + f(x, t) & \text{in } \Omega_T, \\ u(x, t) = 0 & \text{on } \partial\Omega \times (0, T), \\ u(x, 0) = u_0(x) & \text{in } \Omega. \end{cases} \tag{6}$$

We recall that if the data f and u_0 are nonnegative and not both identically zero, there exists a dimension dependent constant Λ_N such that (6) has no solution for $\lambda > \Lambda_N$ (see [10]). More in details, the constant Λ_N is the optimal constant (not attained) in the Hardy’s inequality

$$\Lambda_N \int_{\Omega} \frac{u^2}{|x|^2} dx \leq \int_{\Omega} |\nabla u|^2 dx \quad \text{for every } u \in H_0^1(\Omega) \quad \text{where} \quad \Lambda_N \equiv \left(\frac{N-2}{2}\right)^2, \tag{7}$$

(see [24] and [20]).

Hence, here, in order to guarantee the existence of solutions, we assume $\lambda < \Lambda_N$ in the model case (6) and its generalization

$$\lambda < \alpha \Lambda_N. \tag{8}$$

in the general case (1).

The main aim of this paper is the study of the asymptotic behavior of the solutions of (1).

The peculiarity of these problems is the presence of the singular Hardy potential, also called in literature “inverse-square” potential. This kind of singular potential arises, for example, in the context of combustion theory (see [11, 47] and the references therein) and quantum mechanics (see [10, 44, 47] and the references therein).

There is an extensive literature on problems with Hardy potentials both in the stationary and evolution cases and it is a difficult task to give a complete bibliography. In the elliptic case, more related to our framework are [1, 2, 4, 5, 15, 35, 41, 49, 50] and [7]. In the parabolic case, a mile stone is the pioneer paper [10] which revealing the surprising effects of these singular potentials on the solutions stimulated the study of these problems. More connected to our results are [3, 6, 19, 25, 26, 43, 45, 47] and [38].

In particular, in [43] it is studied the influence of the regularity of the data f and u_0 on the regularity of the solutions of (1), while in [47] and [38], among other results, there is a description of the behavior (in time) of the solutions when $f \equiv 0$.

Hence, here we want to complete these results studying what is the asymptotic behavior of the solutions when f is not identically zero.

We point out that the presence of a singular potential term has a strong influence not only, as recalled above, on the existence theory, but also on the regularity and on the asymptotic behavior, even when the datum f is zero. As a matter of fact, it is well known that if $\lambda \equiv 0 = f$ and the initial datum u_0 is bounded then also the solution of (1) is bounded; moreover, this result remains true in the more general case of non zero data f belonging to $L^r(0, T; L^q(\Omega))$ with r and q satisfying

$$\frac{1}{r} + \frac{N}{2q} < 1 \tag{9}$$

(see [9] and the references therein).

Surprisingly, the previous L^∞ -regularity fails even in the model case (6) as soon as λ becomes positive if f and u_0 are not both identically zero (otherwise $u \equiv 0$ is a bounded solution) since every solution (for nonnegative initial data u_0) satisfies²

$$u(x, t) \geq \frac{C}{|x|^{\alpha_1}} \quad \text{for almost every } (x, t) \in \Omega' \times [\varepsilon, \hat{T}], \tag{10}$$

for every $\varepsilon \in (0, \hat{T})$, $0 < \hat{T} < T$ and $\Omega' \subset\subset \Omega$, where the constant C depends only on $\varepsilon, \hat{T}, \Omega'$ and λ , while α_1 is the smallest root of $z^2 - (N - 2)z + \lambda = 0$.

Indeed, the singular potential term influences the solutions also when the summability coefficients r and q of f do not satisfy (9). As a matter of fact, again the regularity of the solutions in presence of the Hardy potential is different from the classical semilinear case $\lambda = 0$ (see [43] if $\lambda > 0$ and [14, 16, 27, 31, 33, 34] and the references therein if $\lambda = 0$).

Great changes appear also in the behavior in time of the solutions. As a matter of fact, if $\lambda = 0 = f(x, t)$ it is well known that the solutions of (1) become immediately bounded also in presence of unbounded initial data u_0 belonging only to $L^{r_0}(\Omega)$ ($r_0 \geq 1$) and satisfy the same decay estimates of the heat equation

$$\|u(t)\|_{L^\infty(\Omega)} \leq c \frac{\|u_0\|_{L^{r_0}(\Omega)}}{t^{\frac{N}{2r_0}} e^{\sigma t}} \quad \text{for almost every } t \in (0, T), \tag{11}$$

²The proof of (10) can be easily obtained following the outline of the proof of (2.5) of Theorem 2.2 in [10].

where $\sigma = \frac{c}{|\Omega|^{\frac{1}{N}}}$ is a constant depending on the measure of Ω (see [36] and the references therein). The previous bound, or more in general estimates of the type

$$\|u(t)\|_{L^\infty(\Omega)} \leq c \frac{\|u_0\|_{L^{h_0}(\Omega)}^{h_0}}{t^{h_1}} \quad h_0, h_1 > 0, \tag{12}$$

are often referred as ultracontractive estimates and hold for many different kinds of parabolic PDE (degenerate or singular) like, for example, the p-Laplacian equation, the fast diffusion equation, the porous medium equation etc. These estimates are widely studied because they describe the behavior in time of the solutions and often imply also further important properties like, for example, the uniqueness (see for example [8, 12, 17, 18, 21–23, 28–30, 36, 37, 40, 42, 46, 48] and the references therein).

Unfortunately, by estimate (10) above it follows that estimate (11) together with (12) fail in presence of a Hardy potential term. Anyway, in [38] it is proved that if $f \equiv 0$, $\lambda > 0$ and $u_0 \in L^2(\Omega)$, then there exists a solution that satisfies

$$\|u(t)\|_{L^{2\gamma}(\Omega)} \leq c \frac{\|u_0\|_{L^2(\Omega)}}{t^\delta e^{\sigma t}} \quad \text{for almost every } t \in (0, T), \quad \delta = \frac{N(\gamma - 1)}{4\gamma}, \tag{13}$$

for every $\gamma > 1$ satisfying

$$\gamma \in \left(1, \frac{1 + \sqrt{1 - \theta}}{\theta} \right) \quad \text{where } \theta = \frac{\lambda}{\alpha \Lambda_N}.$$

Hence an increasing of regularity appears (depending on the “size” λ of the singular potential), but according with (10), there is not the boundedness of the solutions.

As said above, aim of this paper is to describe what happens when f is not identically zero.

We will show that under the previous assumptions on the operator a and on the data f and u_0 , there exists only one “good” global solution u of (1). Moreover, if v is the global solution of

$$\begin{cases} v_t - \operatorname{div}(a(x, t, \nabla v)) = \lambda \frac{v}{|x|^2} + f(x, t) & \text{in } \Omega \times (0, +\infty), \\ v(x, t) = 0 & \text{on } \partial\Omega \times (0, +\infty), \\ v(x, 0) = v_0(x) & \text{in } \Omega, \end{cases} \tag{14}$$

i.e., v satisfies the same PDE of u (with the same datum f) but verifies the different initial condition $v(x, 0) = v_0 \in L^2(\Omega)$, then the following estimate holds

$$\|u(t) - v(t)\|_{L^2(\Omega)} \leq \frac{\|u_0 - v_0\|_{L^2(\Omega)}}{e^{\sigma t}} \quad \text{for every } t > 0, \tag{15}$$

where σ is a positive constant which depends on λ (see formula (26) below). In particular, it results

$$\lim_{t \rightarrow +\infty} \|u(t) - v(t)\|_{L^2(\Omega)} = 0. \tag{16}$$

Hence, for t large, the initial data do not influence the behavior of the solutions since by (16) it follows that all the global solutions tend to the solution which assumes the null initial datum.

We recall that in absence of the singular potential term we can replace the L^2 -norm in the left-hand side of (15) with the L^∞ -norm and, consequently, together with (16), the following stronger result holds true

$$\lim_{t \rightarrow +\infty} \|u(t) - v(t)\|_{L^\infty(\Omega)} = 0. \tag{17}$$

(see [39]). Thus, the presence of the Hardy potential provokes again a change in the behavior of the solutions since generally the difference of two solutions u and v cannot be bounded if $\lambda > 0$. As a matter of fact, it is sufficient to notice that choosing $f = 0$ and $v = 0$ (which corresponds to the choice $v_0 = 0$) the boundedness of $u-v$ becomes the boundedness of u which by (10) we know to be false in presence of a Hardy potential.

Moreover, in the autonomous case

$$a(x, t, \xi) = a(x, \xi) \quad f(x, t) = f(x)$$

we prove that all the global solutions of (1) (whatever is the value of the initial datum u_0) tend to the solution $w \in H_0^1(\Omega)$ of the associate elliptic problem

$$\begin{cases} -\operatorname{div}(a(x, \nabla w)) = \lambda \frac{w}{|x|^2} + f(x) & \text{in } \Omega, \\ w(x) = 0 & \text{on } \partial\Omega. \end{cases}$$

Indeed, we estimate also the difference $u - v$ between the global solutions of (1) and the global solution v of the different evolution problem (not necessarily of parabolic type)

$$\begin{cases} v_t - \operatorname{div}(b(x, t, \nabla v)) = \lambda \frac{v}{|x|^2} + F(x, t) & \text{in } \Omega \times (0, +\infty), \\ v(x, t) = 0 & \text{on } \partial\Omega \times (0, +\infty), \\ v(x, 0) = v_0(x) & \text{in } \Omega, \end{cases}$$

looking for conditions which guarantee that this difference goes to zero (letting $t \rightarrow +\infty$).

Finally, we estimate also the difference $u - w$ between a global solution of (1) and the solutions w of the stationary problem

$$\begin{cases} -\operatorname{div}(b(x, \nabla w)) = \lambda \frac{w}{|x|^2} + F(x) & \text{in } \Omega, \\ w(x) = 0 & \text{on } \partial\Omega, \end{cases} \tag{18}$$

showing that in the non autonomous case, under suitable “proximity” conditions on the operators a and b and the data f and F , the global solution of (1) tends to the solution w of the stationary problem (18).

The paper is organized as follows: in next section we give the statements of our results in all the details. The proofs can be found in Sect. 4 and make use of some “abstract results” proved in [38] and [32] that, for the convenience of the reader, we recall in Sect. 3.

2 Main Results

Before stating our results, we recall the definitions of solution and global solution of (1).

Definition 1 Assume (2)–(5). A function u in $L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$ is a solution of (1) if it results

$$\int_0^T \int_\Omega \{-u\varphi_t + a(x, t, \nabla u)\nabla\varphi\} dxdt = \int_\Omega u_0\varphi(x, 0) dx + \int_0^T \int_\Omega \left[\lambda \frac{u}{|x|^2} + f\right] \varphi dxdt \tag{19}$$

for every $\varphi \in W^{1,1}(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$ satisfying $\varphi(T) = 0$.

We point out that all the integrals in (19) are well defined. As a matter of fact, by (3) it follows that $a(x, t, \nabla u) \in (L^2(\Omega_T))^N$ and thanks to Hardy’s inequality (7) it results

$$\begin{aligned} \int_0^T \int_\Omega \lambda \frac{u}{|x|^2} \varphi dxdt &\leq \lambda \left(\int_0^T \int_\Omega \frac{u^2}{|x|^2}\right)^{\frac{1}{2}} \left(\int_0^T \int_\Omega \frac{\varphi^2}{|x|^2}\right)^{\frac{1}{2}} \leq \\ &\frac{\lambda}{\Lambda_N} \|\nabla u\|_{L^2(\Omega_T)} \|\nabla \varphi\|_{L^2(\Omega_T)}. \end{aligned} \tag{20}$$

We recall that under the assumptions (2)–(5) and (8) there exist solutions of (1) (see [43]). Now, to extend the previous notion to that of global solution, we assume

$$f \in L^2_{loc}([0, +\infty); L^2(\Omega)) \quad \text{and} \quad \mu \in L^2_{loc}([0, +\infty); L^2(\Omega)) \tag{21}$$

where μ is the function that appears in (3).

Definition 2 By a global solution of (1), or (equivalently) of

$$\begin{cases} u_t - \operatorname{div}(a(x, t, \nabla u)) = \lambda \frac{u}{|x|^2} + f(x, t) & \text{in } \Omega \times (0, +\infty) \\ u(x, t) = 0 & \text{on } \partial\Omega \times (0, +\infty), \\ u(x, 0) = u_0(x) & \text{in } \Omega, \end{cases} \quad (22)$$

we mean a measurable function u that is a solution of (1) for every $T > 0$ arbitrarily chosen.

We point out that (21) together with the previous structure assumptions guarantee that the integrals in (19) are well defined for every choice of $T > 0$. Indeed, there exists only one global solution of (1). In detail, we have:

Theorem 1 *Assume (2)–(5), (8) and (21). Then there exists only one global solution u of (1) belonging to $C_{loc}([0, +\infty); L^2(\Omega)) \cap L^2_{loc}([0, +\infty); H^1_0(\Omega))$. In particular, for every $t > 0$ it results*

$$\begin{aligned} & \int_0^t \int_{\Omega} \{-u\varphi_t + a(x, t, \nabla u)\nabla\varphi\} dxdt + \int_{\Omega} [u(x, t)\varphi(x, t) - u_0\varphi(x, 0)] dx = \\ & \int_0^t \int_{\Omega} \left[\lambda \frac{u}{|x|^2} + f \right] \varphi dxdt, \end{aligned} \quad (23)$$

for every $\varphi \in W^{1,1}_{loc}([0, +\infty); L^2(\Omega)) \cap L^2_{loc}([0, +\infty); H^1_0(\Omega))$.

As noticed in the introduction, if we change the initial data in (1), all the associated global solutions (to these different initial data) have the same asymptotic behavior. In detail, let us consider the following problem

$$\begin{cases} v_t - \operatorname{div}(a(x, t, \nabla v)) = \lambda \frac{v}{|x|^2} + f(x, t) & \text{in } \Omega \times (0, +\infty), \\ v(x, t) = 0 & \text{on } \partial\Omega \times (0, +\infty), \\ v(x, 0) = v_0(x) & \text{in } \Omega. \end{cases} \quad (24)$$

We have the following result:

Theorem 2 *Assume (2)–(5), (8) and (21). If $v_0 \in L^2(\Omega)$, then the global solutions u and v of, respectively, (1) and (24) belonging to $C_{loc}([0, +\infty); L^2(\Omega)) \cap L^2_{loc}([0, +\infty); H^1_0(\Omega))$ satisfy*

$$\|u(t) - v(t)\|_{L^2(\Omega)} \leq \frac{\|u_0 - v_0\|_{L^2(\Omega)}}{e^{\sigma t}} \quad \text{for every } t > 0, \quad (25)$$

where

$$\sigma = \left(\alpha - \frac{\lambda}{\Lambda_N} \right) c_P \quad (26)$$

with c_P Poincaré’s constant.³
 In particular, it results

$$\lim_{t \rightarrow +\infty} \|u(t) - v(t)\|_{L^2(\Omega)} = 0. \tag{28}$$

Remark 1 Notice that in the particular case $f \equiv 0$, choosing as initial datum $v_0 = 0$ we obtain that $v \equiv 0$ is the global solution of (1). With such a choice in (25) it follows that

$$\|u(t)\|_{L^2(\Omega)} \leq \frac{\|u_0\|_{L^2(\Omega)}}{e^{\sigma t}} \quad \text{for every } t > 0.$$

In the model case (6) the previous estimate can be found (among other interesting results) in [47] with $\sigma = \mu_1$ the first eigenvalue (see also [38]). We recall that decay estimates of the solutions in the same Lebesgue space where is the initial datum is not a peculiarity of problems with singular potentials since appear also for other parabolic problems (see [13, 38, 46] and the references therein).

An immediate consequence of Theorem 2 is that in the autonomous case

$$a(x, t, \xi) = a(x, \xi) \quad f(x, t) = f(x) \tag{29}$$

all the global solutions of (1), whatever is the value of the initial datum u_0 , tend (letting $t \rightarrow +\infty$) to the solution w of the associate elliptic problem

$$\begin{cases} -\operatorname{div}(a(x, \nabla w)) = \lambda \frac{w}{|x|^2} + f(x) & \text{in } \Omega, \\ w(x) = 0 & \text{on } \partial\Omega. \end{cases} \tag{30}$$

In detail, we have:

Corollary 1 (Autonomous Case) *Assume (2)–(5), (8), (21) and (29). Let w be the unique solution of (30) in $H_0^1(\Omega)$ and u be the global solution of (1) belonging to $C_{loc}([0, +\infty); L^2(\Omega)) \cap L_{loc}^2([0, +\infty); H_0^1(\Omega))$. Then it results*

$$\|u(t) - w\|_{L^2(\Omega)} \leq \frac{\|u_0 - w\|_{L^2(\Omega)}}{e^{\sigma t}} \quad \text{for every } t > 0, \tag{31}$$

where σ is as in (26).

³Poincaré’s inequality:

$$c_P \int_{\Omega} u^2 dx \leq \int_{\Omega} |\nabla u|^2 dx \quad \text{for every } u \in H_0^1(\Omega), \tag{27}$$

where c_P is a constant depending only on N and on the bounded set Ω .

In particular, it results

$$\lim_{t \rightarrow +\infty} \|u(t) - w\|_{L^2(\Omega)} = 0. \tag{32}$$

We show now that it is also possible to estimate the distance between the global solution u of (1) and the global solution v of the different parabolic problem

$$\begin{cases} v_t - \operatorname{div}(b(x, t, \nabla v)) = \lambda \frac{v}{|x|^2} + F(x, t) & \text{in } \Omega \times (0, +\infty), \\ v(x, t) = 0 & \text{on } \partial\Omega \times (0, +\infty), \\ v(x, 0) = v_0(x) & \text{in } \Omega, \end{cases} \tag{33}$$

where $b(x, t, \xi) : \Omega \times \mathbb{R}^+ \times \mathbb{R}^N \rightarrow \mathbb{R}^N$ is a Caratheodory function satisfying

$$b(x, t, \xi)\xi \geq \alpha_0 |\xi|^2, \quad \alpha_0 > 0, \tag{34}$$

$$|b(x, t, \xi)| \leq \beta_0 [|\xi| + \mu_0(x, t)], \quad \beta_0 > 0, \quad \mu_0 \in L^2_{loc}([0, +\infty); L^2(\Omega)), \tag{35}$$

$$(b(x, t, \xi) - b(x, t, \xi')) \cdot (\xi - \xi') \geq \alpha_0 |\xi - \xi'|^2. \tag{36}$$

$$v_0 \in L^2(\Omega) \quad F \in L^2_{loc}([0, +\infty); L^2(\Omega)). \tag{37}$$

Theorem 3 Assume (2)–(5), (8), (21) and (34)–(37). Then the global solutions u and v of, respectively, (1) and (33) belonging to $C_{loc}([0, +\infty); L^2(\Omega)) \cap L^2_{loc}([0, +\infty); H^1_0(\Omega))$ satisfy

$$\|u(t) - v(t)\|_{L^2(\Omega)}^2 \leq \frac{\|u_0 - v_0\|_{L^2(\Omega)}^2}{e^{2\sigma_0 t}} + \int_0^t g(s) ds \quad \text{for every } t \geq 0, \tag{38}$$

for every choice of

$$\sigma_0 < \sigma \tag{39}$$

where

$$g(s) = \frac{c_P}{\sigma - \sigma_0} \int_{\Omega} \left[|b(x, s, \nabla v(x, s)) - a(x, s, \nabla v(x, s))|^2 + \frac{1}{c_P} |f(x, s) - F(x, s)|^2 \right] dx, \tag{40}$$

with σ and c_P are as in (26). Moreover, if $g \in L^1((0, +\infty))$ then it results

$$\|u(t) - v(t)\|_{L^2(\Omega)}^2 \leq \frac{\Lambda}{e^{\sigma_0 t}} + \int_{\frac{t}{2}}^t g(s) ds \quad \text{for every } t > 0, \tag{41}$$

where

$$\Lambda = \|u_0 - v_0\|_{L^2(\Omega)}^2 + \int_0^{+\infty} g(t)dt .$$

In particular, we have

$$\lim_{t \rightarrow +\infty} \|u(t) - v(t)\|_{L^2(\Omega)} = 0 . \tag{42}$$

Remark 2 The proof of Theorem 3 shows that the structure assumptions (34)–(36) on the operator b can be weakened. In particular, it is sufficient to assume that there exists a global solution v of (33) in $C_{loc}([0, +\infty); L^2(\Omega)) \cap L^2_{loc}([0, +\infty); H^1_0(\Omega))$ satisfying

$$b(x, t, \nabla v) \in L^2_{loc}([0, +\infty); L^2(\Omega)) .$$

Hence, also problems (33) which are not of parabolic type are allowed.

Moreover, with slight changes in the proof, it is also possible to choose a larger class of data f and F . In particular, an alternative option that can be done is $L^2_{loc}([0, +\infty); H^{-1}(\Omega))$.

Examples of operators satisfying all the assumptions of the previous Theorem (and hence for which (42) holds) are

$$\begin{cases} u_t - \operatorname{div}(a(x, t, \nabla u)) = \lambda \frac{u}{|x|^2} + f(x, t) & \text{in } \Omega \times (0, +\infty), \\ u(x) = 0 & \text{on } \partial\Omega \times (0, +\infty) \\ u(x, 0) = u_0 & \text{in } \Omega . \end{cases}$$

and the model case

$$\begin{cases} v_t - \Delta v = \lambda \frac{u}{|x|^2} + F(x, t) & \text{in } \Omega \times (0, +\infty), \\ v(x) = 0 & \text{on } \partial\Omega \times (0, +\infty) \\ v(x, 0) = v_0 & \text{in } \Omega , \end{cases}$$

if we assume

$$[a(x, t, \nabla v) - \nabla v] \in L^2(\Omega \times (0, +\infty)) \quad [f(x, t) - F(x, t)] \in L^2(\Omega \times (0, +\infty)) .$$

Remark 3 We point out that an admissible choice for the parameter λ in Theorem 3 is

$$\lambda = 0 ,$$

i.e., the case of absence of the singular potential. In this particular but also interesting case, the previous result permits to estimate the difference of solutions of different evolution problems. Moreover, as noticed in Remark 2, only one of these two evolution problems is required to be of parabolic type.

A consequence of Theorem 3 is the possibility to estimate also the distance between global solutions of (1) and solutions of stationary problems, for example of elliptic type, without assuming to be in the autonomous case (29). These estimates (see Corollary 2 below) show that if the data and the operators of these different PDE problems (calculated on the solution of the stationary problem) are “sufficiently near”, then the solutions of the evolution problems tend (for every choice of the initial data u_0) to the stationary solution.

In detail, let us consider the following stationary problem

$$\begin{cases} -\operatorname{div}(b(x, \nabla w)) = \lambda \frac{w}{|x|^2} + F(x) & \text{in } \Omega, \\ w(x) = 0 & \text{on } \partial\Omega. \end{cases} \tag{43}$$

To stress the assumptions really needed, in what follows we do not assume any structure condition on b except that $b : \Omega \times \mathbb{R}^N \rightarrow \mathbb{R}^N$ is a Caratheodory function. We have:

Corollary 2 *Assume (2)–(5), (8) and (21). Let F be in $L^2(\Omega)$ and $w \in H_0^1(\Omega)$ be such that*

$$b(x, \nabla w) \in (L^2(\Omega))^N. \tag{44}$$

If w is a solution of (43) and $u \in C_{loc}([0, +\infty); L^2(\Omega)) \cap L^2_{loc}([0, +\infty); H_0^1(\Omega))$ is the global solution of (1), then the following estimate holds true

$$\|u(t) - w\|_{L^2(\Omega)}^2 \leq \frac{\|u_0 - w\|_{L^2(\Omega)}^2}{e^{2\sigma_0 t}} + \int_0^t g(s) ds, \tag{45}$$

for every $t \geq 0$ and for every choice of σ_0 as in (39) where

$$g(s) = \frac{c_P}{\sigma - \sigma_0} \int_{\Omega} \left[|b(x, \nabla w(x)) - a(x, s, \nabla w(x))|^2 + \frac{1}{c_P} |f(x, s) - F(x)|^2 \right] dx. \tag{46}$$

Moreover, if $g \in L^1((0, +\infty))$, then it results

$$\|u(t) - w\|_{L^2(\Omega)}^2 \leq \frac{\Lambda}{e^{\sigma_0 t}} + \int_{\frac{t}{2}}^t g(s) ds \quad \text{for every } t > 0, \tag{47}$$

where

$$\Lambda = \|u_0 - w\|_{L^2(\Omega)}^2 + \int_0^{+\infty} g(t)dt .$$

In particular, it follows

$$\lim_{t \rightarrow +\infty} \|u(t) - w\|_{L^2(\Omega)} = 0 . \tag{48}$$

Examples of operators satisfying all the assumptions of Corollary 2 (and hence for which (48) holds) are

$$\begin{cases} u_t - \operatorname{div}(\alpha(x, t)\nabla u) = \lambda \frac{u}{|x|^2} + f(x, t) & \text{in } \Omega \times (0, +\infty), \\ u(x) = 0 & \text{on } \partial\Omega \times (0, +\infty) \\ u(x, 0) = u_0 & \text{in } \Omega . \end{cases}$$

and

$$\begin{cases} -\Delta w = \lambda \frac{w}{|x|^2} + F(x) & \text{in } \Omega, \\ w(x) = 0 & \text{on } \partial\Omega . \end{cases}$$

with

$$[\alpha(x, t) - 1] \in L^\infty(\Omega \times (0, +\infty)) \quad [f(x, t) - F(x)] \in L^2(\Omega \times (0, +\infty)) \tag{49}$$

or

$$\begin{cases} u_t - \operatorname{div}(\alpha(x, t)b(x, \nabla u)) = \lambda \frac{u}{|x|^2} + f(x, t) & \text{in } \Omega \times (0, +\infty), \\ u(x) = 0 & \text{on } \partial\Omega \times (0, +\infty) \\ u(x, 0) = u_0 & \text{in } \Omega . \end{cases}$$

and

$$\begin{cases} -\operatorname{div}(b(x, \nabla w)) = \lambda \frac{w}{|x|^2} + F(x) & \text{in } \Omega, \\ w(x) = 0 & \text{on } \partial\Omega . \end{cases}$$

with $\alpha(x, t)$ and the data f and F satisfying (49).

3 Preliminary Results

In this section we state two results that will be essential tools in proving the theorems presented above.

Theorem 4 (Theorem 2.8 in [38]) *Let u be in $C((0, T); L^r(\Omega)) \cap L^\infty(0, T; L^{r_0}(\Omega))$ where $0 < r \leq r_0 < \infty$. Suppose also that $|\Omega| < +\infty$ if $r \neq r_0$ (no assumption are needed on $|\Omega|$ if $r = r_0$). If u satisfies*

$$\int_{\Omega} |u|^r(t_2) - \int_{\Omega} |u|^r(t_1) + c_1 \int_{t_1}^{t_2} \|u(t)\|_{L^r(\Omega)}^r dt \leq 0 \quad \text{for every } 0 < t_1 < t_2 < T, \tag{50}$$

and there exists $u_0 \in L^{r_0}(\Omega)$ such that

$$\|u(t)\|_{L^{r_0}(\Omega)} \leq c_2 \|u_0\|_{L^{r_0}(\Omega)} \quad \text{for almost every } t \in (0, T), \tag{51}$$

where $c_i, i = 1, 2$ are real positive numbers, then the following estimate holds true

$$\|u(t)\|_{L^r(\Omega)} \leq c_4 \frac{\|u_0\|_{L^{r_0}(\Omega)}}{e^{\sigma t}} \quad \text{for every } 0 < t < T, \tag{52}$$

where

$$c_4 = \begin{cases} c_2 |\Omega|^{\frac{1}{r} - \frac{1}{r_0}} & \text{if } r < r_0, \\ 1 & \text{if } r = r_0, \end{cases} \quad \sigma = \frac{c_1}{r}.$$

Proposition 1 (Proposition 3.2 in [32]) *Assume $T \in (t_0, +\infty]$ and let $\phi(t)$ a continuous and non negative function defined in $[t_0, T)$ verifying*

$$\phi(t_2) - \phi(t_1) + M \int_{t_1}^{t_2} \phi(t) dt \leq \int_{t_1}^{t_2} g(t) dt$$

for every $t_0 \leq t_1 \leq t_2 < T$ where M is a positive constant and g is a non negative function in $L^1_{loc}([t_0, T))$. Then for every $t \in (t_0, T)$ we get

$$\phi(t) \leq \phi(t_0) e^{-M(t-t_0)} + \int_{t_0}^t g(s) ds \tag{53}$$

Moreover, if $T = +\infty$ and g belongs to $L^1((t_0, +\infty))$ there exists $t_1 \geq t_0$ (for example $t_1 = 2t_0$) such that

$$\phi(t) \leq \Lambda e^{-\frac{M}{2}t} + \int_{\frac{t}{2}}^t g(s) ds \quad \text{for every } t \geq t_1, \tag{54}$$

where

$$\Lambda = \phi(t_0) + \int_{t_0}^{+\infty} g(s)ds .$$

In particular, we get that

$$\lim_{t \rightarrow +\infty} \phi(t) = 0 .$$

4 Proofs of the Results

4.1 Proof of Theorem 1

Let $T > 0$ arbitrarily fixed. The existence of a solution $u \in L^\infty(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$ of (1) can be found in [43]. We point out that since u_t belongs to $L^2(0, T; H^{-1}(\Omega))$ (thanks to the regularity of u and (20)) it follows that u belongs also to $C([0, T]; L^2(\Omega))$. Consequently, it results

$$\begin{aligned} & \int_0^T \int_\Omega \{-u\varphi_t + a(x, t, \nabla u)\nabla\varphi\} dxdt + \int_\Omega [u(x, T)\varphi(x, T) - u_0\varphi(x, 0)] dx = \\ & \int_0^T \int_\Omega \left[\lambda \frac{u}{|x|^2} + f \right] \varphi dxdt , \end{aligned} \tag{55}$$

for every $\varphi \in W^{1,1}(0, T; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$. Moreover, u is the unique solution of (1) belonging to $C([0, T]; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$. As a matter of fact, if there exists an other solution v of (1) in $C([0, T]; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$, taking as test function $u - v$ in the equation satisfied by u and in that satisfied by v and subtracting the results⁴ we deduce (using (4))

$$\begin{aligned} & \frac{1}{2} \int_\Omega [u(x, T) - v(x, T)]^2 + \alpha \int_0^T \int_\Omega |\nabla(u - v)|^2 \leq \\ & \lambda \int_0^T \int_\Omega \frac{[u - v]^2}{|x|^2} . \end{aligned} \tag{56}$$

By the previous estimate and Hardy’s inequality (7) we obtain

$$\frac{1}{2} \int_\Omega [u(x, T) - v(x, T)]^2 + \left(\alpha - \frac{\lambda}{\Lambda_N} \right) \int_0^T \int_\Omega |\nabla(u - v)|^2 \leq 0 .$$

⁴The use here and below of these test functions can be made rigorous by means of Steklov averaging process.

from which the uniqueness follows since by assumption it results $\alpha - \frac{\lambda}{\Lambda_N} > 0$.

Hence, for every arbitrarily fixed $T > 0$ there exists a unique solution of (1) in $C([0, T]; L^2(\Omega)) \cap L^2(0, T; H_0^1(\Omega))$ that we denote $u^{(T)}$.

To conclude the proof, let us construct now the global solution of (1). For every $t \geq 0$ let us define $u(x, t) = u^{(T)}(x, t)$ where T is arbitrarily chosen satisfying $T > t$. We notice that by the uniqueness proved above this definition is well posed. Moreover, by construction this function satisfies the assertions of the theorem. \square

4.2 Proof of Theorem 2

Let u and v be as in the statement of Theorem 2. Taking as test function $u - v$ in (1) and in (24) and subtracting the equations obtained in this way, we deduce (using assumption (4)) that for every $0 \leq t_1 < t_2$ it results

$$\begin{aligned} & \frac{1}{2} \int_{\Omega} [u(x, t_2) - v(x, t_2)]^2 dx - \frac{1}{2} \int_{\Omega} [u(x, t_1) - v(x, t_1)]^2 + \alpha \int_{t_1}^{t_2} \int_{\Omega} |\nabla(u - v)|^2 \leq \\ & \lambda \int_{t_1}^{t_2} \int_{\Omega} \frac{[u - v]^2}{|x|^2}. \end{aligned} \tag{57}$$

Using again Hardy’s inequality (7), from (57) we deduce

$$\int_{\Omega} [u(x, t_2) - v(x, t_2)]^2 dx - \int_{\Omega} [u(x, t_1) - v(x, t_1)]^2 + c_0 \int_{t_1}^{t_2} \int_{\Omega} |\nabla(u - v)|^2 \leq 0, \tag{58}$$

where we have defined

$$c_0 = 2 \left(\alpha - \frac{\lambda}{\Lambda_N} \right). \tag{59}$$

Thanks to Poincaré’s inequality (27) by the previous estimate we get for every $0 \leq t_1 < t_2$

$$\int_{\Omega} [u(x, t_2) - v(x, t_2)]^2 dx - \int_{\Omega} [u(x, t_1) - v(x, t_1)]^2 + c_1 \int_{t_1}^{t_2} \int_{\Omega} |u - v|^2 \leq 0, \tag{60}$$

where $c_1 = c_P c_0$. Notice that by (60) (choosing $t_2 = t$ and $t_1 = 0$) it follows also that

$$\|u(t) - v(t)\|_{L^2(\Omega)} \leq \|u_0 - v_0\|_{L^2(\Omega)}.$$

Now the assert follows applying Theorem 4 with $r = r_0 = 2$. \square

4.3 Proof of Corollary 1

The assertion (31) follows by Theorem 2 once noticed that, thanks to the assumption (29), the solution $w \in H_0^1(\Omega)$ of (30) is also the global solution $w(x, t) \equiv w(x) \in C_{loc}([0, +\infty]; L^2(\Omega)) \cap L_{loc}^2([0, +\infty); H_0^1(\Omega))$ of the following parabolic problem

$$\begin{cases} w_t - \operatorname{div}(a(x, \nabla w)) = \lambda \frac{w}{|x|^2} + f(x) & \text{in } \Omega \times (0, +\infty), \\ w(x, t) = 0 & \text{on } \partial\Omega \times (0, +\infty), \\ w(x, 0) = w(x) & \text{in } \Omega. \end{cases}$$

□

4.4 Proof of Theorem 3

Let u and v be the global solutions in $C_{loc}([0, +\infty]; L^2(\Omega)) \cap L_{loc}^2([0, +\infty); H_0^1(\Omega))$ of, respectively, (1) and (33). Taking $u-v$ in both the problems (1) and (33) and subtracting the results we obtain for every $0 \leq t_1 < t_2$

$$\begin{aligned} & \frac{1}{2} \int_{\Omega} [u(x, t_2) - v(x, t_2)]^2 dx - \frac{1}{2} \int_{\Omega} [u(x, t_1) - v(x, t_1)]^2 + \\ & \int_{t_1}^{t_2} \int_{\Omega} [a(x, t, \nabla u) - b(x, t, \nabla v)] \nabla(u - v) \leq \\ & \lambda \int_{t_1}^{t_2} \int_{\Omega} \frac{(u - v)^2}{|x|^2} + \int_{t_1}^{t_2} \int_{\Omega} (f - F)(u - v), \end{aligned}$$

which is equivalent to the following estimate

$$\begin{aligned} & \frac{1}{2} \int_{\Omega} [u(x, t_2) - v(x, t_2)]^2 dx - \frac{1}{2} \int_{\Omega} [u(x, t_1) - v(x, t_1)]^2 + \\ & \int_{t_1}^{t_2} \int_{\Omega} [a(x, t, \nabla u) - a(x, t, \nabla v)] \nabla(u - v) \leq \lambda \int_{t_1}^{t_2} \int_{\Omega} \frac{(u - v)^2}{|x|^2} + \\ & \int_{t_1}^{t_2} \int_{\Omega} (f - F)(u - v) + \int_{t_1}^{t_2} \int_{\Omega} [b(x, t, \nabla v) - a(x, t, \nabla v)] \nabla(u - v). \quad (61) \end{aligned}$$

By assumption (4), Hardy’s inequality (7) and (61) we deduce

$$\begin{aligned} & \frac{1}{2} \int_{\Omega} [u(x, t_2) - v(x, t_2)]^2 dx - \frac{1}{2} \int_{\Omega} [u(x, t_1) - v(x, t_1)]^2 dx \\ & \quad \left(\alpha - \frac{\lambda}{\Lambda_N} \right) \int_{t_1}^{t_2} \int_{\Omega} |\nabla(u - v)|^2 \leq \tag{62} \\ & \int_{t_1}^{t_2} \int_{\Omega} (f - F)(u - v) + \int_{t_1}^{t_2} \int_{\Omega} [b(x, t, \nabla v) - a(x, t, \nabla v)] \nabla(u - v). \end{aligned}$$

We estimate the last two integrals in (62). Let $\theta \in (0, 1)$ a constant that we will choose below. It results (using Young’s and Poincaré’s inequalities)

$$\begin{aligned} \int_{t_1}^{t_2} \int_{\Omega} (f - F)(u - v) & \leq \frac{\theta}{2} C_0 c_P \int_{t_1}^{t_2} \int_{\Omega} (u - v)^2 + \frac{1}{2\theta C_0 c_P} \int_{t_1}^{t_2} \int_{\Omega} |f - F|^2 \leq \\ \frac{\theta}{2} C_0 \int_{t_1}^{t_2} \int_{\Omega} |\nabla(u - v)|^2 & + \frac{1}{2\theta C_0 c_P} \int_{t_1}^{t_2} \int_{\Omega} |f - F|^2 \end{aligned}$$

where c_P is Poincaré’s constant defined in (27) and $C_0 = \left(\alpha - \frac{\lambda}{\Lambda_N} \right)$. Moreover, we have

$$\begin{aligned} \int_{t_1}^{t_2} \int_{\Omega} [b(x, t, \nabla v) - a(x, t, \nabla v)] \nabla(u - v) & \leq \frac{\theta}{2} C_0 \int_{t_1}^{t_2} \int_{\Omega} |\nabla(u - v)|^2 + \\ \frac{1}{2\theta C_0} \int_{t_1}^{t_2} \int_{\Omega} |b(x, t, \nabla v) - a(x, t, \nabla v)|^2 \end{aligned}$$

By the previous estimates we deduce that

$$\begin{aligned} & \int_{\Omega} [u(x, t_2) - v(x, t_2)]^2 dx - \int_{\Omega} [u(x, t_1) - v(x, t_1)]^2 dx + \\ & \quad 2(1 - \theta) C_0 \int_{t_1}^{t_2} \int_{\Omega} |\nabla(u - v)|^2 \leq \\ & \quad \frac{1}{\theta C_0 c_P} \int_{t_1}^{t_2} \int_{\Omega} |f - F|^2 + \frac{1}{\theta C_0} \int_{t_1}^{t_2} \int_{\Omega} |b(x, t, \nabla v) - a(x, t, \nabla v)|^2. \end{aligned}$$

which implies (again by Poincaré’s inequality)

$$\begin{aligned} & \int_{\Omega} [u(x, t_2) - v(x, t_2)]^2 dx - \int_{\Omega} [u(x, t_1) - v(x, t_1)]^2 dx + \\ & \quad M \int_{t_1}^{t_2} \int_{\Omega} |u - v|^2 \leq \int_{t_1}^{t_2} g(s) ds \tag{63} \end{aligned}$$

where $M = 2c_P(1 - \theta)C_0 = 2(1 - \theta)\sigma$ (where σ is as in (26)) and

$$g(s) = \frac{1}{\theta C_0} \int_{\Omega} \left[\frac{1}{c_P} |f(x, s) - F(x, s)|^2 + |b(x, s, \nabla v) - a(x, s, \nabla v)|^2 \right] dx. \tag{64}$$

Denoting $\sigma_0 = (1 - \theta)\sigma$ (i.e., $\theta = 1 - \frac{\sigma_0}{\sigma}$) and applying Proposition 1 with $\phi(t) = \int_{\Omega} [u(x, t) - v(x, t)]^2 dx$ and $t_0 = 0$, the assertions follow. \square

4.5 Proof of Corollary 2

The asserts follow observing that $w(x, t) = w(x)$ is also a global solution in

$$C_{loc}([0, +\infty]; L^2(\Omega)) \cap L^2_{loc}([0, +\infty); H^1_0(\Omega))$$

of the following evolution problem

$$\begin{cases} w_t - \operatorname{div}(b(x, \nabla w)) = \lambda \frac{w}{|x|^2} + F(x) & \text{in } \Omega \times (0, +\infty), \\ w(x, t) = 0 & \text{on } \partial\Omega \times (0, +\infty), \\ w(x, 0) = w(x) & \text{in } \Omega. \end{cases}$$

\square

References

1. Abdellaoui, B., Peral, I.: Existence and non-existence results for quasilinear elliptic problems involving the p-Laplacian. *Ann. Mat. Pura Appl.* **182**(3), 247–270 (2003)
2. Abdellaoui, B., Peral, I.: Nonexistence results for quasilinear elliptic equations related to Caffarelli-Kohn-Nirenberg inequality. *Commun. Pure Appl. Anal.* **2**(4), 539–566 (2003)
3. Abdellaoui, B., Colorado, E., Peral, I.: Existence and nonexistence results for a class of linear and semilinear parabolic equations related to some Caffarelli–Kohn–Nirenberg inequalities. *J. Eur. Math. Soc.* **6**, 119–148 (2004)
4. Abdellaoui, B., Peral, I., Primo, A.: Elliptic problems with a Hardy potential and critical growth in the gradient: non-resonance and blow-up results. *J. Differ. Equ.* **239**(2), 386–416 (2007)
5. Abdellaoui, B., Peral, I., Primo, A.: Breaking of resonance and regularizing effect of a first order quasi-linear term in some elliptic equations. *Annales de l’Institut Henri Poincaré* **25**, 969–985 (2008)
6. Aguilar Crespo, J.A., Peral, I.: Global behavior of the Cauchy problem for some critical nonlinear parabolic equations. *SIAM J. Math. Anal.* **31**, 1270–1294 (2000)
7. Arcoya, D., Molino, A., Moreno, L.: Existence and regularizing effect of degenerate lower order terms in elliptic equations beyond the Hardy constant (2018, preprint)
8. Aronson, D.G., Peletier, L.A.: Large time behavior of solutions of the porous medium equation in bounded domains. *J. Differ. Equ.* **39**, 378–412 (1981)

9. Aronson, D.G., Serrin, J.: Local behavior of solutions of quasilinear parabolic equations. *Arch. Ration. Mech. Anal.* **25**, 81–122 (1967)
10. Baras, P., Goldstein, J.: The heat equation with singular potential. *Trans. Am. Math. Soc.* **294**, 121–139 (1984)
11. Bebernes, J., Eberly, D.: *Mathematical Problems from Combustion Theory*. Applied Mathematical Sciences, vol. 83. Springer, New York (1989)
12. Benilan, P., Crandall, M.G., Pierre, M.: Solutions of the porous medium in \mathbb{R}^N under optimal conditions on initial values. *Indiana Univ. Math. J.* **33**, 51–87 (1984)
13. Boccardo, L., Porzio, M.M.: Degenerate parabolic equations: existence and decay properties. *Discrete Contin. Dyn. Syst. Ser. S* **7**(4), 617–629 (2014)
14. Boccardo, L., Dall’Aglío, A., Gallouët, T., Orsina, L.: Existence and regularity results for nonlinear parabolic equations. *Adv. Math. Sci. Appl.* **9**, 1017–1031 (1999)
15. Boccardo, L., Orsina, L., Peral, I.: A remark on existence and optimal summability of solutions of elliptic problems involving Hardy potential. *Discrete Contin. Dyn. Syst.* **16**(3), 513 (2006)
16. Boccardo, L., Porzio, M.M., Primo, A.: Summability and existence results for nonlinear parabolic equations. *Nonlinear Anal. TMA* **71**, 978–990 (2009)
17. Bonforte, M., Grillo, G.: Super and ultracontractive bounds for doubly nonlinear evolution equations. *Rev. Mat. Iberoamericana* **22**(1), 11–129 (2006)
18. Brezis, H., Crandall, M.G.: Uniqueness of solutions of the initial-value problems for $u_t - \Delta\varphi(u) = 0$. *J. Math. Pures Appl.* **58**, 153–163 (1979)
19. Cabré, X., Martel, Y.: Existence versus explosion instantanée pour des équations de la chaleur linéaires avec potentiel singulier. *C. R. Acad. Sci. Paris* **329**(11), 973–978 (1999)
20. Caffarelli, L., Kohn, R., Nirenberg, L.: First order interpolation inequalities with weights. *Compos. Math.* **53**, 259–275 (1984)
21. Cipriani, F., Grillo, G.: Uniform bounds for solutions to quasilinear parabolic equations. *J. Differ. Equ.* **177**, 209–234 (2001)
22. Di Benedetto, E., Herrero, M.A.: On the Cauchy problem and initial traces for a degenerate parabolic equation. *Trans. Am. Math. Soc.* **314**, 187–224 (1989)
23. Di Benedetto, E., Herrero, M.A.: Non negative solutions of the evolution p-Laplacian equation. Initial traces and Cauchy problem when $1 < p < 2$. *Arch. Ration. Mech. Anal.* **111**(3), 225–290 (1990)
24. Garcia Azorero, J., Peral, I.: Hardy inequalities and some critical elliptic and parabolic problems. *J. Differ. Equ.* **144**, 441–476 (1998)
25. Goldstein, J., Zhang, Q.S.: On a degenerate heat equation with a singular potential. *J. Funct. Anal.* **186**(2), 342–359 (2001)
26. Goldstein, J., Zhang, Q.S.: Linear parabolic equations with strong singular potentials. *Trans. Am. Math. Soc.* **355**(1), 197–211 (2003)
27. Grenon, N., Mercaldo, A.: Existence and regularity results for solutions to nonlinear parabolic equations. *Adv. Differ. Equ.* **10**(9), 1007–1034 (2005)
28. Grillo, G., Muratori, M., Porzio, M.M.: Porous media equations with two weights: existence, uniqueness, smoothing and decay properties of energy solutions via Poincaré inequalities. *Discrete Contin. Dyn. Syst. A* **33**(8), 3599–3640 (2013)
29. Herrero, M.A., Pierre, M.: The Cauchy problem for $u_t = \Delta u^m$ when $0 < m < 1$. *Trans. Am. Math. Soc.* **291**, 145–158 (1985)
30. Herrero, M.A., Vazquez, J.L.: Asymptotic behavior of the solutions of a strongly nonlinear parabolic problem. *Ann. Fac. Sci., Toulouse Math.* (5) **3**(2), 113–127 (1981)
31. Ladyženskaja, O., Solonnikov, V.A., Ural’ceva, N.N.: *Linear and Quasilinear Equations of Parabolic Type*. Translations of the American Mathematical Society. American Mathematical Society, Providence (1968)
32. Moscariello, G., Porzio, M.M.: Quantitative asymptotic estimates for evolution problems. *Nonlinear Anal. Theory Methods Appl.* **154**, 225–240 (2017)
33. Porzio, M.M.: Existence of solutions for some “noncoercive” parabolic equations. *Discrete Contin. Dyn. Syst.* **5**(3), 553–568 (1999)

34. Porzio, M.M.: Local regularity results for some parabolic equations. *Houst. J. Math.* **25**(4), 769–792 (1999)
35. Porzio, M.M.: On some quasilinear elliptic equations involving Hardy potential. *Rendiconti di matematica e delle sue applicazioni, Serie VII*, **27**, fasc. III–IV, 277–299 (2007)
36. Porzio, M.M.: On decay estimates. *J. Evol. Equ.* **9**(3), 561–591 (2009)
37. Porzio, M.M.: Existence, uniqueness and behavior of solutions for a class of nonlinear parabolic problems. *Nonlinear Anal. Theory Methods Appl.* **74**, 5359–5382 (2011)
38. Porzio, M.M.: On uniform and decay estimates for unbounded solutions of partial differential equations. *J. Differ. Equ.* **259**, 6960–7011 (2015)
39. Porzio, M.M.: Regularity and time behavior of the solutions of linear and quasilinear parabolic equations. *Adv. Differ. Equ.* **23**(5–6), 329–372 (2018)
40. Porzio, M.M.: A new approach to decay estimates. Application to a nonlinear and degenerate parabolic PDE. *Rend. Lincei Mat. Appl.* **29**, 635–659 (2018)
41. Porzio, M.M.: On the lack of summability properties of solutions of elliptic problems with singular potentials. Preprint
42. Porzio, M.M., Pozio, M.A.: Parabolic equations with non-linear, degenerate and space-time dependent operators. *J. Evol. Equ.* **8**, 31–70 (2008)
43. Porzio, M.M., Primo, A.: Summability and existence results for quasilinear parabolic equations with Hardy potential term. *Nonlinear Differ. Equ. Appl.* **20**(1), 65–100 (2013)
44. Reed, M., Simon, B.: *Methods of Modern Mathematical Physics, vol. II.* Academic Press, New York (1979)
45. Vancostenoble, J., Zuazua, E.: Null controllability for the heat equation with singular inverse-square potentials. *J. Funct. Anal.* **254**, 1864–1902 (2008)
46. Vazquez, J.L.: *Smoothing and decay estimates for nonlinear diffusion equations.* Oxford University Press, Oxford (2006)
47. Vazquez, J.L., Zuazua, E.: The Hardy inequality and the asymptotic behavior of the heat equation with an inverse-square potential. *J. Funct. Anal.* **173**, 103–153 (2000)
48. Veron, L.: Effects regularisants des semi-groupes non linéaires dans des espaces de Banach. *Ann. Fac. Sci. Toulouse Math. (5)* **1**(2), 171–200 (1979)
49. Wei, L., Du, Y.: Exact singular behavior of positive solutions to nonlinear elliptic equations with a Hardy potential. *J. Differ. Equ.* **262**, 3864–3886 (2017)
50. Wei, L., Feng, Z.: Isolated singularity for semilinear elliptic equations. *Discrete Contin. Dyn. Syst.* **35**, 3239–3252 (2015)

How to Hear the Corners of a Drum



Medet Nursultanov, Julie Rowlett, and David Sher

Abstract We prove that the existence of corners in a class of planar domain, which includes all simply connected polygonal domains and all smoothly bounded domains, is a spectral invariant of the Laplacian with both Neumann and Robin boundary conditions. The main ingredient in the proof is a locality principle in the spirit of Kac’s “principle of not feeling the boundary,” but which holds uniformly up to the boundary. Albeit previously known for Dirichlet boundary condition, this appears to be new for Robin and Neumann boundary conditions, in the geometric generality presented here. For the case of curvilinear polygons, we describe how the same arguments using the locality principle are insufficient to obtain the analogous result. However, we describe how one may be able to harness powerful microlocal methods and combine these with the locality principles demonstrated here to show that corners are a spectral invariant; this is current work-in-progress (Nursultanov et al., Preprint).

1 Introduction

It is well known that “one cannot hear the shape of a drum” [8, 22, 27]. Mathematically, this means that there exist bounded planar domains which have the same eigenvalues for the Laplacian with Dirichlet boundary condition, in spite of the

M. Nursultanov · J. Rowlett (✉)
Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden

University of Gothenburg, Gothenburg, Sweden
e-mail: medet@chalmers.se; julie.rowlett@chalmers.se

D. Sher
Department of Mathematical Sciences, DePaul University, Chicago, IL, USA
e-mail: dsher@depaul.edu

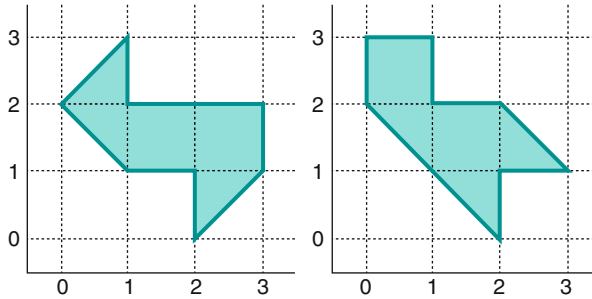


Fig. 1 These two domains were demonstrated by Gordon, Webb, and Wolpert to be isospectral for the Laplacian with Dirichlet boundary condition [9]. This image is from Wikipedia Commons

domains having different shapes. The standard example is shown in Fig. 1. Two geometric characteristics of these domains are immediately apparent:

1. These domains both have corners.
2. Neither of these domains are convex.

This naturally leads to the following two open problems:

Problem 1 Can one hear the shape of a smoothly bounded drum?

Problem 2 Can one hear the shape of a convex drum?

The mathematical formulation of these problems are: if two smoothly bounded (respectively, convex) domains in the plane are isospectral for the Laplacian with Dirichlet boundary condition, then are they the same shape?

One could dare to conjecture that the answer to Problem 1 is yes, based on the isospectrality result of Zelditch [36]. He proved that if two analytically bounded domains both have a bilateral symmetry and are isospectral, then they are in fact the same shape. For certain classes of convex polygonal domains including triangles [5, 11]; parallelograms [15]; and trapezoids [12]; if two such domains are isospectral, then they are indeed the same shape. This could lead one to suppose that perhaps Problem 2 also has a positive answer.

Contemplating these questions led the second author and Z. Lu to investigate whether smoothly bounded domains can be isospectral to domains with corners. In [16], they proved that for the Dirichlet boundary condition, “one can hear the corners of a drum” in the sense that a domain with corners cannot be isospectral to a smoothly bounded domain. Here we generalize that result to both Neumann and Robin boundary conditions.

The key technical tool in the proof is a locality principle for the Neumann and Robin boundary conditions in a general context which includes domains with only piecewise smooth boundary. This locality principle may be of independent interest, because it not only generalizes Kac’s “principle of not feeling the boundary” [13] but also unlike that principle, it holds uniformly up to the boundary. First, we explain Kac’s locality principle. Let Ω be a bounded domain in \mathbb{R}^2 , or more generally

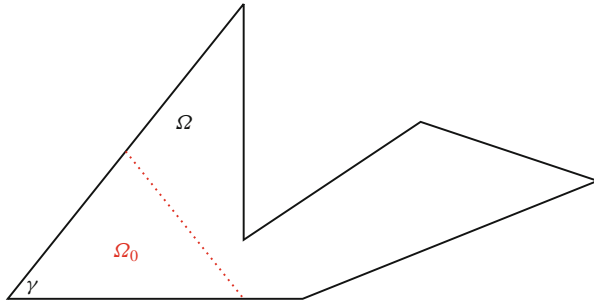


Fig. 2 Above, we have the polygonal domain Ω which contains the triangular domain, Ω_0 . Letting $S = S_\gamma$ be a circular sector of opening angle γ and infinite radius, this is an example of an “exact geometric match,” in the sense that Ω_0 is equal to a piece of S

\mathbb{R}^n , because the argument works in the same way in all dimensions. Assume the Dirichlet boundary condition, and let the corresponding heat kernel for Ω be denoted by H , while the heat kernel for \mathbb{R}^n ,

$$K(t, z, z') = (4\pi t)^{-n/2} e^{-d(z,z')^2/4t}. \tag{1}$$

Let

$$\delta = \min\{d(z, \partial\Omega), d(z', \partial\Omega)\}.$$

Then, there are constants $A, B > 0$ such that

$$|K(t, z, z') - H(t, z, z')| \leq At^{-n/2} e^{-B\delta^2/t}.$$

This means that the heat kernel for Ω is $O(t^\infty)$ ¹ close to the Euclidean heat kernel, as long as we consider points z, z' which are at a positive distance from the boundary. Hence the heat kernel “does not feel the boundary.”

In a similar spirit, a more general locality principle is known to be true. The idea is that one has a collection of sets which are “exact geometric matches” to certain pieces of the domain, Ω . To describe the meaning of an “exact geometric match,” consider a piece of the first quadrant near the origin in \mathbb{R}^2 . A sufficiently small piece is an exact match for a piece of a rectangle near a corner. Similarly, for a surface with exact conical singularities, near a singularity of opening angle γ , a piece of an infinite cone with the same opening angle is an exact geometric match to a piece of the surface near that singularity. For a planar example, see Fig. 2. The locality principle states that if one takes the heat kernels for those “exact geometric matches,” and restricts them to the corresponding pieces of the domain (or manifold), Ω , then those “model heat kernels” are equal to the heat kernel for Ω , restricted to the corresponding pieces of Ω , with error $O(t^\infty)$ as $t \downarrow 0$.

¹By $O(t^\infty)$, we mean $O(t^N)$ for any $N \in \mathbb{N}$.

This locality principle is incredibly useful, because if one has exact geometric matches for which one can explicitly compute the heat kernel, then one can use these to compute the short time asymptotic expansion of the heat trace. Moreover, in addition to being able to compute the heat trace expansion, one can also use this locality principle to compute the zeta regularized determinant of the Laplacian as in [1].

Here, we shall give one application of the locality principle: “how to hear the corners of a drum.”

Theorem 1 *Let $\Omega \subset \mathbb{R}^2$ be a simply connected, bounded, Lipschitz planar domain with piecewise smooth boundary. Moreover, assume that the (finitely many) points at which the boundary is not smooth are exact corners; that is, there exists a neighborhood of each corner in which the boundary of Ω is the union of two straight line segments. Assume that for at least one such corner, the interior angle is not equal to π .*

Then the Laplacian with either Dirichlet,² Neumann, or Robin boundary condition is not isospectral to the Laplacian with the same boundary condition³ on any smoothly bounded domain.

To prove the result, we use a locality principle which is stated and proved in Sect. 2. We next introduce model heat kernels as well as the corresponding Green’s functions for the “exact geometric matches” in Sect. 3. We proceed there to use the models together with our locality principle to compute the short time asymptotic expansion of the heat trace. Theorem 1 is then a consequence of comparing the heat trace expansions in the presence and lack of corners. In conclusion, we explain in how the locality principle *fails* to prove Theorem 1 for the case of curvilinear polygonal domains, in which the corners are not exact. An example of a non-exact corner of interior angle $\pi/2$ is the corner where the straight edge meets the curved edge in a half-circle. This motivates the discussion in Sect. 4 concerning the necessity and utility of microlocal analysis, in particular, the construction of the heat kernel for curvilinear polygonal domains and surfaces via the heat space and heat calculus in those settings. This construction, together with a generalization of Theorem 1 to all boundary conditions (including discontinuous, mixed boundary conditions), as well as to surfaces with both conical singularities and edges, is currently in preparation and shall be presented in forthcoming work [23].

²This result was proven in the Dirichlet case in [16].

³In particular, in the case of Robin boundary conditions, we assume the same Robin parameters for both domains.

2 The Locality Principle

We begin by setting notations and sign conventions and recalling fundamental concepts.

2.1 Geometric and Analytic Preliminaries

To state the locality principle, we make the notion of an “exact geometric match” precise. Let Ω be a domain, possibly infinite, contained in \mathbb{R}^n .

Definition 1 Assume that $\Omega_0 \subset \Omega \subset \mathbb{R}^n$, and $S \subset \mathbb{R}^n$. We say that S and Ω are *exact geometric matches on Ω_0* if there exists a sub-domain $\Omega_c \subseteq \Omega$ which compactly contains Ω_0 and which is isometric to a sub-domain of S (which, abusing notation, we also call Ω_c). Recall that Ω_0 being compactly contained in Ω_c means that the distance from $\overline{\Omega_0}$ to $\overline{\Omega} \setminus \overline{\Omega_c}$ is positive. A planar example is depicted in Fig. 2.

Next, we recall the heat kernel in this context. The heat kernel, H , is the Schwartz kernel of the fundamental solution of the heat equation. It is therefore defined on $\Omega \times \Omega \times [0, \infty)$, and satisfies

$$H(t, z, z') = H(t, z', z), \quad (\partial_t + \Delta)H(t, z, z') = 0 \text{ for } t > 0,$$

$$H(0, z, z') = \delta(z - z'), \quad \text{in the distributional sense.}$$

Throughout we use the sign convention for the Laplacian, Δ , on \mathbb{R}^n , that

$$\Delta = - \sum_{j=1}^n \partial_j^2.$$

We consider two boundary conditions:

- (N) the *Neumann boundary condition*, which requires the normal derivative of the function to vanish on the boundary;
- (R) the *Robin boundary condition*, which requires the function, u , to satisfy the following equation on the boundary:

$$\alpha u + \beta \frac{\partial u}{\partial \nu} = 0, \quad \frac{\partial u}{\partial \nu} \text{ is the outward pointing normal derivative.} \quad (2)$$

For $u_0 \in \mathcal{L}^2(\Omega)$, the heat equation with initial data given by u_0 is then solved by

$$u(t, z) = \int_{\Omega} H(t, z, z') u_0(z') dz'.$$

Moreover, if Ω is a bounded domain, and $\{\phi_k\}_{k \geq 1}$ is an orthonormal basis for $\mathcal{L}^2(\Omega)$ consisting of eigenfunctions of the Laplacian satisfying the appropriate boundary condition, with corresponding eigenvalues $\{\lambda_k\}_{k \geq 1}$, then the heat kernel

$$H(t, z, z') = \sum_{k \geq 1} e^{-\lambda_k t} \phi_k(z) \phi_k(z').$$

2.2 Locality Principle for Dirichlet Boundary Condition

In the general context of domains in \mathbb{R}^n which have only piecewise smooth boundary, the key point is that the locality principle should hold *up to the boundary*. This differs from many previous presentations of a locality principle. For example, in [14, Theorem 1.1], it is proved that without any condition on the regularity of the boundary, for any choice of self-adjoint extension of the Laplacian on $\Omega \subset \mathbb{R}^n$, the heat kernel for this self adjoint extension of the Laplacian on Ω , denoted by H^Ω satisfies

$$|H^\Omega(t, z, z') - H^0(t, z, z')| \leq (C_a \rho(z, z')^{-n} + C_b) \cdot \frac{\exp\left(-\frac{(\rho(z) + \rho(z'))^2}{4t}\right)}{t^{2\lceil \frac{n+1}{2} \rceil - \frac{1}{2}}}.$$

Above, H^0 is the heat kernel for \mathbb{R}^n , $\rho(z) = \text{dist}(z, \partial\Omega)$, $\rho(z, z') = \min(\rho(z), \rho(z'))$. The constants C_a and C_b can also be calculated explicitly according to [14]. Clearly, the estimate loses its utility as one approaches the boundary.

In the case of smoothly bounded domains, there is a result of Lück and Schick [17, Theorem 2.26], which implies the locality principle for both the Dirichlet and Neumann boundary conditions, and which holds all the way up to the boundary. We recall that result.⁴

Theorem 2 (Lück and Schick) *Let N be a Riemannian manifold possibly with boundary which is of bounded geometry. Let $V \subset N$ be a closed subset which carries the structure of a Riemannian manifold of the same dimension as N such that the inclusion of V into N is a smooth map respecting the Riemannian metrics. For fixed $p \geq 0$, let $\Delta[V]$ and $\Delta[N]$ be the Laplacians on p -forms on V and N , considered as unbounded operators with either absolute boundary conditions or with relative boundary conditions (see Definition 2.2 of [17]). Let*

⁴In the original statement of their result, Lück and Schick make the parenthetical remark “We make no assumptions about the boundaries of N and V and how they intersect.” This could easily be misunderstood. If one carefully reads the proof, it is implicit that the boundaries are *smooth*. The arguments break down if the boundaries have singularities, such as corners. For this reason, we have omitted the parenthetical remark from the statement of the theorem.

$\Delta[V]^k e^{-t\Delta[V]}(x, y)$ and $\Delta[N]^k e^{-t\Delta[N]}(x, y)$ be the corresponding smooth integral kernels. Let k be a non-negative integer.

Then there is a monotone decreasing function $C_k(K) : (0, \infty) \rightarrow (0, \infty)$ which depends only on the geometry of N (but not on V, x, y, t) and a constant C_2 depending only on the dimension of N such that for all $K > 0$ and $x, y \in V$ with $d_V(x) := d(x, N \setminus V) \geq K, d_V(y) \geq K$ and all $t > 0$:

$$\left| \Delta[V]^k e^{-t\Delta[V]}(x, y) - \Delta[N]^k e^{-t\Delta[N]}(x, y) \right| \leq C_k(K) e^{-\left(\frac{d_V(x)^2 + d_V(y)^2 + d(x, y)^2}{C_2 t}\right)}.$$

One may therefore compare the heat kernels for the Laplacian acting on functions, noting (see p. 362 of [28]) that relative boundary conditions are Dirichlet boundary conditions, and absolute boundary conditions are Neumann boundary conditions. We present this as a corollary to Lück and Schick’s theorem.

Corollary 1 Assume that S is an exact match for $\Omega_0 \subset \Omega$, for two smoothly bounded domains, Ω and Ω_0 in \mathbb{R}^n . Assume the same boundary condition, either Dirichlet or Neumann, for the Euclidean Laplacian on both domains. Then

$$\left| H^\Omega(t, z, z') - H^S(t, z, z') \right| = O(t^\infty) \text{ as } t \downarrow 0, \quad \text{uniformly for } z, z' \in \Omega_0.$$

Proof We use the theorem of Lück and Schick twice, once with $N = \Omega$ and once with $N = S$, with $V = \Omega_c$ in both cases. We set $k = 0$ and

$$K = \alpha = d(\Omega_0, S \setminus \Omega_c).$$

By the definition of an exact geometric match, $\alpha > 0$. In the $N = S$ case, the theorem reads

$$\left| H^S(t, z, z') - H^{\Omega_c}(t, z, z') \right| \leq C_0(\alpha) e^{-\frac{|\text{dist}(z, S \setminus \Omega_c)|^2}{C_2 t} - \frac{|\text{dist}(z', S \setminus \Omega_c)|^2}{C_2 t}} \leq C_0(\alpha) e^{-\frac{2\alpha^2}{C_2 t}}.$$

We conclude that

$$\left| H^S(t, z, z') - H^{\Omega_c}(t, z, z') \right| = O(t^\infty)$$

uniformly on Ω_0 . The same statement holds with S replaced by Ω , and then the triangle inequality completes the proof. \square

The assumption of smooth boundary is quite restrictive, and the proof in [17] relies heavily on this assumption. To the best of our knowledge, the first locality result which holds all the way up to the boundary and includes domains which have only piecewise smooth boundary, but may have corners, was demonstrated by van den Berg and Srisatkunarahaj [29]. We note that this result is not stated in the precise form below in [29], but upon careful reading, it is straightforward to verify that this result is indeed proven in [29] and is used in several calculations therein.

Theorem 3 (van den Berg and Srisatkunarajah) *Let $\Omega \subset \mathbb{R}^2$ be a polygonal domain. Let H^Ω denote the heat kernel for the Laplacian on Ω with the Dirichlet boundary condition. Then, for $S = S_\gamma$, a sector of opening angle γ , and for any corner of Ω with opening angle γ , there is a neighborhood of the corner \mathcal{N}_γ such that*

$$|H^\Omega(t, z, z') - H^{S_\gamma}(t, z, z')| = O(t^\infty), \quad \text{uniformly } \forall(z, z') \in \mathcal{N}_\gamma \times \mathcal{N}_\gamma,$$

Above, H^{S_γ} denotes the heat kernel for S_γ with the Dirichlet boundary condition. Moreover, for any $\mathcal{N}_e \subset \Omega$ which is at a positive distance to all corners of Ω ,

$$|H^\Omega(t, z, z') - H^{\mathbb{R}_+^2}(t, z, z')| = O(t^\infty), \quad \text{uniformly } \forall(z, z') \in \mathcal{N}_e \times \mathcal{N}_e.$$

Above, $H^{\mathbb{R}_+^2}$ denotes the heat kernel for a half space with the Dirichlet boundary condition.

The proof uses probabilistic methods. We are currently unaware of a generalization to domains with corners in higher dimensions. However, it is reasonable to expect that such a generalization holds. Since Theorem 1 has already been demonstrated for the Dirichlet boundary condition in [16], we are interested in the Neumann and Robin boundary conditions. For this reason, we shall give a proof of a locality principle for both Neumann and Robin boundary conditions which holds in all dimensions, for domains with piecewise smooth boundary (in fact, only piecewise \mathcal{C}^3 boundary is required), as long as we have a suitable estimate on the second fundamental form on the boundary. Moreover, our locality principle, similar to that of [29], allows one to compare the heat kernels all the way up to the boundary. For this reason, the locality principles demonstrated below may be of independent interest.

2.3 Locality Principle for Neumann Boundary Condition

Here we prove a locality principle for the Neumann boundary condition for domains in \mathbb{R}^n with piecewise \mathcal{C}^2 boundary satisfying some relatively general geometric assumptions. Since we consider both bounded and unbounded domains, we require a uniform version of an interior cone condition:

Definition 2 Let $\varepsilon > 0$ and $h > 0$. We say that a domain $\Omega \subset \mathbb{R}^n$ satisfies the (ε, h) -cone condition if, for every $x \in \partial\Omega$, there exists a ball $B(x, \delta)$ centered at x of radius δ , and a direction ξ_x , such that for all $y \in B(x, \delta) \cap \Omega$, the cone with vertex y directed by ξ_x of opening angle ε and height h is contained in Ω .

Definition 3 Let $\varepsilon > 0$ and $h > 0$. We say that a domain $\Omega \subset \mathbb{R}^n$ satisfies the two-sided (ε, h) -cone condition if both Ω and $\mathbb{R}^n \setminus \Omega$ satisfy the (ε, h) -cone condition.

Theorem 4 (Locality Principle for Neumann Boundary Condition) *Let Ω , Ω_0 , and S be domains in \mathbb{R}^n such that S and Ω are exact geometric matches on Ω_0 , as in Definition 1. Assume that both Ω and S satisfy the two-sided (ε, h) -cone condition for some $\varepsilon > 0$ and $h > 0$. Let H^Ω denote the heat kernel associated to the Laplacian on Ω , and let H^S denote the heat kernel on S , with the same boundary condition for ∂S as taken on $\partial\Omega$. Moreover, assume that there exists $\sigma \in \mathbb{R}$ such that the second fundamental form $\mathbb{I} \geq \sigma$ holds on all the \mathcal{C}^2 pieces of $\partial\Omega$ and ∂S . Then*

$$\left| H^\Omega(t, z, z') - H^S(t, z, z') \right| = O(t^\infty) \text{ as } t \downarrow 0, \quad \text{uniformly for } z, z' \in \Omega_0.$$

Proof We use a patchwork parametrix construction, as discussed in section 3.2 of [1]. This is a general technique to construct heat kernels whenever one has exact geometric matches for each part of a domain; see for example [1], [18] and [26].

Let $\{\chi_j\}_{j=1}^2$ be a \mathcal{C}^∞ partition of unity on Ω . Assume that $\tilde{\chi}_j \in \mathcal{C}^\infty(\Omega)$ is identically 1 on a small neighborhood of the support of χ_j and vanishes outside a slightly larger neighborhood. In particular, we choose χ_1 to be identically equal to one on Ω_0 . Choose $\tilde{\chi}_1$ to be identically one on a strictly larger neighborhood and to have its support equal to Ω_c . We assume that the support of $\tilde{\chi}_2$ does not intersect Ω_0 . We then define the patchwork heat kernel

$$G(t, z, z') := \sum_{j=1}^2 \tilde{\chi}_j(z) H^S(t, z, z') \chi_j(z').$$

We claim that uniformly for all $z, z' \in \Omega_0$,

$$|H^\Omega(t, z, z') - G(t, z, z')| = O(t^\infty), \quad t \downarrow 0.$$

That is, we claim that the patchwork heat kernel is equal to the true heat kernel with an error that is $O(t^\infty)$ for small time. This claim immediately implies our result, since on Ω_0 , $\chi_1 = 1$, and $\tilde{\chi}_1 = 1$, whereas χ_2 and $\tilde{\chi}_2$ both vanish, and thus $G(t, z, z') = H^S(t, z, z')$.

To prove the claim, we follow the usual template. Observe that

$$E(t, z, z') := (\partial_t + \Delta)G(t, z, z') = \sum_{j=1}^2 [\Delta, \tilde{\chi}_j(z)] H^S(t, z, z') \chi_j(z').$$

Each commutator $[\Delta, \tilde{\chi}_j(z)]$ is a first-order differential operator with support a positive distance from the support of χ_j . Thus $E(t, z, z')$ is a sum of model heat kernels and their first derivatives, cut off so that their spatial arguments are a positive distance from the diagonal. We claim each such term is $O(t^\infty)$. To obtain this

estimate, we use [30, Theorem 1.1], which gives the estimate

$$|\nabla H^D(t, z, z')| \leq \frac{C_\alpha}{t^{(n+1)/2}} \exp\left(-\frac{|z - z'|}{C_\beta t}\right), \quad z, z' \in D,$$

for some constants $C_\alpha, C_\beta > 0$, for $D = \Omega$ and $D = S$. The setting there is not identical, so we note the places in the proof where minor modifications are required. First, the assumption that Ω is compact is used there to obtain estimates for all $t > 0$. In particular, the discreteness of the spectrum is used to obtain long time estimates by exploiting the first positive eigenvalue in (2.1) of [30]. Since we are only interested in $t \downarrow 0$, this long time estimate is not required. Next, compactness is used to be able to estimate the volume of balls, $|B(x, \sqrt{t})| \geq C_\varepsilon t^{\frac{n}{2}}$, for a uniform constant C_ε . However, we have this estimate due to the two-sided (ε, h) -cone condition which is satisfied for both Ω and S which are contained in \mathbb{R}^n . Moreover, we have verified (Wang, private communication) that the assumption of piecewise \mathcal{C}^2 boundary (rather than \mathcal{C}^2 boundary) is sufficient for the proof of [30, Theorem 1.1], as well as the references used therein: [31–33].⁵

Since the domains S and Ω satisfy the two-sided (ε, h) -cone condition, there are Gaussian upper bounds for the corresponding Neumann heat kernels. Specifically, as a result of [4, Theorems 6.1, 4.4], for any $T > 0$, there exist $C_1, C_2 > 0$ such that

$$|H^S(t, z, z')| \leq C_1 t^{-\frac{n}{2}} e^{-\frac{|z-z'|^2}{C_2 t}}, \quad |H^\Omega(t, z, z')| \leq C_1 t^{-\frac{n}{2}} e^{-\frac{|z-z'|^2}{C_2 t}} \tag{3}$$

on $(0, T] \times S \times S$ and $(0, T] \times \Omega \times \Omega$ respectively. The upshot is that each term in the sum defining $E(t, z, z')$ is uniformly $O(t^\infty)$ for all z and z' in Ω , and therefore

$$|E(t, z, z')| = O(t^\infty).$$

From here, the error may be iterated away using the usual Neumann series argument, as in [19] or Section 4 of [26]. Letting $*$ denote the operation of convolution in time and composition in space, define

$$K := E - E * E + E * E * E - \dots$$

It is an exercise in induction to see that $K(t, z, z')$ is well-defined and also $O(t^\infty)$ as t goes to zero, see for example the proof of parts a) and b) of Lemma 13 of [26]. Note that Ω is compact, which is key. Then the difference of the true heat kernel and the patchwork heat kernel is

$$H^\Omega(t, z, z') - G(t, z, z') = -(G * K)(t, z, z').$$

⁵We have also verified in private communication with F. Y. Wang that the arguments in [30–33] apply equally well under the curvature assumption $\mathbb{I} \geq -\sigma$ for piecewise \mathcal{C}^2 boundary.

As in Lemma 14 of [26], this can also be bounded in straightforward fashion by $O(t^\infty)$, which completes the proof. \square

The key ingredients in this patchwork construction are: (1) the model heat kernels satisfy off-diagonal decay estimates, and (2) the gradients of these model heat kernels satisfy similar estimates. The argument can therefore be replicated in any situation where all models satisfy those estimates. Here is one generalization:

Corollary 2 *Using the notation of Theorem 4, suppose that Ω is compact and that the heat kernels on both Ω and S satisfy off-diagonal bounds of the following form: if A and B are any two sets with $d(A, B) > 0$, then uniformly for $z \in A$ and $z' \in B$, we have*

$$|H(t, z, z')| + |\nabla H(t, z, z')| = O(t^\infty) \text{ as } t \rightarrow 0. \tag{4}$$

Then the conclusion of Theorem 4 holds.

Proof Apply the same method, with a partition of unity on Ω consisting of just two components, one cutoff function for Ω_0 where we use the model heat kernel H^S , and one cutoff function for the rest of Ω where we use H^Ω . The result follows.

Remark 1 The bounds (4) are satisfied, for example, by Neumann heat kernels on compact, convex domains with no smoothness assumptions on the boundary [30], as well as by both Dirichlet and Neumann heat kernels on sectors, half-spaces, and Euclidean space.

2.4 Locality for Robin Boundary Condition

In this section, we determine when locality results similar to those of Theorem 4 hold for the Robin problem. The answer is that in many cases they may be deduced from locality of the Neumann heat kernels. We consider a generalization of the classical Robin boundary condition (2)

$$\frac{\partial}{\partial n}u(x) + c(x)u(x) = 0, \quad x \in \text{the smooth pieces of } \partial D. \tag{5}$$

In the first version of the locality principle, to simplify the proof, we shall assume that $\Omega \subset S$, and that Ω is bounded. We note, however, that both of these assumptions can be removed in the corollary to the theorem. The statement of the theorem may appear somewhat technical, so we explain the geometric interpretations of the assumptions. Conditions (1) and (2) below are clear; they are required to apply our references [30] and [4]. Items (3), (4), and (5) mean that the (possibly unbounded) domain, $D = S$ (and as in the corollary, in which Ω may be unbounded, $D = \Omega$) has boundary which does not oscillate too wildly or “bunch up” and become space-filling. These assumptions are immediately satisfied when

the domains are bounded or if the boundary consists of finitely many straight pieces (like a sector in \mathbb{R}^2 , for example).

Theorem 5 (Locality Principle for Robin Boundary Condition) *Assume that Ω and S are exact geometric matches on Ω_0 , as in Definition 1, with $\Omega_0 \subset \Omega \subset S \subset \mathbb{R}^n$. Assume that Ω is bounded. Let $K^S(t, x, y)$ and $K^\Omega(t, x, y)$ be the heat kernels for the Robin Laplacian with boundary condition (5) for $D = S$ and $D = \Omega$, respectively, for the same $c(x) \in \mathcal{L}^\infty(\partial S \cup \partial \Omega)$. Let $\alpha := \text{dist}(\Omega_0, S \setminus \Omega)$, and note that $\alpha > 0$ by our assumption of an exact geometric match. Define the auxiliary domain*

$$W := \{x \in \Omega : d(x, \Omega_0) \leq \alpha/2\}.$$

We make the following geometric assumptions:

1. Both S and Ω satisfy the two-sided (ε, h) -cone condition;
2. Both S and Ω have piecewise \mathcal{C}^3 boundaries, and there exists a constant $\sigma \in \mathbb{R}$ such that the second fundamental form satisfies $\mathbb{I} \geq -\sigma$ on all the \mathcal{C}^3 pieces of both ∂S and $\partial \Omega$.
3. For any sufficiently small $r > 0$ and any $t > 0$, we have

$$\sup_{x \in W} \int_0^t \int_{\partial S \setminus B(x,r)} \frac{1}{s^{\frac{n}{2}}} e^{-\frac{|x-z|^2}{s}} \sigma(dz) ds < \infty; \tag{6}$$

4. For all $r > 0$ and all $x \in \mathbb{R}^n$, and both $D = S$ and $D = \Omega$, there is a constant C_D such that

$$\mathcal{H}^{n-1}(\partial D \cap (B(x, r))) \leq C_D \text{Vol}_{n-1}(B_{n-1}(x, r)), \tag{7}$$

where \mathcal{H}^{n-1} denotes the $n - 1$ dimensional Hausdorff measure;

5. If $G_n(x, y)$ is the free Green’s function on \mathbb{R}^n , we have

$$\sup_{x \in W} \int_{\partial \Omega} G_n(x, y) \sigma(dy) < \infty. \tag{8}$$

Then, uniformly on $\overline{\Omega}_0 \times \overline{\Omega}_0$, we have Robin locality:

$$|K^S(t, z, z') - K^\Omega(t, z, z')| = O(t^\infty), \quad \forall z, z' \in \Omega_0, \quad t \rightarrow 0.$$

The assumptions that $\Omega \subset S$ and that Ω is bounded can both be removed:

Corollary 3 *Suppose we have an exact geometric match between Ω and S on the bounded domain Ω_0 , and the Robin coefficient $c(x)$ agrees on a common open, bounded neighborhood Ω_c of Ω_0 in Ω and S . Then, as long as Theorem 5 holds for the pairs (Ω_0, Ω) and (Ω_0, S) , the conclusion of Theorem 5 holds for the pair (Ω, S) .*

Proof Apply Theorem 5 to the pairs (Ω_0, Ω) and (Ω_0, S) , using the same W , then use the triangle inequality. \square

Before we prove Theorem 5, we discuss the geometric assumptions (6), (7), and (8), and give some sufficient conditions for them to hold. First, observe that regardless of what W is, (6) is immediately valid if S is a bounded domain whose boundary has finite $n - 1$ dimensional Lebesgue measure. It is also valid if S is an infinite circular sector, by a direct computation, part of which is presented below.

Example 1 Let $S = S_\gamma \subset \mathbb{R}^2$ be a circular sector of opening angle γ and infinite radius. Assume that W and Ω are bounded domains such that $W \subset \Omega \subset S$, and assume for simplicity that W contains the corner of S ; see Fig. 2 (the case where this does not happen is similar.) Then (6) holds. Indeed, let $r \in (0, \alpha/2)$ and $t > 0$, then

$$\begin{aligned} & \sup_{x \in W} \int_0^t \int_{\partial S \setminus B(x,r)} \frac{1}{s} e^{-\frac{|x-z|^2}{s}} \sigma(dz) ds \\ & \leq \sup_{x \in W} \int_0^t \int_{\partial S \setminus \partial \Omega} \frac{1}{s} e^{-\frac{|x-z|^2}{s}} \sigma(dz) ds + \sup_{x \in W} \int_0^t \int_{(\partial S \cap \partial \Omega) \setminus B(x,r)} \frac{1}{s} e^{-\frac{|x-z|^2}{s}} \sigma(dz) ds \\ & \leq 2 \int_0^t \int_0^\infty \frac{1}{s} e^{-\frac{\tau^2}{s}} d\tau ds + \int_0^t \frac{1}{s} e^{-\frac{r^2}{s}} \int_{(\partial S \cap \partial \Omega)} \sigma(dz) ds < \infty. \end{aligned}$$

Moreover, recalling the Green’s function in two dimensions (9), we also have

$$\sup_{x \in W} \int_{\partial \Omega \cap B(x,r)} G_n(x, y) \sigma(dy) = \sup_{x \in W} \int_{\partial \Omega \cap B(x,r)} |\ln |x - z|| \sigma(dz) \leq \int_0^\alpha |\ln \tau| d\tau < \infty.$$

As for (7), this is automatic if D is a bounded domain with piecewise \mathcal{C}^1 boundary. It is also true if D is a circular sector (in fact here $C_D = 2$).

The condition (8) is also easy to satisfy:

Proposition 1 *Assume that Ω is a bounded domain in \mathbb{R}^n which has piecewise \mathcal{C}^3 boundary. Let $W \subset \Omega$ be a compact set; then (8) holds.*

Proof Recall that

$$G_n(x, y) = \begin{cases} |\ln |x - y||, & \text{if } n = 2; \\ |x - y|^{2-n}, & \text{if } n \geq 3. \end{cases} \tag{9}$$

Since W is compact, it is enough to prove that

$$x \mapsto \int_{\partial \Omega} G_n(x, y) \sigma(dy) \tag{10}$$

is a continuous function on W .

Fix $x \in W$. Let $\varepsilon > 0$ and $\{x_j\}_{j=1}^\infty \subset W$ be a sequence such that $x_j \rightarrow x$. Since $\partial\Omega$ is piecewise \mathcal{C}^3 , and $G_n(x, y)$ is in \mathcal{L}^1_{loc} , we can choose $\delta > 0$ such that

$$\int_{\partial\Omega \cap B(x, 2\delta)} G_n(x, y)\sigma(dy) < \varepsilon, \quad \int_{\partial\Omega \cap B(x, 2\delta)} G_n(x_j, y)\sigma(dy) < \varepsilon, \quad (11)$$

for sufficiently large $j \in \mathbb{N}$, such that for these j we also have $|x - x_j| < \delta$. To see this, we note that $G_n(x, y) = G_n(|x - y|) = G_n(r)$, where $r = |x - y|$, and similarly, $G_n(x_j, y) = G_n(r_j)$ with $r_j = |x_j - y|$. Thus, choosing the radius, 2δ , sufficiently small, since G_n is locally $\mathcal{L}^1(\partial\Omega)$ integrable, and $\partial\Omega$ is piecewise \mathcal{C}^3 , we can make the above integrals as small as we like.

Now, we note that $G_n(x_j, y) \rightarrow G_n(x, y)$ as $j \rightarrow \infty$, for $y \in \partial\Omega \setminus B(x, 2\delta)$. Moreover, since Ω and thus $\partial\Omega$ are both compact, $G_n(x_j, y) < C = C(\delta)$ for $y \in \partial\Omega \setminus B(x, 2\delta)$. The Dominated Convergence Theorem therefore implies

$$\left| \int_{\partial\Omega \setminus B(x, 2\delta)} (G_n(x, y) - G_n(x_j, y))\sigma(dy) \right| < \varepsilon$$

for sufficiently large $j \in \mathbb{N}$. This, together with (11), implies that the function (10) is continuous on W .

In summary, we have

Corollary 4 *The locality principle, Theorem 5, holds in the case where Ω is a bounded domain in \mathbb{R}^n with piecewise \mathcal{C}^3 boundary, and S is any domain with piecewise \mathcal{C}^3 boundary such that Ω and S are an exact geometric match on the bounded subdomain Ω_0 as in Definition 1. Moreover, we assume that:*

1. Both S and Ω satisfy the two-sided (ε, h) -cone condition;
2. There exists a constant $\sigma \in \mathbb{R}$ such that the second fundamental form satisfies $\mathbb{I} \geq -\sigma$ on both ∂S and $\partial\Omega$;
3. S satisfies (6) and (7);
4. The Robin coefficient $c(x) \in \mathcal{L}^\infty(\partial S \cup \partial\Omega)$ agrees on a common open bounded neighborhood Ω_c in Ω and S .

Remark 2 In particular, all assumptions are satisfied if Ω is a bounded polygonal domain in \mathbb{R}^2 , and S is a circular sector in \mathbb{R}^2 .

The proof of Theorem 5 is accomplished by proving several estimates, in the form of lemmas and propositions below. Since the domains S and Ω satisfy the two-sided (ε, h) -cone condition, there are Gaussian upper bounds for the corresponding

Neumann heat kernels as given in (3). With this in mind, define

$$\begin{aligned}
 F_1(t) &:= \sup_{(s,x,z) \in (0,t] \times W \times (\overline{S \cap \Omega})} \left| H^S(s, x, z) - H^\Omega(s, x, z) \right|, \\
 F_2(t) &:= \sup_{(s,x,z) \in (0,t] \times W \times S \setminus \Omega} \left| H^S(s, x, z) \right|, \\
 F_3(t) &:= \sup_{(s,x,z) \in (0,t] \times W \times \partial\Omega \setminus \partial S} \left| H^\Omega(s, x, z) \right|.
 \end{aligned}$$

It now follows from (3) and Theorem 4 that

$$F(t) := \max(F_1(t), F_2(t), F_3(t)) = O(t^\infty), \quad t \rightarrow 0. \tag{12}$$

The reason we require the Neumann heat kernels is because, as in [24, 35],⁶ the Robin heat kernels, $K^S(t, x, y)$ and $K^\Omega(t, x, y)$, can be expressed in terms of $H^S(t, x, y)$ and $H^\Omega(t, x, y)$ in the following way. Define

$$k_0^D(t, x, y) = H^D(t, x, y), \quad D = S \text{ and } D = \Omega,$$

and

$$k_m^D(t, x, y) = \int_0^t \int_{\partial D} H^D(s, x, z) c(z) k_{m-1}^D(t-s, z, y) \sigma(dz) ds \tag{13}$$

for $m \in \mathbb{N}$. Then

$$K^S(t, x, y) = \sum_{m=1}^\infty k_m^S(t, x, y), \quad K^\Omega(t, x, y) = \sum_{m=1}^\infty k_m^\Omega(t, x, y).$$

Let us define the function

$$\begin{aligned}
 A(t, x) &:= \int_0^t \int_{\partial S} \left| H^S(s, x, z) c(z) \right| \sigma(dz) ds + \int_0^t \int_{\partial\Omega} \left| H^\Omega(s, x, z) c(z) \right| \sigma(dz) ds \\
 &=: A_1(t, x) + A_2(t, x)
 \end{aligned} \tag{14}$$

on $(0, 1] \times W$. The following lemma, in particular, shows that $A(t, x)$ is a well defined function.

Lemma 1 *The function $A(t, x)$ is uniformly bounded on $(0, 1] \times W$.*

⁶We note that the result is stated for compact domains. However, the construction is purely formal and works as long as the series converges. Under our assumptions, we shall prove that it does.

Proof For $n = 1$ the lemma follows from (3). Hence, we assume here $n \geq 2$. For any $x \in \overline{W}$, $A_j(t, x)$, $j = 1, 2$, is an increasing function with respect to the variable $t \in (0, 1]$. Therefore, it is sufficient to prove that $A_j(x) := A_j(1, x)$ is bounded on W , for $j = 1, 2$.

Let us choose $0 < \rho < \min(\alpha/2, 1)$. Without loss of generality, setting $C_1 = C_2 = 1$ in (3), we obtain

$$\begin{aligned} & A_1(x) + A_2(x) \\ & \leq \int_0^1 \int_{\partial S \setminus B(x, \rho)} s^{-\frac{n}{2}} e^{-\frac{|x-z|^2}{s}} |c(z)| \sigma(dz) ds + \int_0^1 \int_{\partial S \cap B(x, \rho)} s^{-\frac{n}{2}} e^{-\frac{|x-z|^2}{s}} |c(z)| \sigma(dz) ds \\ & + \int_0^1 \int_{\partial \Omega \setminus B(x, \rho)} s^{-\frac{n}{2}} e^{-\frac{|x-z|^2}{s}} |c(z)| \sigma(dz) ds + \int_0^1 \int_{\partial \Omega \cap B(x, \rho)} s^{-\frac{n}{2}} e^{-\frac{|x-z|^2}{s}} |c(z)| \sigma(dz) ds \\ & =: J_1(x) + J_2(x) + J_3(x) + J_4(x). \end{aligned}$$

The boundedness of $J_1(x)$ on W follows from (6) and the assumption that $c(z) \in \mathcal{L}^\infty$. For $J_3(x)$ we estimate using only that $\partial \Omega$ is bounded and thus, since it is piecewise \mathcal{C}^3 , has finite measure,

$$J_3(x) \leq \|c\|_\infty \int_0^1 \frac{1}{s^{\frac{n}{2}}} e^{-\frac{\rho^2}{s}} \int_{\partial \Omega \setminus B(x, \rho)} \sigma(dz) ds < \infty.$$

Since $\rho < \alpha/2$, $\partial S \cap B(x, \rho) = \partial \Omega \cap B(x, \rho)$ for $x \in W$, and hence by Fubini's theorem and a change of variables

$$\begin{aligned} J_2(x) = J_4(x) &= \int_0^1 \int_{\partial \Omega \cap B(x, \rho)} s^{-\frac{n}{2}} e^{-\frac{|x-z|^2}{s}} |c(z)| \sigma(dz) ds \\ &\leq \|c\|_\infty \int_{\partial \Omega \cap B(x, \rho)} \frac{1}{|x-z|^{n-2}} \int_{|x-z|^2}^{+\infty} \tau^{\frac{n}{2}-2} e^{-\tau} d\tau \sigma(dz). \end{aligned}$$

For $n > 2$, the second integral is uniformly bounded, and hence, (8) implies that $J_2(x)$ and $J_4(x)$ are bounded on W . If on the other hand $n = 2$, then

$$J_2(x) = J_4(x) = \int_{\partial \Omega \cap B(x, \rho)} |c(z)| \int_{|x-z|^2}^{+\infty} \tau^{-1} e^{-\tau} d\tau \sigma(dz).$$

Since $\rho < 1$, $\rho^2 < \rho < 1$, so we can write

$$\begin{aligned} J_2(x) = J_4(x) &\leq \int_{\partial \Omega \cap B(x, \rho)} |c(z)| \int_{|x-z|^2}^1 \tau^{-1} d\tau \sigma(dz) + \int_{\partial \Omega \cap B(x, \rho)} |c(z)| \int_1^{+\infty} e^{-\tau} d\tau \sigma(dz) \\ &\leq \|c\|_\infty \int_{\partial \Omega \cap B(x, \rho)} |\ln|x-z|^2| \sigma(dz) + \|c\|_\infty \int_{\partial \Omega \cap B(x, \rho)} \sigma(dz), \end{aligned}$$

which is finite by (8), the boundedness of $\partial\Omega$ and the piecewise \mathcal{C}^3 smoothness of the boundary. \square

Corollary 5 *In the notation of Lemma 1, we have*

$$\lim_{T \rightarrow 0} \sup_{(t,x) \in (0,T] \times W} A(t, x) = 0.$$

Proof Consider the functions $A_j(t, x)$. They are monotone increasing in t for each x , and they are continuous in x for each t by continuity of solutions to the heat equation. We claim that as $t \rightarrow 0$, $A(t, x)$ approaches zero pointwise. To see this write the time integral from 0 to t in each $A_j(t, x)$, $j = 1, 2$, as a time integral over $[0, 1]$ by multiplying the integrand by the characteristic function $\chi_{[0,t]}$. For example,

$$A_1(t, x) = \int_0^1 \int_{\partial S} \chi_{[0,t]} |H^S(s, x, z)c(z)| \sigma(dz) ds.$$

The integrands are bounded by $|H^S(s, x, z)c(z)|$, which is integrable by Lemma 1. For each x , they converge to zero as $t \rightarrow 0$. So by the Dominated Convergence Theorem applied to each $A_j(t, x)$, we see that $A(t, x) \rightarrow 0$ as $t \rightarrow 0$ for each x .

Now we have a monotone family of continuous functions converging pointwise to a continuous function (zero) on the compact set W . By Dini’s theorem, this convergence is in fact uniform, which is precisely what we want. \square

To use this, fix a small number A to be chosen later. Then Corollary 5 allows us to find $T > 0$ such that

$$A(t, x) < A, \quad (t, x) \in (0, T] \times W. \tag{15}$$

Next we prove the following two auxiliary propositions.

Proposition 2 *The following inequality holds with $D = S$ and $D = \Omega$:*

$$\int_0^t \int_{\partial D} |k_m^D(s, x, z)c(z)| \sigma(dz) ds \leq 2^{m+1} A^{m+1} \tag{16}$$

on $(0, T] \times W$, for any $m \in \mathbb{N}$. Moreover, an identical inequality holds when $k_m^D(s, x, z)$ is replaced by $k_m^D(s, z, x)$.

Proof By induction. For $m = 0$, recalling the definition of $A(t, x)$, (14),

$$\int_0^t \int_{\partial D} |k_0^D(s, x, z)c(z)| \sigma(dz) ds = \int_0^t \int_{\partial D} |H^D(s, x, z)c(z)| \sigma(dz) ds \leq A(t, x) < A.$$

We have thus verified the base case. Now, we assume that (16) holds for $k \leq m$. Consider $k = m + 1$:

$$\begin{aligned} & \int_0^t \int_{\partial D} \left| k_{m+1}^D(s, x, z)c(z) \right| \sigma(dz) ds \\ &= \int_0^t \int_{\partial D} \int_0^s \int_{\partial D} \left| H^D(\tau, z, \zeta) k_m^D(s - \tau, \zeta, x) c(\zeta) c(z) \right| \sigma(d\zeta) d\tau \sigma(dz) ds. \end{aligned}$$

Changing variables:

$$\begin{aligned} & \int_0^t \int_{\partial D} \int_0^s \int_{\partial D} \left| H^D(\tau, z, \zeta) k_m^D(s - \tau, \zeta, x) c(\zeta) c(z) \right| \sigma(d\zeta) d\tau \sigma(dz) ds \\ & \leq \int_0^t \int_{\partial D} \int_0^s \int_{\partial D} \left| H^D(s - \tau, z, \zeta) k_m^D(\tau, \zeta, x) c(\zeta) c(z) \right| \sigma(d\zeta) d\tau \sigma(dz) ds \\ & \leq \int_0^t \int_{\partial D} \int_0^t \int_{\partial D} \left| H^D(|s - \tau|, z, \zeta) k_m^D(\tau, \zeta, x) c(\zeta) c(z) \right| \sigma(d\zeta) d\tau \sigma(dz) ds \\ & \leq \int_0^t \int_{\partial D} \left(\int_0^t \int_{\partial D} \left| H^D(|s - \tau|, z, \zeta) c(z) \right| \sigma(dz) ds \right) \left| k_m^D(\tau, \zeta, x) c(\zeta) \right| \sigma(d\zeta) d\tau. \end{aligned} \tag{17}$$

For the integrand, we compute

$$\begin{aligned} & \int_0^t \int_{\partial D} \left| H^D(|s - \tau|, z, \zeta) c(z) \right| \sigma(dz) ds \\ &= \int_0^\tau \int_{\partial D} \left| H^D(|s - \tau|, z, \zeta) c(z) \right| \sigma(dz) ds + \int_\tau^t \int_{\partial D} \left| H^D(|s - \tau|, z, \zeta) c(z) \right| \sigma(dz) ds \\ &= \int_0^\tau \int_{\partial D} \left| H^D(\tau - s, z, \zeta) c(z) \right| \sigma(dz) ds + \int_0^{t-\tau} \int_{\partial D} \left| H^D(s, z, \zeta) c(z) \right| \sigma(dz) ds < 2A. \end{aligned}$$

Therefore, from the induction hypothesis and (17), we obtain

$$\begin{aligned} & \int_0^t \int_{\partial D} \int_0^s \int_{\partial D} \left| H^D(\tau, z, \zeta) k_m^D(s - \tau, \zeta, x) c(\zeta) c(z) \right| \sigma(d\zeta) d\tau \sigma(dz) ds \\ & \leq 2A \int_0^t \int_{\partial D} \left| k_m^D(\tau, \zeta, x) c(\zeta) \right| \sigma(d\zeta) d\tau \leq 2A \cdot 2^{m+1} A^{m+1} = 2^{m+2} A^{m+2}, \end{aligned}$$

as desired.

The estimates with x and z reversed are proved similarly. Note in particular that the base case works because $k_0^D = H_0^D$ is a Neumann heat kernel and is thus symmetric in its spatial arguments. \square

We need one more lemma concerning pointwise bounds for k_m^D , which uses the geometric assumption (7).

Lemma 2 *Let $D = S$ or Ω . There exists $T_0 > 0$ such that for all m , all $t < T_0$, all $x \in D$, and all $y \in D$,*

$$|k_m^D(t, x, y)| \leq \frac{C_1}{2^m} t^{-\frac{n}{2}} e^{-\frac{|x-y|^2}{C_2 t}}.$$

Proof The proof proceeds by induction. The base case is $m = 0$, which is (3).

Now assume we have the result for $k = m$. Using the iterative formula (13), we have

$$|k_{m+1}^D(t, x, y)| \leq \|c\|_\infty \int_0^t \int_{\partial D} |H^D(s, x, z) k_m^D(t-s, z, y)| \sigma(dz) ds. \tag{18}$$

Using (3) and the inductive hypothesis, we see that the integrand is bounded by

$$C_1 C_1 2^{-m} s^{-n/2} (t-s)^{-n/2} e^{-\frac{1}{C_2} \left(\frac{|x-z|^2}{s} + \frac{|z-y|^2}{t-s} \right)}.$$

First assume that D is a half-space. We do the estimate in the case $n = 2$, because the general case is analogous. Hence, we use the coordinates $x = (x_1, x_2)$, $y = (y_1, y_2)$, $z = (z_1, z_2)$, and estimate using $\{z_2 = 0\} \subset \mathbb{R}^2$ for ∂D . Dropping the constant factors, and saving the integral with respect to time for later, we therefore estimate

$$\int_{\mathbb{R}} s^{-1} (t-s)^{-1} e^{-\frac{|x-z|^2}{C_2 s} - \frac{|y-z|^2}{C_2 (t-s)}} dz_1.$$

Without loss of generality, we shall assume that $x = (0, 0)$. Then we are estimating

$$\int_{\mathbb{R}} s^{-1} (t-s)^{-1} e^{-\frac{z_1^2 (t-s) - s|y-z|^2}{C_2 s (t-s)}} dz_1.$$

Since $z \in \partial D$, we have $z_2 = 0$. For the sake of simplicity, set $y_2 = 0$; the case where y_2 is nonzero is similar. Given this assumption, we set

$$z := z_1, \quad y := y_1,$$

and estimate

$$\int_{\mathbb{R}} s^{-1} (t-s)^{-1} e^{-\frac{-z^2 (t-s) - s(y-z)^2}{C_2 s (t-s)}} dz.$$

We do the standard trick of completing the square in the exponent. This gives

$$\int_{\mathbb{R}} s^{-1}(t-s)^{-1} \exp \left[- \left(\frac{\sqrt{t}z - \frac{sy}{\sqrt{t}}}{\sqrt{C_2}\sqrt{s}\sqrt{t-s}} \right)^2 - \frac{y^2}{C_2(t-s)} + \frac{sy^2}{C_2t(t-s)} \right] dz.$$

We therefore compute the integral over \mathbb{R} in the standard way, obtaining

$$\begin{aligned} s^{-1/2}(t-s)^{-1/2} \sqrt{\frac{C_2\pi}{t}} e^{-\frac{y^2}{C_2(t-s)} + \frac{sy^2}{C_2t(t-s)}} &= s^{-1/2}(t-s)^{-1/2} \sqrt{\frac{C_2\pi}{t}} e^{\frac{-ty^2+sy^2}{C_2t(t-s)}} \\ &= s^{-1/2}(t-s)^{-1/2} \sqrt{\frac{C_2\pi}{t}} e^{-\frac{y^2}{C_2t}}. \end{aligned}$$

Finally, we compute the integral with respect to s ,

$$\int_0^t \frac{1}{\sqrt{s}} \frac{1}{\sqrt{t-s}} ds = \pi.$$

Hence, the total expression is bounded from above by

$$\pi \sqrt{\frac{C_2\pi}{t}} e^{-\frac{y^2}{C_2t}}.$$

Since we had assumed that $x = 0$, we see that this is indeed

$$\pi \sqrt{\frac{C_2\pi}{t}} e^{-\frac{|x-y|^2}{C_2t}}.$$

Recalling the constant factors, we have

$$|k_{m+1}^D(t, x, y)| \leq C_1 C_1 \|c\|_{\infty} 2^{-m} \pi \sqrt{\frac{C_2\pi}{t}} e^{-\frac{|x-y|^2}{C_2t}}.$$

Now we note that the power of t is $t^{-(n-1)/2}$ for dimension $n = 2$. Hence, we re-write the above estimate as

$$|k_{m+1}^D(t, x, y)| \leq C_1 C_1 \|c\|_{\infty} 2^{-m} \pi \sqrt{t} t^{-1} \sqrt{C_2\pi} e^{-\frac{|x-y|^2}{C_2t}}.$$

We then may choose for example

$$\begin{aligned} t \leq T_0 &= \frac{1}{4(C_1 + 1)^2 (\|c\|_{\infty} + 1)^2 \pi^3 (C_2 + 1)} \\ \implies \sqrt{t} &\leq \frac{1}{2(C_1 + 1) (\|c\|_{\infty} + 1) \pi^{\frac{3}{2}} \sqrt{C_2 + 1}}. \end{aligned}$$

This ensures that

$$|k_{m+1}^D(t, x, y)| \leq C_1 2^{-(m+1)} t^{-\frac{n}{2}} e^{-\frac{|x-y|^2}{C_2 t}}, \quad n = 2.$$

We note that in general, for \mathbb{R}^n , by estimating analogously, noting that the integral will be over \mathbb{R}^{n-1} , we obtain

$$|k_{m+1}^D(t, x, y)| \leq \|c\|_\infty 2^{-m} \pi (C_2 \pi)^{n/2} t^{-(n-1)/2} e^{-\frac{|x-y|^2}{C_2 t}}.$$

So, in the general- n case, we let

$$T_0 = \frac{1}{4(C_1 + 1)^2 (\|c\|_\infty + 1)^2 \pi^{2+n} (C_2 + 1)^n}.$$

Then, for all $t \leq T_0$, we have

$$|k_{m+1}^D(t, x, y)| \leq C_1 2^{-m-1} t^{-\frac{n}{2}} e^{-\frac{|x-y|^2}{C_2 t}}.$$

Now consider the case where D is a general domain, not necessarily a half-space. As before, we have

$$|k_{m+1}^D(t, x, y)| \leq \frac{\|c\|_\infty C_1^2}{2^m} \int_0^t \int_{\partial D} s^{-n/2} (t-s)^{-n/2} e^{-\frac{1}{C_2} \left(\frac{|x-z|^2}{s} + \frac{|z-y|^2}{t-s} \right)} \sigma(dz) ds. \tag{19}$$

We claim that the right-hand side of (19) is less than or equal to C_D , the constant from (7), times the corresponding integral in the case where D is a half-plane through x and y . Assuming this claim, we get the same bound as for a half-plane, but with an extra C_D , and adjusting T_0 to absorb C_D as well, by putting an extra $(C_D + 1)^2$ in the denominator, completes the proof.

To prove this claim, we use the so-called layer cake representation: rewrite the right-hand side of (19), without the outside constants, as

$$\int_0^t s^{-n/2} (t-s)^{-n/2} \int_{\partial D} \int_0^\infty \chi_{\{f(s,t,x,y,z) < a\}} e^{-a} da \sigma(dz) ds, \tag{20}$$

where naturally

$$f(s, t, x, y, z) := \frac{1}{C_2} \left(\frac{|x-z|^2}{s} + \frac{|z-y|^2}{t-s} \right).$$

The representation (20) may seem odd at first but reverts to (19) upon integration in a . Switching the order of integration in (20) (valid by Fubini-Tonelli, since

everything is positive) and evaluating the z -integral, this becomes

$$\int_0^t s^{-n/2}(t-s)^{-n/2} \int_0^\infty \mathcal{H}^{n-1}(\partial D \cap \{z : f(s, t, x, y, z) < a\}) e^{-a} \, dads. \quad (21)$$

Let us more closely examine the set $\{z : f(s, t, x, y, z) < a\}$. It is the set where

$$\left(1 - \frac{s}{t}\right) |x - z|^2 + \frac{s}{t} |z - y|^2 < \frac{1}{t} C_2 a s (t - s).$$

It is straightforward to compute that this set is in fact a ball centered at the point $P(s, t, x, y) := (1 - \frac{s}{t})x + \frac{s}{t}y$, with radius squared equal to

$$R^2(s, t, x, y) := \max \left\{ 0, \frac{1}{t} C_2 a s (t - s) - \frac{s}{t} \left(1 - \frac{s}{t}\right) |y - x|^2 \right\}.$$

Therefore (21) equals

$$\int_0^t s^{-n/2}(t-s)^{-n/2} \int_0^\infty \mathcal{H}^{n-1}(\partial D \cap B_n(P, R)) e^{-a} \, dads. \quad (22)$$

By the assumption (7), this is bounded by

$$C_D \int_0^t s^{-n/2}(t-s)^{-n/2} \int_0^\infty \text{Vol}_{n-1}(B_{n-1}(P, R)) e^{-a} \, dads. \quad (23)$$

However, in the event that D is a half-space with x and $y \in \partial D$ (so also $P \in \partial D$), we have $\partial D \cap B_n(P, R) = B_{n-1}(P, R)$, so (22) equals

$$\int_0^t s^{-n/2}(t-s)^{-n/2} \int_0^\infty \text{Vol}_{n-1}(B_{n-1}(P, R)) e^{-a} \, dads. \quad (24)$$

Therefore, the integral (22) for general D is bounded by C_D times the integral (22) for a half-space. Since (22) is equal to the right-hand side of (19) without the preceding constants, the claim is proven. This completes the proof of Lemma 2. \square

Remark 3 The key is that the integral is half an order better in t than the true heat kernel, which is a critical feature of the difference between Robin and Neumann heat kernels. It allows us to utilize the extra \sqrt{t} to obtain the additional factor of 2^{-m} which is required for the induction step in the next proposition.

Now, we establish the main estimate to prove Theorem 5. Let

$$G(t) = \max \left\{ F(t), 2C_1 t^{-(n/2)} e^{-\frac{(\alpha/2)^2}{C_2 t}} \right\}.$$

We note that of course we still have $G(t) = O(t^\infty)$ as $t \downarrow 0$.

Proposition 3 *There exists $T > 0$ such that the estimate*

$$|k_m^S(t, x, y) - k_m^\Omega(t, x, y)| \leq G(t) \cdot 7 \cdot 2^{-m} \tag{25}$$

holds for all $(t, x, y) \in (0, T] \times W \times \overline{\Omega}_0$.

Proof We choose T small enough so that $T < T_0$ in Proposition 3 and so that (15) holds with $A = 1/4$.

Now proceed by induction. The base case is instantaneous by definition of k_0 and of $F(t)$, using our locality principle for the Neumann case. So assume that (25) holds for $k = m$; we will prove it for $k = m + 1$. Using some algebraic manipulations,

$$\begin{aligned} I &:= |k_{m+1}^S(t, x, y) - k_{m+1}^\Omega(t, x, y)| \leq I_1 + I_2 + I_3 \\ &:= \int_0^t \int_{\partial S \cap \partial \Omega} \left| H^S(s, x, z) k_m^S(t-s, z, y) - H^{\Omega}(s, x, z) k_m^\Omega(t-s, z, y) \right| |c(z)| \sigma(dz) ds \\ &\quad + \int_0^t \int_{\partial S \setminus \partial \Omega} |H^S(s, x, z) k_m^S(t-s, z, y) c(z)| \sigma(dz) ds \\ &\quad + \int_0^t \int_{\partial \Omega \setminus \partial S} |H^{\Omega}(s, x, z) k_m^\Omega(t-s, z, y) c(z)| \sigma(dz) ds. \end{aligned}$$

We estimate these terms separately, beginning with I_1 .

$$\begin{aligned} I_1 &\leq \int_0^t \int_{\partial S \cap \partial \Omega} \left| H^S(s, x, z) - H^{\Omega}(s, x, z) \right| |k_m^S(t-s, z, y)| |c(z)| \sigma(dz) ds \\ &\quad + \int_0^t \int_{W \cap \partial S \cap \partial \Omega} \left| k_m^S(t-s, z, y) - k_m^\Omega(t-s, z, y) \right| |H^{\Omega}(s, x, z)| |c(z)| \sigma(dz) ds \\ &\quad + \int_0^t \int_{(\Omega \setminus W) \cap \partial S \cap \partial \Omega} \left| k_m^S(t-s, z, y) - k_m^\Omega(t-s, z, y) \right| |H^{\Omega}(s, x, z)| |c(z)| \sigma(dz) ds. \end{aligned}$$

The first term in the first integral is bounded by $F(t)$, since $x \in W$ and $z \in \overline{\Omega}$, so we may pull it out. We estimate the other term with Proposition 2 and get a bound of $F(t) \cdot 2^{m+1} A^{m+1} = F(t) \cdot 2^{-(m+1)}$ for the first integral.

For the second integral, we pull out the supremum of the first term using the inductive hypothesis. We estimate the other term using the definition of A and we get a bound of $G(t) \cdot 7 \cdot 2^{-m-2}$.

For the third integral, we use Lemma 2 to pull out the first term, ignoring the difference and just estimating both k terms separately. Since $|z - y| \geq \alpha/2$ on this region, the supremum is less than $2^{-m} G(t)$ by Lemma 2. We estimate the other term using the definition of A and we get $1/4$, giving a bound of $2^{-m-2} G(t)$. Overall, we have

$$I_1 \leq G(t)(2^{-m-1} + 7 \cdot 2^{-m-2} + 2^{-m-2}).$$

Next we estimate the terms I_2, I_3 . In each, we pull out the supremum of $H^S(s, x, z)$ over the relevant region, and observe that it is bounded above by $F(t)$. For the term remaining in the integral we use Proposition 2. Since $F(t) \leq G(t)$, we obtain a bound of $G(t) \cdot 2^{-m-1}$ for each of these two terms. Putting it all together, we see

$$I \leq G(t)(3 \cdot 2^{-m-1} + 7 \cdot 2^{-m-2} + 2^{-m-2}) = G(t) \cdot 2^{-m-1} \left(3 + \frac{7}{2} + \frac{1}{2} \right) = G(t) \cdot 7 \cdot 2^{-m-1},$$

as desired. □

Proof Finally, we prove Theorem 5. By Proposition 3,

$$\begin{aligned} |K^S(t, x, y) - K^\Omega(t, x, y)| &\leq \sum_{m=0}^{\infty} |k_m^S(t, x, y) - k_m^\Omega(t, x, y)| \\ &\leq \sum_{m=0}^{\infty} 7G(t)2^{-m} = 14 \cdot G(t), \end{aligned}$$

which is $O(t^\infty)$ as $t \rightarrow 0$. □

3 Hearing the Corners of a Drum

As a consequence of the work in the previous section, the locality principle holds for both the Neumann and Robin boundary conditions when Ω is a bounded domain as described in Theorem 1, and S is either a whole space, a half-space, a circular sector, or a smoothly bounded domain which is an exact geometric match for some piece of Ω . Therefore, to compute the heat trace expansion for $\Omega \subset \mathbb{R}^2$ satisfying the hypotheses of Theorem 1, it suffices to chop the domain into pieces and, depending on the piece, replace the true heat kernel with one of the following:

- the heat kernel for an infinite circular sector with the same opening angle and boundary conditions near a corner of Ω ,
- the heat kernel for a smoothly bounded domain which is an exact match to Ω away from all the corners. Note that such a domain can be produced by rounding off each corner.

Henceforth we consider the Neumann boundary condition or the classical Robin boundary condition as in (2), so that in (5), $c(x)$ is a constant, specifically $c(x) = \alpha/\beta$. In [23] we prove that the “corner contribution” for the Robin boundary condition is identical to that for the Neumann boundary condition. In the aforementioned work, we determine the Green’s function for a circular sector of

infinite radius and opening angle γ , with the Neumann boundary condition in polar coordinates:

$$G_N(s, r, \phi, r_0, \phi_0) = \frac{1}{\pi^2} \int_0^\infty K_{i\mu}(r\sqrt{s})K_{i\mu}(r_0\sqrt{s}) \times \left\{ \cosh(\pi - |\phi_0 - \phi|)\mu + \frac{\sinh \pi \mu}{\sinh \gamma \mu} \cosh(\phi + \phi_0 - \gamma)\mu + \frac{\sinh(\pi - \gamma)\mu}{\sinh \gamma \mu} \cosh(\phi - \phi_0)\mu \right\} d\mu. \tag{26}$$

Above, $K_{i\mu}$ is the modified Bessel function of the second kind, and s is the spectral parameter of the resolvent, $(\Delta + s)^{-1}$. The derivation of these formulas stems from Fedosov’s study of Kontorovich-Lebedev transforms [7] and shall be presented in our forthcoming work [23]. Using functional calculus techniques, as we do in [23], one may rigorously justify the statement that

$$H(t, r, \phi, r_0, \phi_0) = \mathcal{L}^{-1} (G(s, r, \phi, r_0, \phi_0)) (t),$$

where H denotes the heat kernel and \mathcal{L}^{-1} denotes the inverse Laplace transform taken with respect to s . This allows us to pass from the Green’s functions to the heat kernels on a sector, and we may then compute the short time asymptotic expansions of the heat traces using our locality principles.

3.1 Heat Trace Calculations

Let Ω be a domain with corners as described in Theorem 1. Assume that Ω has n corners. Let \mathcal{N}_i be a neighborhood of the i^{th} corner consisting of a circular sector of radius R , with R sufficiently small so that each \mathcal{N}_i can be taken to equal to Ω_0 in the definition of exact geometric match corresponding to S_i , where S_i is the infinite circular sector of interior angle equal to the angle θ_i . Then let U be a smoothly bounded domain such that U can be taken equal to S in the definition of exact geometric match, with $\Omega_0 = \Omega \setminus \{\mathcal{N}_i\}_{i=1}^n$. By our locality principles, the heat trace

$$\int_{\Omega} H^{\Omega}(t, z, z) dz = \int_{\Omega \setminus \{\mathcal{N}_i\}_{i=1}^n} H^U(t, z, z) dz + \sum_{i=1}^n \int_{\mathcal{N}_i} H^{S_i}(t, z, z) dz + O(t^\infty). \tag{27}$$

The calculation of the asymptotics of the integral of $H^U(t, z, z)$ is well-known and may be extracted from [21], [25] and [35]. More interesting is the calculation

of the heat trace near the corners. Let us define:

$$\begin{aligned}
 A &:= \int_0^\infty K_{i\mu}(r\sqrt{s})K_{i\mu}(r_0\sqrt{s}) \cosh(\pi - |\phi_0 - \phi|)\mu d\mu, \\
 B &:= \int_0^\infty K_{i\mu}(r\sqrt{s})K_{i\mu}(r_0\sqrt{s}) \frac{\sinh \pi\mu}{\sinh \gamma\mu} \cosh(\phi + \phi_0 - \gamma)\mu d\mu \\
 C &:= \int_0^\infty K_{i\mu}(r\sqrt{s})K_{i\mu}(r_0\sqrt{s}) \frac{\sinh(\pi - \gamma)\mu}{\sinh \gamma\mu} \cosh(\phi - \phi_0)\mu d\mu,
 \end{aligned}$$

With this terminology, the Neumann Green’s function for an infinite sector is given by

$$\frac{1}{\pi^2} (A + B + C).$$

We shall compute the heat trace contributions from each of these terms. In each calculation, we will take the inverse Laplace transform, restrict to the angular diagonal $\phi = \phi_0$ (which commutes with \mathcal{L}^{-1}), integrate in ϕ (same), restrict to $r = r_0$, and integrate in r .

3.1.1 Heat Trace Contribution from the A Term

Setting $\phi = \phi_0$, we have by Gradshteyn and Ryzhik [10, 6.794.1]

$$\int_0^\infty K_{ix}(r\sqrt{s})K_{ix}(r_0\sqrt{s}) \cosh(\pi x)dx = \frac{\pi}{2} K_0(\sqrt{(r - r_0)^2s}).$$

Then, by Erdelyi et al. [6, 5.16.35], we have

$$\mathcal{L}^{-1} [A] = \mathcal{L}^{-1} \left[\frac{\pi}{2} K_0(\sqrt{(r - r_0)^2s}) \right] = \frac{\pi}{2} \frac{1}{2} \frac{1}{t} e^{-\frac{(r-r_0)^2}{4t}}.$$

Hence for $\phi = \phi_0$,

$$\frac{1}{\pi^2} \mathcal{L}^{-1}(A) = \frac{e^{-\frac{(r-r_0)^2}{4t}}}{4\pi t}. \tag{28}$$

Setting $r = r_0$ gives $(4\pi t)^{-1}$, and integrating over \mathcal{N}_i , the contribution from this term to the heat trace is the usual area term:

$$\frac{A(\mathcal{N}_i)}{4\pi t}.$$

3.1.2 Heat Trace Contribution from the B Term

Now we investigate the contribution from B . The first simplification is to restrict to $\phi = \phi_0$, then compute

$$\int_0^\gamma B|_{\phi=\phi_0} d\phi = \int_0^\gamma K_{ix}(r\sqrt{s})K_{ix}(r_0\sqrt{s})\frac{\sinh \pi x}{\sinh \gamma x} \cosh(2\phi - \gamma)x d\phi.$$

The only dependence on the angle is in the cosh term, which may be explicitly integrated, and we obtain

$$\int_0^\gamma B|_{\phi=\phi_0} d\phi = \int_0^\infty K_{ix}(r\sqrt{s})K_{ix}(r_0\sqrt{s})\frac{\sinh \pi x}{x} dx = \frac{\pi^2}{2} I_0(r_0\sqrt{s})K_0(r\sqrt{s}),$$

where in the last equality we have used [10, 6.794.10]. Now take the inverse Laplace transform:

$$\mathcal{L}^{-1}\left[\int_0^\gamma B|_{\phi=\phi_0} d\phi\right] = \mathcal{L}^{-1}\left[\frac{\pi^2}{2} I_0(r_0\sqrt{s})K_0(r\sqrt{s})\right] = \frac{\pi^2}{2} \frac{1}{2t} e^{-\frac{r^2+r_0^2}{4t}} I_0\left(\frac{rr_0}{2t}\right). \tag{29}$$

Thus, we see that

$$\frac{1}{\pi^2} \mathcal{L}^{-1}\left[\int_0^\gamma B|_{\phi=\phi_0} d\phi\right] = \frac{1}{4t} e^{-\frac{r^2+r_0^2}{4t}} I_0\left(\frac{rr_0}{2t}\right). \tag{30}$$

To compute the trace, we make a change of variables, by setting

$$u = \frac{r^2}{2t}, \quad du = \frac{r}{t} dr.$$

Therefore,

$$\frac{1}{4t} \int_0^R e^{-r^2/2t} I_0\left(\frac{r^2}{2t}\right) r dr = \frac{1}{4} \int_0^{\frac{R^2}{2t}} e^{-u} I_0(u) du.$$

By Watson [34, p. 79 (3)] with $\nu = 1$,

$$uI_1'(u) + I_1(u) = uI_0(u). \tag{31}$$

By Watson [34, p. 79 (4)] with $\nu = 0$,

$$uI_0'(u) = uI_1(u). \tag{32}$$

We use these to compute

$$\begin{aligned} \frac{d}{du} (e^{-u}u(I_0(u) + I_1(u))) &= e^{-u} (-uI_0'(u) - I_0(u) + I_0(u) + I_1(u) + uI_0'(u) + uI_1'(u)) \\ &= e^{-u} (-uI_1'(u) + I_0(u) + uI_0'(u)) \quad (\text{by (31)}) \\ &= e^{-u}I_0(u) \quad (\text{by (32)}). \end{aligned}$$

Next, define

$$g(u) := e^{-u}u(I_0(u) + I_1(u)), \tag{33}$$

and note that we have computed

$$g'(u) = e^{-u}I_0(u).$$

We therefore have

$$\int_0^{R^2/2t} e^{-u}I_0(u)du = (g(R^2/2t) - g(0)).$$

Since $I_0(0) = 1$ and $I_1(0) = 0$ [34], it follows that $g(0) = 0$, and we therefore compute that

$$\int_0^{R^2/2t} e^{-u}I_0(u)du = g(R^2/2t) = e^{-R^2/2t} \frac{R^2}{2t} (I_0(R^2/2t) + I_1(R^2/2t)).$$

For large arguments, the Bessel functions admit the following asymptotic expansions (see [34])

$$I_j(x) = \frac{e^x}{\sqrt{2\pi x}} \left(1 - \frac{1}{2x} \left(j^2 - \frac{1}{4} \right) + \sum_{k=2}^{\infty} c_{j,k} x^{-k} \right), \quad x \gg 0, \quad j = 0, 1.$$

Consequently, for $x = R^2/2t$,

$$\begin{aligned} g(R^2/2t) &= \frac{R^2}{2t} e^{-R^2/2t} (I_0(R^2/2t) + I_1(R^2/2t)) = \frac{R^2}{2t} \left(\frac{2}{\sqrt{2\pi(R^2/2t)}} \right) - O\left(\frac{1}{(R^2/2t)^{3/2}} \right) \\ &= \frac{R}{\sqrt{\pi t}} + O(\sqrt{t}), \quad t \downarrow 0. \end{aligned}$$

Recalling the factor of $\frac{1}{4}$, we see that the trace of B contributes

$$\frac{R}{4\sqrt{\pi t}} + O(\sqrt{t}), \quad t \downarrow 0. \tag{34}$$

Observe that this is precisely the usual perimeter term:

$$\frac{\ell(\mathcal{N}_i \cap \partial\Omega)}{8\sqrt{\pi t}} + O(\sqrt{t}).$$

3.1.3 Heat Trace Contribution from the C Term

Next, we compute the trace of the C term. This is done following [29]. The cosh term drops out when $\phi = \phi_0$. Integrating with respect to the angle gives a factor of γ . We define

$$R(t) = -\mathcal{L}^{-1} \left(\frac{\gamma}{\pi^2} \int_0^\infty \frac{\sinh(\pi - \gamma)x}{\sinh(\gamma x)} \int_R^\infty K_{ix}^2(r\sqrt{s})rdr \right).$$

It is shown in [29] that

$$R(t) = O(e^{-c/t}),$$

and in fact an estimate is also obtained there for the constant $c > 0$. Hence, it suffices to compute

$$\mathcal{L}^{-1} \left(\frac{\gamma}{\pi^2} \int_0^\infty dx \frac{\sinh(\pi - \gamma)x}{\sinh \gamma x} \int_0^\infty K_{ix}^2(r\sqrt{s})rdr \right).$$

Here we use [10, 6.521.3]. As in that notation we have $a = s = b$, we must compute instead the limit of the expression as $b \rightarrow a$,

$$\lim_{b \rightarrow a} \frac{\pi(ab)^{-v}(a^v + b^v)}{2 \sin(v\pi)(a + b)} \frac{f(a) - f(b)}{a - b}, \quad f(t) = t^v.$$

Then, since

$$f'(t) = vt^{v-1}$$

we have

$$\lim_{b \rightarrow a} \frac{\pi(ab)^{-v}(a^v + b^v)}{2 \sin(v\pi)(a + b)} \frac{f(a) - f(b)}{a - b} = \frac{\pi a^{-2v}(2a^v)}{4 \sin(v\pi)a} v a^{v-1} = \frac{\pi v}{2 \sin(v\pi)a^2}.$$

Inserting our parameters, we have that

$$\int_0^\infty K_{ix}^2(r\sqrt{s})rdr = \frac{\pi x}{2 \sinh(\pi x)s}.$$

So we must compute

$$\mathcal{L}^{-1} \left\{ \frac{\gamma}{\pi^2} \int_0^\infty \frac{\sinh(\pi - \gamma)x}{\sinh \gamma x} \frac{\pi x}{2s \sinh(\pi x)} dx \right\}.$$

This calculation has been done in [29, p. 122] using [10]; we have independently verified these calculations as well. The result is given in [29, (2.10)]:

$$\frac{\pi^2 - \gamma^2}{24\pi \gamma}.$$

Thus, we see that C contributes to the trace the usual “corner contribution”:

$$\frac{\pi^2 - \gamma^2}{24\pi \gamma} + O(t^\infty). \tag{35}$$

3.2 Robin Boundary Condition

The Robin heat kernel has an additional contribution from the boundary. In [2, (3.19)], and more classically [3, §14.2], the Robin heat kernel for a half-space is computed, and it is equal to the Neumann heat kernel plus one additional term. This term is

$$E = -\frac{1}{\sqrt{4\pi t}} e^{-\frac{(x-x')^2}{4t}} \frac{\alpha}{\beta} e^{\frac{\alpha(y+y')}{\beta}} e^{\frac{\alpha^2 t}{\beta^2}} \operatorname{erfc} \left(\frac{y+y'}{\sqrt{4t}} + \frac{\alpha}{\beta} \sqrt{t} \right).$$

3.2.1 Trace of the E Term

The restriction of E to the diagonal yields

$$-\frac{1}{\sqrt{4\pi t}} \frac{\alpha}{\beta} e^{\frac{2\alpha y}{\beta}} e^{\frac{\alpha^2 t}{\beta^2}} \operatorname{erfc} \left(\frac{y}{\sqrt{t}} + \frac{\alpha}{\beta} \sqrt{t} \right).$$

We have to integrate this over the semicircle $x^2 + y^2 \leq R^2$. Doing the x -integration first yields

$$\int_0^R -\frac{\sqrt{R^2 - y^2}}{\sqrt{\pi t}} \alpha \frac{2\alpha y}{\beta} e^{\frac{\alpha^2 t}{\beta^2}} \operatorname{erfc}\left(\frac{y}{\sqrt{t}} + \frac{\alpha}{\beta}\sqrt{t}\right) dy.$$

We make a substitution by setting

$$u = \frac{y}{\sqrt{t}},$$

so we obtain

$$-\frac{\alpha}{\sqrt{\pi}\beta} e^{\frac{\alpha^2 t}{\beta^2}} \int_0^{R/\sqrt{t}} e^{2\alpha u\sqrt{t}/\beta} \sqrt{R^2 - t^2 u^2} \operatorname{erfc}\left(u + \frac{\alpha}{\beta}\sqrt{t}\right) du.$$

We shall use integration by parts, noting that

$$\frac{d}{dz} \left(z \operatorname{erfc}(z) - \frac{e^{-z^2}}{\sqrt{\pi}} \right) = \operatorname{erfc}(z).$$

So,

$$\begin{aligned} & \int_0^{R/\sqrt{t}} e^{2\alpha u\sqrt{t}/\beta} \sqrt{R^2 - t^2 u^2} \operatorname{erfc}\left(u + \frac{\alpha}{\beta}\sqrt{t}\right) du \\ &= e^{2\alpha\sqrt{t}u/\beta} \sqrt{R^2 - t^2 u^2} \left[\left(u + \frac{\alpha\sqrt{t}}{\beta}\right) \operatorname{erfc}\left(u + \frac{\alpha\sqrt{t}}{\beta}\right) - \frac{e^{-(u+\alpha\sqrt{t}/\beta)^2}}{\sqrt{\pi}} \right]_{u=0}^{R/\sqrt{t}} \\ & - \int_0^{R/\sqrt{t}} \left(\frac{2\alpha\sqrt{t}}{\beta} - \frac{t(tu)}{R^2 - t^2 u^2} \right) e^{2\alpha\sqrt{t}u/\beta} \sqrt{R^2 - t^2 u^2} \operatorname{erfc}\left(u + \frac{\alpha\sqrt{t}}{\beta}\right) du. \end{aligned}$$

It is a straightforward exercise to prove that

$$\int_0^{R/\sqrt{t}} e^{2\alpha\sqrt{t}u/\beta} \operatorname{erfc}\left(u + \frac{\alpha\sqrt{t}}{\beta}\right) du$$

is uniformly bounded as $t \downarrow 0$. Hence the second term is $O(\sqrt{t})$ as $t \downarrow 0$, for we can pull out a factor of \sqrt{t} and keep every other term in the integrand bounded above by a constant. However, as $t \downarrow 0$, the first term converges to $R\pi^{-1/2}$. Hence, the E

term gives a contribution of

$$-\frac{\alpha R}{\pi\beta} = -\frac{\alpha}{2\pi\beta}\ell(\mathcal{N}_i \cap \partial\Omega).$$

This is the usual perimeter term in the Robin setting [35].

The Robin heat trace asymptotics also have a contribution from the corners. Though we cannot calculate it explicitly here, we do so using geometric microlocal analysis in [23]. We obtain that the ‘‘corner contribution’’ from an interior angle γ is the same as in the Neumann case,

$$\frac{\pi^2 - \gamma^2}{24\pi\gamma}.$$

3.3 Heat Trace Expansions and Proof of Theorem 1

It is now straightforward to use (27) to compute the heat trace asymptotics for Ω by combining our explicit computations of the integrals over \mathcal{N}_i with the known asymptotics [21, 35] for the integrals over $\Omega \setminus \{\mathcal{N}_i\}$. In addition to the ‘‘corner contribution’’ from the parametrix at the corner, there is also a contribution from the model used for the smooth parts of the domain, as turning the corner contributes to the curvature:

$$-\frac{\pi - \theta_k}{12\pi}, \text{ for an interior angle } \theta_k.$$

We therefore obtain:

(N) for the Neumann boundary condition,

$$\text{tr}e^{-t\Delta} \sim \frac{|\Omega|}{4\pi t} + \frac{|\partial\Omega|}{8\sqrt{\pi t}} + \frac{\chi(\Omega)}{6} - \frac{n}{12} + \sum_{k=1}^n \frac{\pi^2 + \theta_k^2}{24\pi\theta_k} + O(\sqrt{t}),$$

(R) for the Robin boundary condition,

$$\text{tr}e^{-t\Delta} \sim \frac{|\Omega|}{4\pi t} + \frac{|\partial\Omega|}{8\sqrt{\pi t}} + \frac{\chi(\Omega)}{6} - \frac{|\partial\Omega|\alpha}{2\pi\beta} - \frac{n}{12} + \sum_{k=1}^n \frac{\pi^2 + \theta_k^2}{24\pi\theta_k} + O(\sqrt{t}).$$

Now let $\tilde{\Omega}$ be a smoothly bounded domain in the plane. The heat trace expansions have been computed by McKean and Singer [21] for the Neumann boundary condition and [35] for the Robin condition. These are, respectively,

(N) for the Neumann boundary condition,

$$\text{tr}e^{-t\Delta} \sim \frac{|\tilde{\Omega}|}{4\pi t} + \frac{|\partial\tilde{\Omega}|}{8\sqrt{\pi t}} + \frac{\chi(\tilde{\Omega})}{6} + O(\sqrt{t}),$$

(R) for the Robin boundary condition,

$$\text{tr}e^{-t\Delta} \sim \frac{|\tilde{\Omega}|}{4\pi t} + \frac{|\partial\tilde{\Omega}|}{8\sqrt{\pi t}} + \frac{\chi(\tilde{\Omega})}{6} - \frac{|\partial\tilde{\Omega}|\alpha}{2\pi\beta} + O(\sqrt{t}).$$

Proof (Theorem 1) If two domains are isospectral, then they have the same heat trace. Hence, for each power of t in such an expansion, the coefficient must be identical for both domains. Now, let us assume that Ω satisfies the assumptions in Theorem 1, so it has at least one corner of interior angle not equal to π . Let the interior angles at the corners be $\{\theta_k\}_{k=1}^n$. Let us assume that $\tilde{\Omega}$ is a smoothly bounded domain, and that we have taken the same boundary condition for the Laplacian for both Ω and $\tilde{\Omega}$. Assume for the sake of contradiction that Ω and $\tilde{\Omega}$ are isospectral. Therefore, their heat trace coefficients coincide. Hence, they have the same area and perimeter. Since the same boundary condition is taken for both domains, and thus the same values of α and β in the Robin case, we should have

$$\frac{\chi(\Omega)}{6} - \frac{n}{12} + \sum_{k=1}^n \frac{\pi^2 + \theta_k^2}{24\pi\theta_k} = \frac{\chi(\tilde{\Omega})}{6}. \tag{36}$$

We have assumed that Ω is simply connected, but we make no such assumption on $\tilde{\Omega}$. Hence

$$\chi(\Omega) = 1, \quad \chi(\tilde{\Omega}) \leq 1.$$

Following the argument on p. 91–92 of [16],

$$\frac{\chi(\Omega)}{6} - \frac{n}{12} + \sum_{k=1}^n \frac{\pi^2 + \theta_k^2}{24\pi\theta_k} > \frac{1}{6} \geq \frac{\chi(\tilde{\Omega})}{6},$$

which violates (36). □

4 Microlocal Analysis in the Curvilinear Case

It turns out that the heat trace expansions above are also valid for curvilinear polygons, once terms accounting for the curvature of the boundary away from the corners have been included. Although this has been demonstrated in [16] for the

Dirichlet boundary condition using monotonicity, it becomes a much more subtle matter for the Neumann and Robin boundary conditions.

The main problem is that for curvilinear polygons, we no longer have an exact geometric match. Hence, we can no longer use the locality principle to compute the heat trace expansion, because there are no known expressions for the heat kernels. For classical polygons, one may compute the Neumann heat trace using the Dirichlet heat trace together with the trace of a Euclidean surface with conical singularities created by doubling the polygon. However, this technique fails once the edges of the polygon are no longer necessarily straight near the corners. Therefore, in order to compute the short time asymptotic expansion of the heat trace without exact geometric matches, we turn to the robust techniques of geometric microlocal analysis. This allows us to give a full description of the Dirichlet, Neumann, and Robin heat kernels on a curvilinear polygon in all asymptotic regimes. Restricting to the diagonal and integrating yields the heat trace.

In order to describe the heat kernel in all asymptotic regimes, we build a space, called the *heat space* or *double heat space*, on which the heat kernel is well-behaved. This space is built by blowing up various p -submanifolds of $\Omega \times \Omega \times [0, \infty)$. To see why this is needed, first consider the heat kernel (1) on \mathbb{R}^n . At the diagonal in $\mathbb{R}^n \times \mathbb{R}^n \times [0, \infty)$, the heat kernel behaves as $O(t^{-n/2})$ as $t \downarrow 0$. However, as long as $d(z, z') \geq \varepsilon > 0$, the heat kernel behaves as $O(t^\infty)$ as $t \downarrow 0$. So the heat kernel fails to be well-behaved at $\{z = z', t = 0\}$. This is the motivation for “blowing up” the diagonal $\{z = z'\}$ at $t = 0$, which means replacing this diagonal with its inward pointing spherical normal bundle, corresponding to the introduction of “polar coordinates”. The precise meaning of “blowing up” is explained in [20], and in this particular case of blowing up $\{z = z'\}$ at $t = 0$ in $\mathbb{R}^n \times \mathbb{R}^n \times [0, \infty)$, see [20, Chapter 7].

For the case of a curvilinear polygonal domain $\Omega \subset \mathbb{R}^2$, we begin with $\Omega \times \Omega \times [0, \infty)$ and perform a sequence of blow-ups. Our construction is inspired by the construction of the heat kernel on manifolds with wedge singularities performed by Mazzeo and Vertman in [19]. We leave the details to our forthcoming work [23].

Once the double heat space has been constructed, the heat kernel may be built in a similar spirit to the Duhamel’s principle construction of the Robin heat kernel in the proof of Theorem 5. We start with a parametrix, or initial guess, and then use Duhamel’s principle to iterate away the error. This requires the proof of a composition result for operators whose kernels are well-behaved on our double heat space, and that in turn requires some fairly involved technical machinery (a proof “by hand” without using this machinery would be entirely unreadable). However, it works out and gives us a very precise description of the heat kernel on a curvilinear polygon, with any combination of Dirichlet, Neumann, and Robin conditions. Moreover, we are able to generalize our techniques and results to surfaces which have boundary, edges, corners, and conical singularities.

The details of this sort of geometric microlocal analysis construction are intricate, but its utility is undeniable. In settings such as this, where exact geometric matches are lacking, but instead, one has *asymptotic geometric matches*, these microlocal techniques may be helpful. For the full story in the case of curvilinear polygons

and their heat kernels, please stay tuned for our forthcoming work [23]. We have seen here that the heat kernels for circular sectors gives the same angular contribution, arising from the so-called “C term” for both Neumann and Robin boundary conditions. Moreover, this is the same in the Dirichlet case as well [29]. Interestingly, it appears that for mixed boundary conditions, there is a sudden change in this corner contribution. We are in the process of obtaining a small collection of negative isospectrality results in these general settings in the spirit of Theorem 1, including a generalization of Theorem 1 which removes the hypothesis that the corners are exact; see [23] for the full story.

Acknowledgements The authors extend their sincere gratitude to the organizers of the Matrix workshop, “Elliptic Partial Differential Equations of Second Order: Celebrating 40 Years of Gilbarg and Trudinger’s Book.” Many thanks to Lucio Boccardo, Florica-Corina Cîrstea, Julie Clutterbuck, L. Craig Evans, Enrico Valdinoci, Paul Bryan, and of course, Neil Trudinger! We also gratefully acknowledge Lashi Bandara, Liangpan Li, Grigori Rozenblum, Alexander Strohmaier, and Feng-Yu Wang for helpful and insightful correspondences. The first author was partially supported by the Ministry of Education Science of the Republic of Kazakhstan under grant AP05132071. The second author is supported by the Swedish Research Council Grant, 2018-03873 (GAAME). The third author was partially supported by a grant from the College of Science and Health at DePaul University.

References

1. Aldana, C., Rowlett, J.: A Polyakov formula for sectors. *J. Geom. Anal.* (2017) <https://doi.org/10.1007/s12220-017-9888-y>
2. Bondurant, J.D., Fulling, S.A.: The Dirichlet-to-Robin Transform (2004, Preprint). arXiv:math-ph/0408054v1
3. Carslaw, H.S., Jaeger, J.C.: *Conduction of Heat in Solids*, 2nd edn. Oxford, Clarendon (1959)
4. Daners, D.: Heat kernel estimates for operators with boundary conditions. *Mathematische Nachrichten* **217**(1), 13–41 (2000)
5. Durso, C.: On the inverse spectral problem for polygonal domains, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge (1988)
6. Erdelyi, A., Magnus, W., Oberhettinger, F., Tricomi, F.G.: *Tables of Integral Transforms*. McGraw-Hill, New York (1954)
7. Fedosov, B.: Asymptotic formulas for eigenvalues of the Laplacian in a polyhedron. *Doklady Akad. Nauk SSSR* **157**, 536–538 (1964)
8. Gordon, C., Webb, D.L., Wolpert, S.: One cannot hear the shape of a drum. *Bull. Am. Math. Soc. (NS)* **27**(1), 134–138 (1992)
9. Gordon, C., Webb, D.L., Wolpert, S.: Isospectral plane domains and surfaces via Riemannian orbifolds. *Invent. Math.* **110**(1), 1–22 (1992)
10. Gradshteyn, I.S., Ryzhik, I.M.: *Table of Integrals, Series, and Products*, 8th edn. Elsevier/Academic Press, Amsterdam (2015)
11. Grieser, D., Maronna, S.: Hearing the shape of a triangle. *Notices Am. Math. Soc.* **60**(11), 1440–1447 (2013)
12. Hezari, H., Lu, Z., Rowlett, J.: The Neumann Isospectral Problem for Trapezoids. *Ann. Henri Poincaré* **18**(12), 3759–3792 (2017)
13. Kac, M.: Can one hear the shape of a drum? *Am. Math. Mon.* **73**(1), 1–23 (1966)
14. Li, L., Strohmaier, A.: Heat kernel estimates for general boundary problems. *J. Spectr. Theory* **6**, 903–919 (2016). arxiv:1604.00784v1

15. Lu, Z., Rowlett, J.: The sound of symmetry. *Am. Math. Mon.* **122**(9), 815–835 (2015)
16. Lu, Z., Rowlett, J.: One can hear the corners of a drum. *Bull. Lond. Math. Soc.* **48**(1), 85–93 (2016)
17. Lück, W., Schick, T.: L^2 -torsion of hyperbolic manifolds of finite volume. *Geom. Funct. Anal.* **9**(3), 518–567 (1999)
18. Mazzeo, R., Rowlett, J.: A heat trace anomaly on polygons. *Math. Proc. Camb. Philos. Soc.* **159**(02), 303–319 (2015)
19. Mazzeo, R., Vertman, B.: Analytic torsion on manifolds with edges. *Adv. Math.* **231**(2), 1000–1040 (2012)
20. Melrose, R.: The Atiyah-Patodi-Singer Index Theorem, *Research Notes in Mathematics*, vol. 4. A K Peters, Wellesley (1993)
21. McKean Jr., H.P., Singer, I.M.: Curvature and the eigenvalues of the Laplacian. *J. Differ. Geom.* **1**(1), 43–69 (1967)
22. Milnor, J.: Eigenvalues of the Laplace operator on certain manifolds. *Proc. Natl. Acad. Sci. USA* **51**, 542 (1964)
23. Nursultanov, M., Rowlett, J., Sher, D.: The heat kernel and geometric spectral invariants on surfaces with corners. Preprint
24. Papanicolaou, V.G.: The probabilistic solution of the third boundary value problem for second order elliptic equations. *Probab. Theory Relat. Fields* **87**(1), 27–77 (1990)
25. Pleijel, Å.: A study of certain Green's functions with applications in the theory of vibrating membranes. *Ark. Mat.* **2**, 553–569 (1954)
26. Sher, D.: Conic degeneration and the determinant of the Laplacian. *J. Anal. Math.* **126**(2), 175–226 (2015)
27. Sunada, T.: Riemannian coverings and isospectral manifolds. *Ann. Math. (2)* **121**(1), 169–186 (1985)
28. Taylor, M.: *Partial Differential Equations I, Basic Theory*. Applied Mathematical Sciences, vol. 115, 2nd edn. Springer, New York (2011)
29. van den Berg, M., Srisatkunarajah, S.: Heat equation for a region in \mathbb{R}^2 with polygonal boundary. *J. Lond. Math. Soc. (2)* **2**(1), 119–127 (1988)
30. Wang, F.Y., Yan, L.: Gradient estimate on the Neumann semigroup and applications, arxiv:1009.1965v2
31. Wang, F.Y.: Gradient estimates and the first Neumann eigenvalue on manifolds with boundary. *Stoch. Process. Appl.* **115**, 1475–1486 (2005)
32. Wang, F.Y.: Semigroup properties for the second fundamental form. *Doc. Math.* **15**, 543–559 (2010)
33. Wang, F.Y.: Gradient and Harnack inequalities on noncompact manifolds with boundary. *Pac. J. Math.* **245**, 185–200 (2010)
34. Watson, G.N.: *A Treatise on the Theory of Bessel Functions*. Cambridge Mathematical Library. Cambridge University Press, Cambridge (1995). Reprint of the second (1944) edition
35. Zayed, E.M.: Short time asymptotics of the heat kernel of the Laplacian of a bounded domain with Robin boundary conditions. *Houst. J. Math.* **24**(2), 377–385 (1998)
36. Zelditch, S.: Spectral determination of analytic bi-axisymmetric plane domains. *Geom. Funct. Anal.* **10**(3), 628–677 (2000)

Nonparametric Bayesian Volatility Estimation



Shota Gugushvili, Frank van der Meulen, Moritz Schauer, and Peter Spreij

Abstract Given discrete time observations over a fixed time interval, we study a nonparametric Bayesian approach to estimation of the volatility coefficient of a stochastic differential equation. We postulate a histogram-type prior on the volatility with piecewise constant realisations on bins forming a partition of the time interval. The values on the bins are assigned an inverse Gamma Markov chain (IGMC) prior. Posterior inference is straightforward to implement via Gibbs sampling, as the full conditional distributions are available explicitly and turn out to be inverse Gamma. We also discuss in detail the hyperparameter selection for our method. Our nonparametric Bayesian approach leads to good practical results in representative simulation examples. Finally, we apply it on a classical data set in change-point analysis: weekly closings of the Dow-Jones industrial averages.

S. Gugushvili · M. Schauer
Mathematical Institute, Leiden University, Leiden, The Netherlands
e-mail: shota.gugushvili@math.leidenuniv.nl; m.r.schauer@math.leidenuniv.nl

F. van der Meulen
Delft Institute of Applied Mathematics, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, The Netherlands
e-mail: f.h.vandermeulen@tudelft.nl

P. Spreij (✉)
Korteweg-de Vries Institute for Mathematics, University of Amsterdam, Amsterdam, The Netherlands

Institute for Mathematics, Astrophysics and Particle Physics, Radboud University, Nijmegen, The Netherlands
e-mail: spreij@uva.nl

1 Introduction

1.1 Problem Formulation

Consider a one-dimensional stochastic differential equation (SDE)

$$dX_t = b_0(t, X_t) dt + s_0(t) dW_t, \quad X_0 = x, \quad t \in [0, T], \quad (1)$$

where b_0 is the drift coefficient, s_0 the deterministic dispersion coefficient or volatility, and x is a deterministic initial condition. Here W is a standard Brownian motion. Assume that standard conditions for existence and uniqueness of a strong solution to (1) are satisfied (see, e.g., [47]), and observations

$$\mathcal{X}_n = \{X_{t_{0,n}}, \dots, X_{t_{n,n}}\}$$

are available, where $t_{i,n} = iT/n$, $i = 0, \dots, n$. Using a nonparametric Bayesian approach, our aim is to estimate the volatility function s_0 . In a financial context, knowledge of the volatility is of fundamental importance e.g. in pricing financial derivatives; see [4] and [52]. However, SDEs have applications far beyond the financial context as well, e.g. in physics, biology, life sciences, neuroscience and engineering (see [1, 25, 40] and [76]). Note that by Itô's formula, using a simple transformation of the state variable, also an SDE of the form

$$dX_t = b_0(t, X_t) dt + s_0(t) f_0(X_t) dW_t, \quad X_0 = x, \quad t \in [0, T],$$

can be reduced to the form (1), provided the function f_0 is known and regular enough; see, e.g., p. 186 in [66]. Some classical examples that fall under our statistical framework are the geometric Brownian motion and the Ornstein-Uhlenbeck process. Note also that as we allow the drift in (1) to be non-linear, marginal distributions of X are not necessarily Gaussian and may thus exhibit heavy tails, which is attractive in financial modelling.

A nonparametric approach guards one against model misspecification and is an excellent tool for a preliminary, exploratory data analysis, see, e.g., [65]. Commonly acknowledged advantages of a Bayesian approach include automatic uncertainty quantification in parameter estimates via Bayesian credible sets, and the fact that it is a fundamentally likelihood-based method. In [51] it has been argued that a nonparametric Bayesian approach is important for honest representation of uncertainties in inferential conclusions. Furthermore, use of a prior allows one to easily incorporate the available external, a priori information into the estimation procedure, which is not straightforward to achieve with frequentist approaches. For instance, this a priori information could be an increasing or decreasing trend in the volatility.

1.2 Literature Overview

Literature on nonparametric Bayesian volatility estimation in SDE models is scarce. We can list theoretical contributions [34, 36, 54], and the practically oriented paper [2]. The model in the former two papers is close to the one considered in the present work, but from the methodological point of view different Bayesian priors are used and practical usefulness of the corresponding Bayesian approaches is limited. On the other hand, the models considered in [54] and [2] are rather different from ours, and so are the corresponding Bayesian approaches. The nearest predecessor of the model and the method in our paper is the one studied in [37]. In the sequel we will explain in what aspects the present contribution differs from that one and what the current improvements are. We note in passing that there exists a solid body of literature on nonparametric Bayesian estimation of the drift coefficient, see, e.g., [35, 55, 57, 63, 70, 71] and the review article [72], but Bayesian volatility estimation requires use of substantially different ideas. We also note existence of works dealing with parametric Bayesian estimation in discrete-time stochastic volatility models, see, e.g., [45] and [46], but again, these are not directly related to the problem we study in this paper.

1.3 Approach and Results

The main potential difficulties facing a Bayesian approach to inference in SDE models from discrete observations are an intractable likelihood and absence of a closed form expression for the posterior distribution; see, e.g., [21, 25, 62] and [69]. Typically, these difficulties necessitate the use of a data augmentation device (see [67]) and some intricate form of a Markov chain Monte Carlo (MCMC) sampler (see [61]). In [37], these difficulties are circumvented by intentionally setting the drift coefficient to zero, and employing a (conjugate) histogram-type prior on the diffusion coefficient, that has piecewise constant realisations on bins forming a partition of $[0, T]$. Specifically, the (squared) volatility is modelled a priori as a function $s^2 = \sum_{k=1}^N \theta_k \mathbf{1}_{B_k}$, with independent and identically distributed inverse gamma coefficients θ_k 's, and the prior Π is defined as the law of s^2 . Here B_1, \dots, B_N are bins forming a partition of $[0, T]$. With this independent inverse Gamma (IIG) prior, $\theta_1, \dots, \theta_N$ are independent, conditional on the data, and of inverse gamma type. Therefore, this approach results in a fast and simple to understand and implement Bayesian procedure. A study of its favourable practical performance, as well as its theoretical validation was recently undertaken in [37]. As shown there under precise regularity conditions, misspecification of the drift is asymptotically, as the sample size $n \rightarrow \infty$, harmless for consistent estimation of the volatility coefficient.

Despite a good practical performance of the method in [37], there are some limitations associated with it too. Thus, the method offers limited possibilities for adaptation to the local structure of the volatility coefficient, which may become an issue if the volatility has a wildly varying curvature on the time interval $[0, T]$. A possible fix to this would be to equip the number of bins N forming a partition of $[0, T]$ with a prior, and choose the endpoints of bins B_k also according to a prior. However, this would force one to go beyond the conjugate Bayesian setting as in [37], and posterior inference in practice would require, for instance, the use of a reversible jump MCMC algorithm (see [32]). Even in the incomparably simpler setting of intensity function estimation for nonhomogeneous Poisson processes with histogram-type priors, this is very challenging, as observed in [77]. Principal difficulties include designing moves between models of differing dimensions that result in MCMC algorithms that *mix well*, and *assessment of convergence* of Markov chains (see [22], p. 204). Thus, e.g., the inferential conclusions in [32] and [33] are different on the same real data example using the same reversible jump method, since it turned out that in the first paper the chain was not run long enough. Cf. also the remarks on Bayesian histograms in [27], p. 546.

Here we propose an alternative approach, inspired by ideas in [7] in the context of audio signal modelling different from the SDE setting that we consider; see also [8, 9, 15, 16] and [73]. Namely, instead of using a prior on the (squared) volatility that has piecewise constant realisations on $[0, T]$ with independent coefficients θ_k 's, we will assume that the sequence $\{\theta_k\}$ forms a suitably defined Markov chain. An immediately apparent advantage of using such an approach is that it induces extra smoothing via dependence in prior realisations of the volatility function across different bins. Arguing heuristically, with a large number N of bins B_k it is then possible to closely mimic the local structure of the volatility: in those parts of the interval $[0, T]$, where the volatility has a high curvature or is subject to abrupt changes, a large number of (narrow) bins is required to adequately capture these features. However, the grid used to define the bins B_k 's is uniform, and if $\theta_1, \dots, \theta_N$ are a priori independent, a large N may induce spurious variability in the volatility estimates in those regions of $[0, T]$ where the volatility in fact varies slowly. As we will see in the sequel, this problem may be alleviated using a priori dependent θ_k 's.

In the subsequent sections we detail our approach, and study its practical performance via simulation and real data examples. Specifically, we implement our method via a straightforward version of the Gibbs sampler, employing the fact that full conditional distributions of θ_k 's are known in closed form (and are in fact inverse gamma). Unlike [37], posterior inference in our new approach requires the use of MCMC. However, this is offset by the advantages of our new approach outlined above, and in fact the additional computational complexity of our new method is modest in comparison to [37]. The prior in our new method depends on hyperparameters, and we will also discuss several ways of their choice in practice.

1.4 Organisation of This Paper

In Sect. 2 we supply a detailed description of our nonparametric Bayesian approach to volatility estimation. In Sect. 3 we study the performance of our method via extensive simulation examples. In Sect. 4 we apply the method on a real data example. Section 5 summarises our findings and provides an outlook on our results. Finally, Sect. 6 contains some additional technical details of our procedure.

1.5 Notation

We denote the prior distribution on the (squared) volatility function by Π and write the posterior measure given data \mathcal{X}_n as $\Pi(\cdot \mid \mathcal{X}_n)$. We use the notation $\text{IG}(\alpha, \beta)$ for the inverse gamma distribution with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$. This distribution has a density

$$x \mapsto \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x}, \quad x > 0. \quad (2)$$

For two sequences $\{a_n\}, \{b_n\}$, the notation $a_n \asymp b_n$ will be used to denote the fact that the sequences are asymptotically (as $n \rightarrow \infty$) of the same order. Finally, for a density f and a function g , the notation $f \propto g$ will mean that f is proportional to g , with proportionality constant on the righthand side recovered as $(\int g)^{-1}$, where the integral is over the domain of definition of g (and of f). The function g can be referred to as an unnormalised probability density.

2 Nonparametric Bayesian Approach

2.1 Generalities

Our starting point is the same as in [37]. Namely, we misspecify the drift coefficient b_0 by intentionally setting it to zero (see also [49] for a similar idea of ‘misspecification on purpose’). The theoretical justification for this under the ‘infill’ asymptotics, with the time horizon T staying fixed and the observation times $t_{i,n} = iT/n$, $i = 1, \dots, n$, filling up the interval $[0, T]$ as $n \rightarrow \infty$, is provided in [37], to which we refer for further details (the argument there ultimately relies on Girsanov’s theorem). Similar ideas are also encountered in the non-Bayesian setting in the econometrics literature on high-frequency financial data, see, e.g., [53].

Set $Y_{i,n} = X_{t_{i,n}} - X_{t_{i-1,n}}$. With the assumption $b_0 = 0$, the pseudo-likelihood of our observations is tractable, in fact Gaussian,

$$L_n(s^2) = \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi \int_{t_{i-1,n}}^{t_{i,n}} s^2(u) du}} \psi \left(\frac{Y_{i,n}}{\sqrt{\int_{t_{i-1,n}}^{t_{i,n}} s^2(u) du}} \right) \right\}, \tag{3}$$

where $\psi(u) = \exp(-u^2/2)$. The posterior probability of any measurable set S of volatility functions can be computed via Bayes' theorem as

$$\Pi(S | \mathcal{X}_n) = \frac{\int_S L_n(s^2) \Pi(ds)}{\int L_n(s^2) \Pi(ds)}.$$

Here the denominator is the normalising constant, the integral over the whole space on which the prior Π is defined, which ensures that the posterior is a *probability* measure (i.e. integrates to one).

2.2 Prior Construction

Our prior Π is constructed similarly to [37], with an important difference to be noted below. Fix an integer $m < n$. Then $n = mN + r$ with $0 \leq r < m$, where $N = \lfloor \frac{n}{m} \rfloor$. Now define bins $B_k = [t_{m(k-1),n}, t_{mk,n}]$, $k = 1, \dots, N - 1$, and $B_N = [t_{m(N-1),n}, T]$. Thus the first $N - 1$ bins are of length mT/n , whereas the last bin B_N has length $T - t_{m(N-1),n} = n^{-1}(r + m)T < n^{-1}2mT$. The parameter N (equivalently, m) is a hyperparameter of our prior. We model s as piecewise constant on bins B_k , thus $s = \sum_{k=1}^N \xi_k \mathbf{1}_{B_k}$. The prior Π on the volatility s can now be defined by assigning a prior to the coefficients ξ_k 's.

Let $\theta_k = \xi_k^2$. Since the bins B_k are disjoint,

$$s^2 = \sum_{k=1}^N \xi_k^2 \mathbf{1}_{B_k} = \sum_{k=1}^N \theta_k \mathbf{1}_{B_k}.$$

As the likelihood depends on s only through its square s^2 , it suffices to assign the prior to the coefficients θ_k 's of s^2 . This is the point where we fundamentally diverge from [37]. Whereas in [37] it is assumed that $\{\theta_k\}$ is an i.i.d. sequence of inverse gamma random variables, here we suppose that $\{\theta_k\}$ forms a Markov chain. This will be referred to as an inverse Gamma Markov chain (IGMC) prior (see [7]), and is defined as follows. Introduce auxiliary variables ζ_k , $k = 2, \dots, N$, and define a Markov chain using the time ordering $\theta_1, \zeta_2, \theta_2, \dots, \zeta_k, \theta_k, \dots, \zeta_N, \theta_N$. Transition

distributions of this chain are defined as follows: fix hyperparameters α_1, α_ζ and α , and set

$$\theta_1 \sim \text{IG}(\alpha_1, \alpha_1), \quad \zeta_{k+1} | \theta_k \sim \text{IG}(\alpha_\zeta, \alpha_\zeta \theta_k^{-1}), \quad \theta_{k+1} | \zeta_{k+1} \sim \text{IG}(\alpha, \alpha \zeta_{k+1}^{-1}). \quad (4)$$

The name of the chain reflects the fact that these distributions are inverse Gamma.

Remark 1 Our definition of the IGMC prior differs from the one in [7] in the choice of the initial distribution of θ_1 , which is important to alleviate possible ‘edge effects’ in volatility estimates in a neighbourhood of $t = 0$. The parameter α_1 determines the initial distribution of the inverse Gamma Markov chain. Letting $\alpha_1 \rightarrow 0$ (which corresponds to a vague prior) ‘releases’ the chain at the time origin. \square

Remark 2 As observed in [7], there are various ways of defining an inverse Gamma Markov chain. The point to be kept in mind is that the resulting posterior should be computationally tractable, and the prior on θ_k ’s should have a capability of producing realisations with positive correlation structures, as this introduces smoothing among the θ_k ’s in adjacent bins. This latter property is not possible to attain with arbitrary constructions of inverse Gamma Markov chains, such as e.g. a natural construction $\theta_k | \theta_{k-1} \sim \text{IG}(\alpha, \theta_{k-1} / \alpha)$. On the other hand, positive correlation between realisations θ_k ’s can be achieved e.g. by setting $\theta_k | \theta_{k-1} \sim \text{IG}(\alpha, (\alpha \theta_{k-1})^{-1})$, but this results in intractable posterior computations. The definition of the IGMC prior in the present work, that employs latent variables ζ_k ’s, takes care of both these important points. For an additional discussion see [7]. \square

Remark 3 Setting the drift coefficient b_0 to zero effectively results in pretending that the process X has independent (Gaussian) increments. In reality, since the drift in practical applications is typically nonzero, increments of the process are dependent, and hence all observations $Y_{i,n}$ contain some indirect information on the value of the volatility s^2 at each time point $t \in [0, T]$. On the other hand, assuming the IGMC prior on s^2 yields a posteriori dependence of coefficients $\{\theta_k\}$, which should be of help in inference with smaller sample sizes n . See Sect. 4 for an illustration. \square

2.3 Gibbs Sampler

It can be verified by direct computations employing (4) that the full conditional distributions of θ_k ’s and ζ_k ’s are inverse gamma,

$$\theta_k | \zeta_k, \zeta_{k+1} \sim \text{IG} \left(\alpha + \alpha_\zeta, \frac{\alpha}{\zeta_k} + \frac{\alpha_\zeta}{\zeta_{k+1}} \right), \quad k = 2, \dots, N - 1, \quad (5)$$

$$\theta_1 | \zeta_2 \sim \text{IG} \left(\alpha_1 + \alpha_\zeta, \alpha_1 + \frac{\alpha_\zeta}{\zeta_2} \right), \quad (6)$$

$$\theta_N | \zeta_N \sim \text{IG} \left(\alpha, \frac{\alpha}{\zeta_N} \right), \tag{7}$$

$$\zeta_k | \theta_k, \theta_{k-1} \sim \text{IG} \left(\alpha_\zeta + \alpha, \frac{\alpha_\zeta}{\theta_{k-1}} + \frac{\alpha}{\theta_k} \right), \quad k = 2, \dots, N. \tag{8}$$

See Sect. 6 for details. Next, the effective transition kernel of the Markov chain $\{\theta_k\}$ is given by formula (4) in [7], and is a scale mixture of inverse gamma distributions; however, its exact expression is of no direct concern for our purposes. As noted in [7], p. 700, depending on the parameter values α, α_ζ , it is possible for the chain $\{\theta_k\}$ to exhibit either an increasing or decreasing trend. We illustrate this point by plotting realisations of $\{\theta_k\}$ in Fig. 1 for different values of α and α_ζ . In the context of volatility estimation this feature is attractive, if prior information on the volatility trend is available.

Inference in [7] is performed using a mean-field variational Bayes approach, see, e.g., [5]. Here we describe instead a fully Bayesian approach relying on Gibbs sampling (see, e.g., [26] and [29]), cf. [9].

The algorithm is initialised at values ζ_2, \dots, ζ_N , e.g. generated from the prior specification (4). In order to derive update formulae for the full conditionals of the θ_k 's, define

$$Z_k = \sum_{i=(k-1)m+1}^{km} Y_{i,n}^2, \quad k = 1, \dots, N - 1,$$

$$Z_N = \sum_{i=(N-1)m+1}^n Y_{i,n}^2.$$

With this notation, the likelihood from (3) satisfies

$$L_n(\theta) \propto \theta_N^{-(m+r)/2} \exp \left(-\frac{nZ_N}{2T\theta_N} \right) \prod_{k=1}^{N-1} \theta_k^{-m/2} \exp \left(-\frac{nZ_k}{2T\theta_k} \right).$$

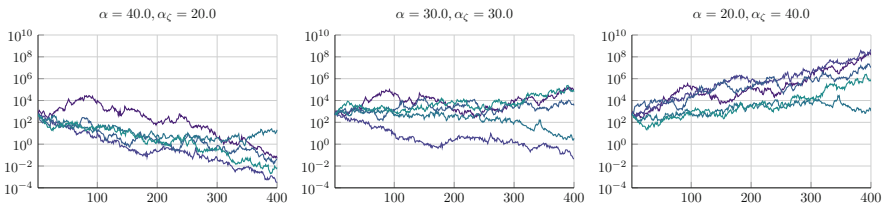


Fig. 1 Realisations of the Markov chain $\{\theta_k\}$ with $\alpha = 40, \alpha_\zeta = 20$ (left panel) and $\alpha = 30, \alpha_\zeta = 30$ (center panel) and $\alpha = 20, \alpha_\zeta = 40$ (right panel). In all cases, θ_1 is fixed to 500

Using this formula and Eq.(5), and recalling the form of the inverse gamma density (2), it is seen that the update distribution for θ_k , $k = 2, \dots, N - 1$, is

$$\text{IG} \left(\alpha + \alpha_\zeta + \frac{m}{2}, \frac{\alpha}{\zeta_k} + \frac{\alpha_\zeta}{\zeta_{k+1}} + \frac{nZ_k}{2T} \right),$$

whereas by (7) the ones for θ_1 and θ_N are

$$\text{IG} \left(\alpha_1 + \alpha_\zeta + \frac{m}{2}, \alpha_1 + \frac{\alpha_\zeta}{\zeta_2} + \frac{nZ_1}{2T} \right), \quad \text{IG} \left(\alpha + \frac{m+r}{2}, \frac{\alpha}{\zeta_N} + \frac{nZ_N}{2T} \right),$$

respectively.

Next, the latent variables ζ_k 's will be updated using formula (8). This update step for ζ_k 's does not directly involve the data \mathcal{X}_n , except through the previous values of θ_k 's.

Finally, one iterates these two Gibbs steps for θ_k 's and ζ_k 's a large number of times (until chains can be assessed as reasonably converged), which gives posterior samples of the θ_k 's. Using the latter, the posterior inference can proceed in the usual way, e.g. by computing the sample posterior mean of θ_k 's, as well as sample quantiles, that provide, respectively, a point estimate and uncertainty quantification via marginal Bayesian credible bands for the squared volatility s^2 . Similar calculations on the square roots of the posterior samples can be used to obtain point estimates and credible bands for the volatility function s itself.

2.4 Hyperparameter Choice

We first assume the number of bins N has been chosen in some way, and we only have to deal with hyperparameters α , α_ζ and α_1 , that govern properties of the Markov chain prior. In [7], where an IGMC prior was introduced, guidance on the hyperparameter selection is not discussed. In [8], the hyperparameters are fine-tuned by hand in specific problems studied there (audio denoising and single channel audio source separation). Another practical solution is to try several different fixed combinations of the hyperparameters α , α_ζ and α_1 , if only to verify sensitivity of inferential conclusions with respect to variations in the hyperparameters. Some further methods for hyperparameter optimisation are discussed in [16]. In [8] optimisation of the hyperparameters via the maximum likelihood method is suggested; practical implementation relies on the EM algorithm (see [13]), and some additional details are given in [15]. Put in other terms, the proposal in [15] amounts to using an empirical Bayes method (see, e.g., [20], [59] and [60]). The use of the latter is widespread and often leads to good practical results, but the method is still insufficiently understood theoretically, except in toy models like the white noise model (see, however, [17] and [56] for some results in other contexts). On the practical side, in our case, given that the dimension of the sequences $\{\zeta_k\}$ and $\{\theta_k\}$ is rather high, namely $2N - 1$ with N large, and the marginal likelihood is not available

in closed form, this approach is expected to be computationally intensive. Therefore, a priori there is no reason not to try instead a fully Bayesian approach by equipping the hyperparameters with a prior, and this is in fact our default approach in the present work. However, the corresponding full conditional distribution turns out to be nonstandard, which necessitates the use of a Metropolis-Hastings step within the Gibbs sampler (see, e.g., [39], [50] and [68]). We provide the necessary details in Sect. 6.

Finally, we briefly discuss the choice of the hyperparameter N . As argued in [37], in practice it is recommended to use the theoretical results in [37] (that suggest to take $N \asymp n^{\lambda/(2\lambda+1)}$, if the true volatility function s_0 is λ -Hölder smooth) and try several values of N simultaneously. Different N 's all provide information on the unknown volatility, but at different resolution levels; see Section 5 in [37] for an additional discussion. As we will see in simulation examples in Sect. 3, inferential conclusions with the IGMC prior are quite robust with respect to the choice of N . This is because through the hyperparameters α and α_ζ , the IGMC prior has an additional layer for controlling the amount of applied smoothness; when α and α_ζ are equipped with a prior (as above), they can in fact be learned from the data.

3 Synthetic Data Examples

Computations in this section have been done in the programming language Julia, see [3]. In order to test the practical performance of our estimation method, we use a challenging example with the blocks function from [18]. As a second example, we consider the case of the Cox-Ross-Ingersoll model. Precise details are given in the subsections below.

We used the Euler scheme on a grid with 800,001 equidistant points on the interval $[0, 1]$ to obtain realisations of a solution to (1) for different combinations of the drift and dispersion coefficients. These were then subsampled to obtain $n = 4000$ observations in each example.

The hyperparameter α_1 was set to 0.1, whereas for the other two hyperparameters we assumed that $\alpha = \alpha_\zeta$ and used a diffuse IG(0.3, 0.3) prior, except in specially noted cases below. Inference was performed using the Gibbs sampler from Sect. 2, with a Metropolis–Hastings step to update the hyperparameter α . The latter used an independent Gaussian random walk proposal with a scaling to ensure the acceptance rate of ca. 50%; see Sect. 6. The Gibbs sampler was run for 200,000 iterations and we used a burn-in of 1000 samples. In each example we plotted 95% marginal credible bands obtained from the central posterior intervals for the coefficients $\xi_k = \sqrt{\theta_k}$.

3.1 Blocks Function

As our first example, we considered the case when the volatility function was given by the blocks function from [18]. With a vertical shift for positivity, this is defined

as follows:

$$s(t) = 10 + 3.655606 \times \sum_{j=1}^{11} h_j K(t - t_j), \quad t \in [0, 1], \tag{9}$$

where $K(t) = (1 + \text{sgn}(t))/2$, and

$$\{t_j\} = (0.1, 0.13, 0.15, 0.23, 0.25, 0.4, 0.44, 0.65, 0.76, 0.78, 0.81),$$

$$\{h_j\} = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2).$$

The function serves as a challenging benchmark example in nonparametric regression: it is mostly very smooth, but spatially inhomogeneous and characterised by abrupt changes (cf. Chap.9 in [74]). Unlike nonparametric regression, the noise (Wiener process) in our setting should be thought of as multiplicative and proportional to s rather than additive, which combined with the fact that s takes rather large values further complicates the inference problem. Our main goal here was to compare the performance of the IGMC prior-based approach to the IIG prior-based one from [37]. To complete the SDE specification, our drift coefficient was chosen to be a rather strong linear drift $b_0(x) = -10x + 20$.

In Fig. 2 we plot the blocks function (9) and the corresponding realisation of the process X used in this simulation run.

The left and right panels of Fig. 3 contrast the results obtained using the IGMC prior with $N = 160$ and $\alpha = \alpha_\zeta = 20$ versus $N = 320$ and $\alpha = \alpha_\zeta = 40$. These plots illustrate the fact that increasing N has the effect of undersmoothing prior realisations, that can be balanced by increasing the values of α_ζ, α , which has the opposite smoothing effect. Because of this, in fact, both plots look quite similar.

The top left and top right panels of Fig. 4 give estimation results obtained with the IIG prior-based approach from [37]. The number of bins was again $N = 160$ and $N = 320$, and in both these cases we used diffuse independent $\text{IG}(0.1, 0.1)$ priors on the coefficients of the (squared) volatility function (see [37] for details).

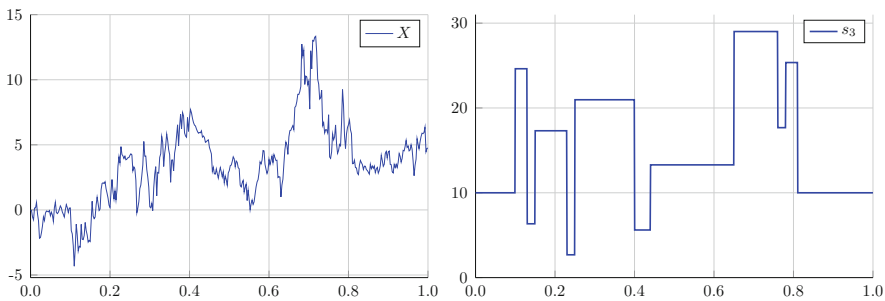


Fig. 2 The sample path of the process X from (9) (left panel) and the corresponding volatility function s (right panel)

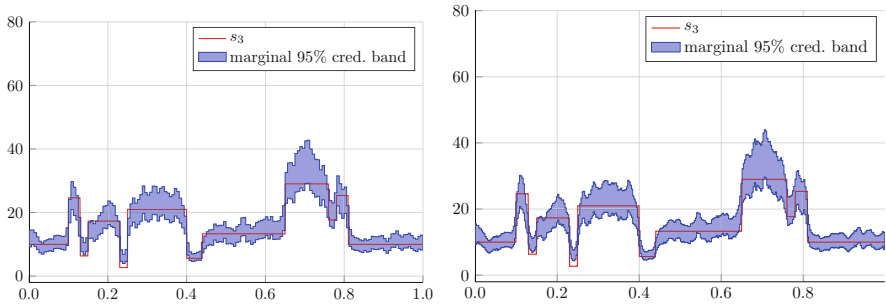


Fig. 3 Volatility function s from (9) with superimposed 95% marginal credible band for the IGMC prior, using $N = 160$, $\alpha = \alpha_\zeta = 20$ (left panel) and $N = 320$, $\alpha = \alpha_\zeta = 40$ (right panel); in both cases, $\alpha_1 = 0.1$

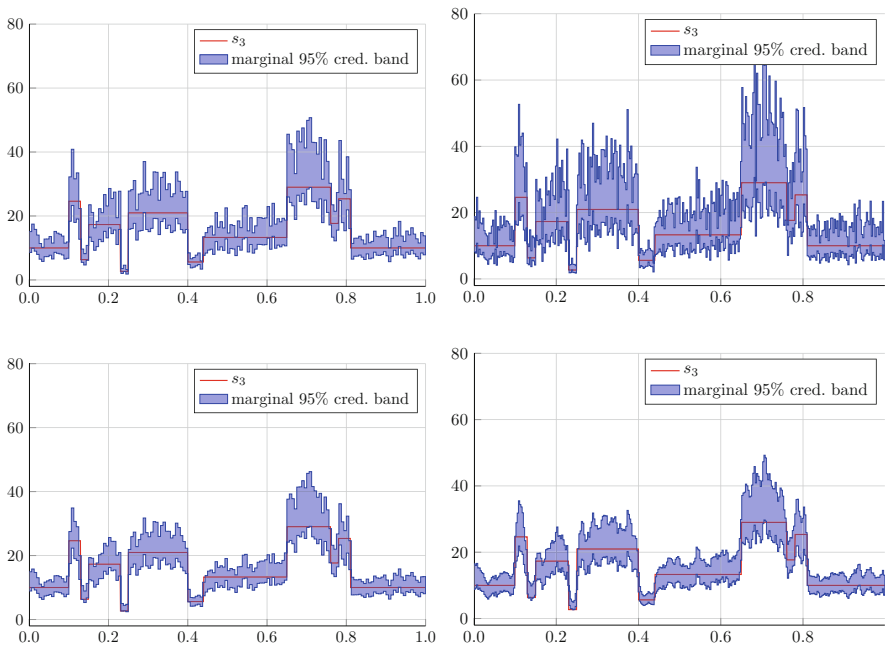


Fig. 4 Volatility function s with superimposed 95% marginal credible band for the IIG prior $IG(0.1, 0.1)$, using $N = 160$ (top left panel) and $N = 320$ bins (top right panel). Volatility function s from (9) with superimposed 95% marginal credible band for the IGMC prior, using $N = 160$ (bottom left panel) and $N = 320$ bins (bottom right panel); in both cases, $\alpha_1 = 0.1$ and $\alpha = \alpha_\zeta \sim IG(0.3, 0.3)$

These results have to be contrasted to those obtained with the IGMC prior, plotted in the bottom left and bottom right panels of Fig. 4, where we assumed $\alpha_1 = 0.1$ and $\alpha = \alpha_\zeta \sim IG(0.3, 0.3)$. The following conclusions emerge from Fig. 4:

- Although both the IGMC and IIG approaches recover globally the shape of the volatility function, the IIG approach results in much greater uncertainty in inferential conclusions, as reflected in wider marginal confidence bands. The effect is especially pronounced in the case $N = 320$, where the width of the band for the IIG prior renders it almost useless for inference.
- The bands based on the IGMC prior look more ‘regular’ than the ones for the IIG prior.
- Comparing the results to Fig. 3, we see the benefits of equipping the hyperparameters α, α_ζ with a prior: credible bands in Fig. 3 do not adequately capture two dips of the function s right before and after the point $t = 0.2$, since s completely falls outside the credible bands there. Thus, an incorrect amount of smoothing is used in Fig. 3.
- The method based on the IIG prior is sensitive to the bin number selection: compare the top left panel of Fig. 4 using $N = 160$ bins to the top right panel using $N = 320$ bins, where the credible band is much wider. On the other hand, the method based on the IGMC prior automatically rebalances the amount of smoothing it uses with different numbers of bins N , thanks to the hyperprior on the parameters α, α_ζ ; in fact, the bottom two plots in Fig. 4 look similar to each other.

3.2 CIR Model

Our core estimation procedure, as described in the previous sections, assumes that the volatility function is deterministic. In this subsection, however, in order to test the limits of applicability of our method and possibilities for future extensions, we applied it to a case where the volatility function was stochastic. The study in [53] lends support to this approach, but here we concentrate on practical aspects and defer the corresponding theoretical investigation until another occasion.

Specifically, we considered the Cox-Ross-Ingersoll (CIR) model or the square root process,

$$dX_t = (\eta_1 - \eta_2 X_t)dt + \eta_3 \sqrt{X_t}dW_t, \quad X_0 = x > 0, \quad t \in [0, T]. \quad (10)$$

Here $\eta_1, \eta_2, \eta_3 > 0$ are parameters of the model. This diffusion process was introduced in [23] and [24], and gained popularity in finance as a model for short-term interest rates, see [11]. The condition $2\eta_1 > \eta_3^2$ ensures strict positivity and ergodicity of X . The volatility function s_0 from (1) now corresponds to a realisation of a stochastic process $t \mapsto \eta_3 \sqrt{X_t}$, where X solves the CIR equation (10).

We took arbitrary parameter values

$$\eta_1 = 6, \quad \eta_2 = 3, \quad \eta_3 = 2, \quad x = 1. \quad (11)$$

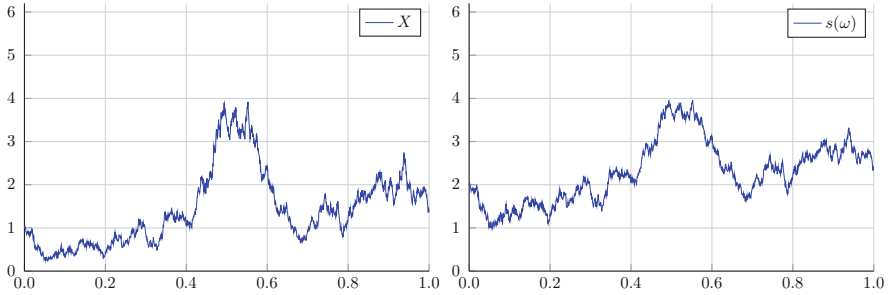


Fig. 5 The sample path of the process X from (10) (left panel) and the corresponding realised volatility function $s(\omega)$ (right panel). The parameter values are given in (11)

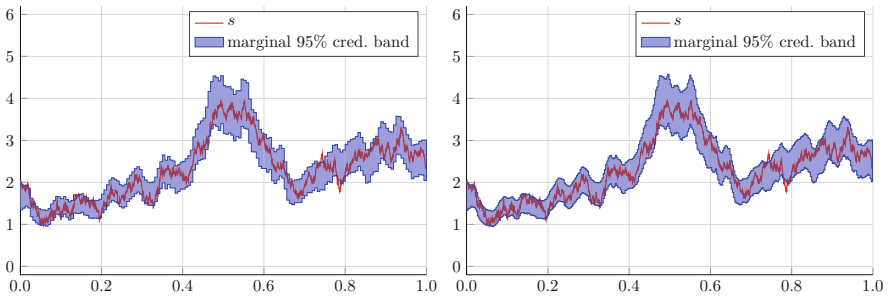


Fig. 6 Volatility function s from (10) with superimposed 95% marginal credible band for the IGMC prior, using $N = 160$ (left panel) and $N = 320$ bins (right panel); in both cases, $\alpha_1 = 0.1$ and $\alpha = \alpha_\zeta \sim \text{IG}(0.3, 0.3)$

A sample path of X is plotted in the left panel of Fig. 5, whereas the corresponding volatility is given in the right panel of the same figure. In Fig. 6 we display estimation results obtained with the IGMC prior, using $N = 160$ and $N = 320$ bins and hyperparameter specifications $\alpha_1 = 0.1$ and $\alpha = \alpha_\zeta \sim \text{IG}(0.3, 0.3)$. A conclusion that emerges from this figure is that our Bayesian method captures the overall shape of the realised volatility in a rather satisfactory manner.

4 Dow-Jones Industrial Averages

In this section we provide a reanalysis of a classical dataset in change-point detection in time series; see, e.g., [10, 14, 41, 42] and [43]. Specifically, we consider weekly closing values of the Dow-Jones industrial averages in the period 2 July 1971–2 August 1974. In total there are 162 observations available, which constitute a relatively small sample, and thus the inference problem is rather nontrivial. The data can be accessed as the dataset DWJ in the `sde` package (see [44]) in **R** (see [58]). See the left panel of Fig. 7 for a visualisation. In [43] the weekly data X_{t_i} ,

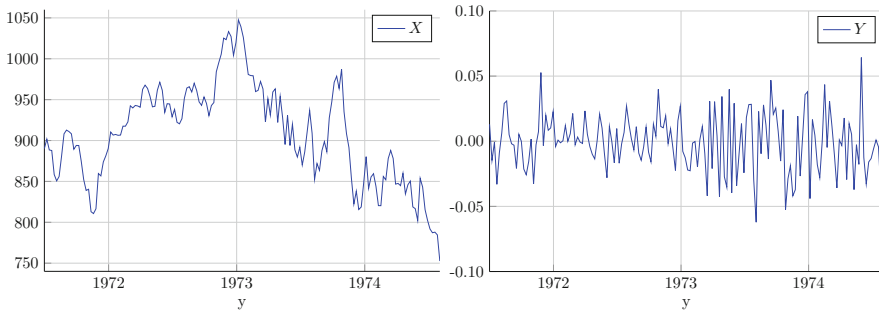


Fig. 7 Dow-Jones weekly closings of industrial averages over the period 2 July 1971–2 August 1974 (left panel) and the corresponding returns (right panel)

$i = 1, \dots, n$, are transformed into returns $Y_{t_i} = (X_{t_i} - X_{t_{i-1}})/X_{t_{i-1}}$, and the least squares change-point estimation procedure from [12] has been performed. Reproducing the corresponding computer code in **R** results in a change-point estimate of 16 March 1973. That author speculates that this change-point is related to the Watergate scandal.

Similar to [43], parametric change-point analyses in [10, 14] and [42] give a change-point in the third week of March 1973. However, as noted in [43], examination of the plot of the time series Y_{t_i} (see Fig. 7, the right panel) indicates that another change-point may be present in the data. Then dropping observations after 16 March 1973 and analysing the data for existence of a change-point using only the initial segment of the time series, the author discovers another change-point on 17 December 1971, which he associates with suspending the convertibility of the US dollar into gold under President Richard Nixon’s administration.

From the above discussion it should be clear that nonparametric modelling of the volatility may provide additional insights for this dataset. We first informally investigated the fact whether an SDE driven by the Wiener process is a suitable model for the data at hand. Many of such models, e.g. the geometric Brownian motion, a classical model for evolution of asset prices over time (also referred to as the Samuelson or Black–Scholes model), rely on an old tenet that returns of asset prices follow a normal distribution. Although the assumption has been empirically disproved for high-frequency financial data (daily or intraday data; see, e.g., [6, 19] and [48]), its violation is less severe for widely spaced data in time (e.g. weekly data, as in our case). In fact, the Shapiro–Wilk test that we performed in **R** on the returns past the change-point 16 March 1973 did not reject the null hypothesis of normality (p -value 0.4). On the other hand, the quantile-quantile (QQ) plot of the same data does perhaps give an indication of a certain mild deviation from normality, see Fig. 8, where we also plotted a kernel density estimate of the data (obtained via the command `density` in **R**, with bandwidth determined automatically through cross-validation).

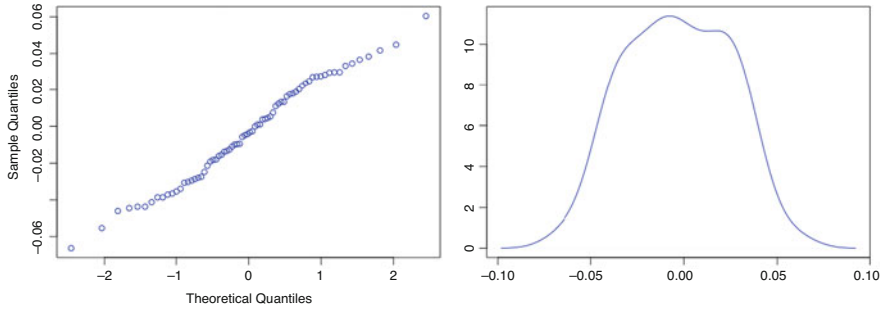


Fig. 8 QQ plot of the returns of Dow-Jones weekly closings of industrial averages over the period 16 March 1973–2 August 1974 (left panel) and a kernel density estimate of the same data (right panel)

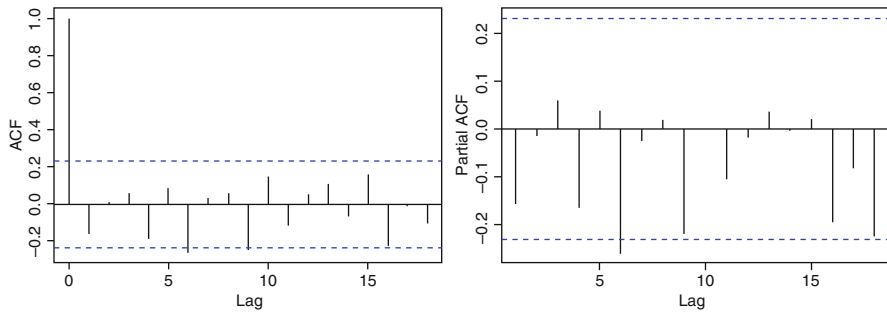


Fig. 9 Sample autocorrelation (left panel) and partial autocorrelation functions of the returns of Dow-Jones weekly closings of industrial averages over the period 16 March 1973–2 August 1974

In Fig. 9 we plot the sample autocorrelation and partial autocorrelation functions based on returns Y_{t_i} 's past the change-point 16 March 1973. These do not give decisive evidence against the assumption of independence of Y_{t_i} 's. Neither does the Ljung–Box test (the test is implemented in **R** via the command `Box.test`), which yields a p -value 0.057 when applied with 10 lags (the p -value is certainly small, but not overwhelmingly so).

Summarising our findings, we detected only a mild evidence against the assumption that the returns of the Dow-Jones weekly closings of industrial averages (over the period 16 March 1973–2 August 1974, but similar conclusions can be reached also over the other subperiods covered by the DWJ dataset) are approximately independent and follow a normal distribution. Thus there is no strong a priori reason to believe that a geometric Brownian motion is an outright unsuitable model in this setting: it can be used as a first approximation. To account for time-variability of volatility (as suggested by the change-point analysis), we incorporate a time-dependent volatility function in the model, and for additional modelling flexibility

we also allow a state-dependent drift. Setting $Z_t = \log(X_t/X_0)$, our model is thus given by

$$dZ_t = b_0(t, Z_t)dt + s_0(t)dW_t, \quad Z_0 = 0. \quad (12)$$

An alternative here is to directly (i.e. without any preliminary transformation) model the Dow-Jones data using Eq. (1). We consider both possibilities, starting with the model (12).

We used a vague prior on θ_1 corresponding to the limit $\alpha_1 \rightarrow 0$, whereas for the other two hyperparameters we assumed $\alpha = \alpha_\zeta \sim \text{IG}(0.3, 0.3)$. The scaling in the independent Gaussian random walk proposal in the Metropolis–Hastings step was chosen in such a way so as to yield an acceptance rate of ca. 50%. The Gibbs sampler was run for 200,000 iterations, and the first 1000 samples were dropped as a burn-in. We present the estimation results we obtained using $N = 13$ and $N = 26$ bins, see Fig. 10. Although the sample size n is quite small in this example, the data are informative enough to yield nontrivial inferential conclusions even with diffuse priors. Both plots in Fig. 10 are qualitatively similar and suggest:

- A decrease in volatility at the end of 1971, which can be taken as corresponding to the change-point in December 1971 identified in [43]. Unlike that author, we do not directly associate it with suspending the convertibility of the US dollar into gold (that took place in August 1971 rather than December 1971).
- A gradual increase in volatility over the subsequent period stretching until the end of 1973. Rather than only the Watergate scandal (and a change-point in March 1973 as in [43]), there could be further economic causes for that, such as the 1973 oil crisis and the 1973–1974 stock market crash.
- A decrease in volatility starting in early 1974, compared to the immediately preceding period.

In general, in this work we do not aim at identifying causes for changes in volatility regimes, but prefer to present our inference results, that may subsequently be used in econometric analyses.

Now we move to the Bayesian analysis of the data using model (1). The prior settings were as in the previous case, and we display the results in Fig. 11. The overall shapes of the inferred volatility functions are the same in both Figs. 10 and 11, and hence similar conclusions apply.

Finally, we stress the fact that our nonparametric Bayesian approach and change-point estimation are different in their scope: whereas our method aims at estimation of the entire volatility function, change-point estimation (as its name actually suggests) concentrates on identifying change-points in the variance of the observed time series, which is a particular feature of the volatility. To that end it assumes the (true) volatility function is piecewise constant, which on the other hand is not an assumption required in our method. Both techniques are useful, and each can provide insights that may be difficult to obtain from another.

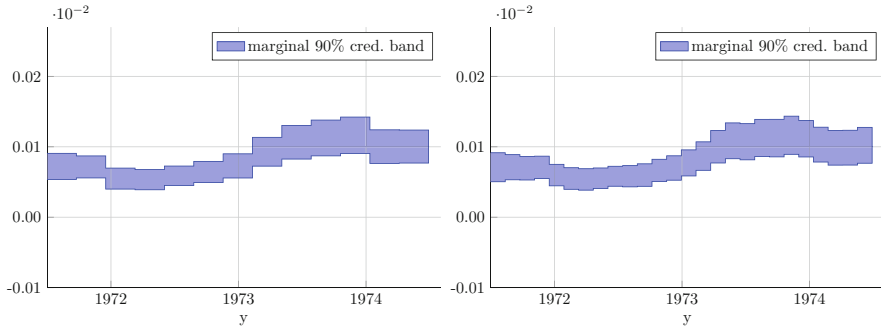


Fig. 10 Marginal 90% credible bands for the volatility function of the log Dow-Jones industrial averages data. The left panel corresponds to $N = 13$ bins, while the right panel to $N = 26$ bins

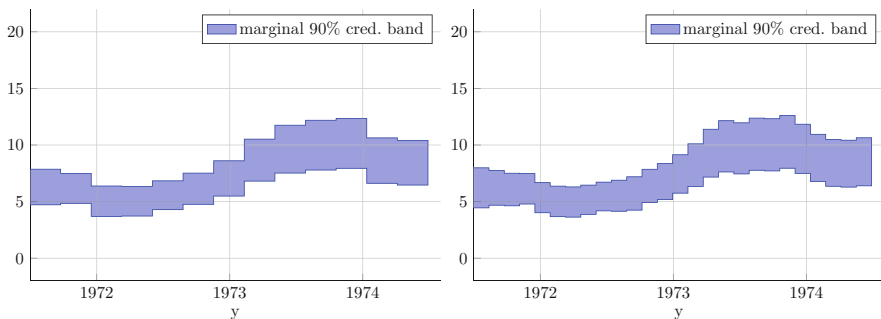


Fig. 11 Marginal 90% credible bands for the volatility function of the Dow-Jones industrial averages data. The left panel corresponds to $N = 13$ bins, while the right panel to $N = 26$ bins

5 Conclusions

Bayesian inference for SDEs from discrete-time observations is a difficult task, owing to intractability of the likelihood and the fact that the posterior is not available in closed form. Posterior inference therefore typically requires the use of intricate MCMC samplers. Designing algorithms that result in Markov chains that mix well and explore efficiently the posterior surface is a highly nontrivial problem. Inspired by some ideas from the audio signal processing literature and our earlier work [37], in this paper we introduced a novel nonparametric Bayesian approach to estimation of the volatility coefficient of an SDE. Our method is easy to understand and straightforward to implement via Gibbs sampling, and performs well in practice. Thereby our hope is that our work will contribute to further dissemination and popularisation of a nonparametric Bayesian approach to inference in SDEs, specifically with financial applications in mind. In that respect, see [38], that builds upon the present paper and deals with Bayesian volatility estimation under market microstructure noise. Our work can also be viewed as a

partial fulfillment of anticipation in [31] that some ideas developed originally in the context of audio and music processing “will also find use in other areas of science and engineering, such as financial or biomedical data analysis”.

As a final remark, we do not attempt to provide a theoretical, i.e. asymptotic frequentist analysis of our new approach here (see, e.g., the recent monograph [30], and specifically [37] for such an analysis in the SDE setting), but leave this as a topic of future research.

6 Formulae for Parameter Updates

In this section we present additional details on the derivation of the update formulae for the Gibbs sampler from Sect. 2. The starting point is to employ the Markov property from (4), and using the standard Bayesian notation, to write the joint density of $\{\zeta_k\}$ and $\{\theta_k\}$ as

$$p(\theta_1) \prod_{k=2}^N p(\zeta_k | \theta_{k-1}) p(\theta_k | \zeta_k). \tag{13}$$

6.1 Full Conditional Distributions

We first indicate how (5) was derived. Insert expressions for the individual terms in (13) from (4) and collect separately terms that depend on θ_k only, to see that the density of the full conditional distribution of $\theta_k, k = 2, \dots, N - 1$, is proportional to

$$\theta_k^{-\alpha-1} e^{-\alpha/(\theta_k \zeta_k)} \theta_k^{-\alpha \zeta} e^{-\alpha \zeta / (\theta_k \zeta_{k+1})}.$$

Upon normalisation, this expression is the density of the $\text{IG}(\alpha + \alpha \zeta, \alpha \zeta_k^{-1} + \alpha \zeta \zeta_{k+1}^{-1})$ distribution, which proves formula (5). Formula (7) follows directly from the last expression in (4). Formula (8) is proved analogously to (5). Finally, (6) follows from (4) and Bayes’ formula. Cf. also [15], Appendix B.6.

6.2 Metropolis-Within-Gibbs Step

Now we describe the Metropolis–Hastings step within the Gibbs sampler, that is used to update the hyperparameters of our algorithm, in case the latter are equipped with a prior. For simplicity, assume $\alpha = \alpha_\zeta$ (we note that such a choice is used in practical examples in [8]), and suppose α is equipped with a prior, $\alpha \sim \pi$. Let the hyperparameter α_1 be fixed. Obviously, α_1 could have been equipped with a prior as well, but this would have further slowed down our estimation procedure,

whereas the practical results in Sects. 3 and 4 we obtained are already satisfactory with α_1 fixed. Using (4) and (13), one sees that the joint density of $\{\zeta_k\}$, $\{\theta_k\}$ and α is proportional to

$$\pi(\alpha) \times \theta_1^{-\alpha_1-1} \times e^{-\alpha_1\theta_1^{-1}} \times \prod_{k=2}^N \left\{ \frac{\alpha^\alpha}{\Gamma(\alpha)\theta_{k-1}^\alpha} \zeta_k^{-\alpha-1} e^{-\alpha/(\theta_{k-1}\zeta_k)} \frac{\alpha^\alpha}{\Gamma(\alpha)\zeta_k^\alpha} \theta_k^{-\alpha-1} e^{-\alpha/(\theta_k\zeta_k)} \right\}.$$

This in turn is proportional to

$$q(\alpha) = \pi(\alpha) \times \left(\frac{\alpha^\alpha}{\Gamma(\alpha)} \right)^{2(N-1)} \times \prod_{k=2}^N (\theta_{k-1}\theta_k\zeta_k^2)^{-\alpha} \times \exp\left(-\alpha \sum_{k=2}^N \frac{1}{\zeta_k} \left(\frac{1}{\theta_{k-1}} + \frac{1}{\theta_k} \right)\right).$$

The latter expression is an unnormalised full conditional density of α , and can be used in the Metropolis-within-Gibbs step to update α .

The rest of the Metropolis–Hastings step is standard, and the following approach was used in our practical examples: pick a proposal kernel $g(\alpha' | \alpha)$, for instance a Gaussian random walk proposal $g(\alpha' | \alpha) = \phi_\sigma(\alpha' - \alpha)$, where ϕ_σ is the density of a normal random variable with mean zero and variance σ^2 . Note that this specific choice may result in proposing a negative value α' , which needs to be rejected straightaway as invalid. Then, for computational efficiency, instead of moving to another step within the Gibbs sampler, one keeps on proposing new values α' until a positive one is proposed. This is then accepted with probability

$$A = \min\left(1, \frac{q(\alpha') \Phi_\sigma(\alpha)}{q(\alpha) \Phi_\sigma(\alpha')}\right),$$

where $\Phi_\sigma(\cdot)$ is the cumulative distribution function of a normal random variable with mean zero and variance σ^2 ; otherwise the current value α is retained. See [75] for additional details and derivations. Finally, one moves to other steps in the Gibbs sampler, namely to updating ζ_k 's and θ_k 's, and iterates the procedure. The acceptance rate in the Metropolis–Hastings step can be controlled through the scale parameter σ of the proposal density ϕ_σ . Some practical rules for determination of an optimal acceptance rate in the Metropolis–Hastings algorithm are given in [28], and those for the Metropolis-within-Gibbs algorithm in [64].

Acknowledgements The research leading to the results in this paper has received funding from the European Research Council under ERC Grant Agreement 320637.

References

1. Allen, E.: Modeling with Itô stochastic differential equations. *Mathematical Modelling: Theory and Applications*, vol. 22. Springer, Dordrecht (2007)
2. Batz, P., Ruttor, A., Opper, M.: Approximate Bayes learning of stochastic differential equations. *Phys. Rev. E* **98**, 022109 (2018)
3. Bezanson, J., Edelman, A., Karpinski, S., Shah, V.B.: Julia: a fresh approach to numerical computing. *SIAM Rev.* **59**(1), 65–98 (2017)
4. Björk, T.: *Arbitrage Theory in Continuous Time*, 3rd edn. Oxford University Press, Oxford (2009)
5. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. *J. Am. Stat. Assoc.* **112**(518), 859–877 (2017)
6. Carr, P., Geman, H., Madan, D.B., Yor, M.: The fine structure of asset returns: an empirical investigation. *J. Bus.* **75**(2), 305–332 (2002)
7. Cemgil, A.T., Dikmen, O.: Conjugate gamma Markov random fields for modelling nonstationary sources. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) *International Conference on Independent Component Analysis and Signal Separation (ICA 2007)*, pp. 697–705. Springer, Berlin (2007)
8. Cemgil, A.T., Dikmen, O.: Inference and learning in gamma chains for Bayesian audio processing. *J. Acoust. Soc. Am.* **123**(5), 3585–3585 (2008)
9. Cemgil, A.T., Févotte, C., Godsill, S.J.: Variational and stochastic inference for Bayesian source separation. *Digital Signal Process.* **17**(5), 891–913 (2007). Special Issue on Bayesian Source Separation
10. Chen, J., Gupta, A.K.: *Parametric Statistical Change Point Analysis*, 2nd edn. Birkhäuser/Springer, New York (2012). With applications to genetics, medicine, and finance
11. Cox, J.C., Ingersoll Jr., J.E., Ross, S.A.: A theory of the term structure of interest rates. *Econometrica* **53**(2), 385–407 (1985)
12. De Gregorio, A., Iacus, S.M.: Least squares volatility change point estimation for partially observed diffusion processes. *Commun. Stat. Theory Methods* **37**(13–15), 2342–2357 (2008)
13. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B: Methodol.* **39**(1), 1–38 (1977)
14. Díaz, J.: Bayesian detection of a change of scale parameter in sequences of independent gamma random variables. *J. Econom.* **19**(1), 23–29 (1982)
15. Dikmen, O., Cemgil, A.T.: Inference and parameter estimation in Gamma chains. Tech. Rep. CUED/F-INFENG/TR.596, University of Cambridge (2008)
16. Dikmen, O., Cemgil, A.T.: Gamma Markov random fields for audio source modeling. *IEEE Trans. Audio Speech Lang. Process.* **18**(3), 589–601 (2010)
17. Donnet, S., Rivoirard, V., Rousseau, J., Scricciolo, C.: Posterior concentration rates for empirical Bayes procedures with applications to Dirichlet process mixtures. *Bernoulli* **24**(1), 231–256 (2018)
18. Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. *J. Am. Stat. Assoc.* **90**(432), 1200–1224 (1995)
19. Eberlein, E., Keller, U.: Hyperbolic distributions in finance. *Bernoulli* **1**(3), 281–299 (1995)
20. Efron, B.: *Large-scale inference*, Institute of Mathematical Statistics (IMS) Monographs, vol. 1. Cambridge University Press, Cambridge (2010). Empirical Bayes methods for estimation, testing, and prediction
21. Elerian, O., Chib, S., Shephard, N.: Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* **69**(4), 959–993 (2001)
22. Fearnhead, P.: Exact and efficient Bayesian inference for multiple changepoint problems. *Stat. Comput.* **16**(2), 203–213 (2006)
23. Feller, W.: Diffusion processes in genetics. In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pp. 227–246. University of California Press, Berkeley (1951)

24. Feller, W.: Two singular diffusion problems. *Ann. Math.* **54**(1), 173–182 (1951)
25. Fuchs, C.: *Inference for Diffusion Processes*. Springer, Heidelberg (2013). With applications in life sciences, With a foreword by Ludwig Fahrmeir
26. Gelfand, A.E., Smith, A.F.M.: Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **85**(410), 398–409 (1990)
27. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: *Bayesian Data Analysis*, 3rd edn. Texts in Statistical Science Series. CRC Press, Boca Raton (2014)
28. Gelman, A., Roberts, G.O., Gilks, W.R.: Efficient Metropolis jumping rules. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) *Bayesian Statistics*, vol. 5, pp. 599–607. Oxford University Press, Oxford (1996)
29. Geman, S., Geman, D.: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**(6), 721–741 (1984)
30. Ghosal, S., van der Vaart, A.: *Fundamentals of nonparametric Bayesian inference*. Cambridge Series in Statistical and Probabilistic Mathematics, vol. 44. Cambridge University Press, Cambridge (2017)
31. Godsill, S.J., Cemgil, A.T., Févotte, C., Wolfe, P.J.: Bayesian computational methods for sparse audio and music processing. In: 15th European Signal Processing Conference (EURASIP), pp. 345–349. IEEE, Poznan (2007)
32. Green, P.J.: Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**(4), 711–732 (1995)
33. Green, P.J.: Trans-dimensional Markov chain Monte Carlo. In: *Highly Structured Stochastic Systems*, Oxford Statistical Science Series, vol. 27, pp. 179–206. Oxford University Press, Oxford (2003). With part A by Simon J. Godsill and part B by Juha Heikkinen
34. Gugushvili, S., Spreij, P.: Consistent non-parametric Bayesian estimation for a time-inhomogeneous Brownian motion. *ESAIM Probab. Stat.* **18**, 332–341 (2014)
35. Gugushvili, S., Spreij, P.: Nonparametric Bayesian drift estimation for multidimensional stochastic differential equations. *Lith. Math. J.* **54**(2), 127–141 (2014)
36. Gugushvili, S., Spreij, P.: Posterior contraction rate for non-parametric Bayesian estimation of the dispersion coefficient of a stochastic differential equation. *ESAIM Probab. Stat.* **20**, 143–153 (2016)
37. Gugushvili, S., van der Meulen, F., Schauer, M., Spreij, P.: Nonparametric Bayesian estimation of a Hölder continuous diffusion coefficient (2017). Preprint. arXiv:1706.07449v4
38. Gugushvili, S., van der Meulen, F., Schauer, M., Spreij, P.: Nonparametric Bayesian volatility learning under microstructure noise (2018). arXiv:1805.05606v1
39. Hastings, W.K.: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109 (1970)
40. Hindriks, R.: *Empirical dynamics of neuronal rhythms: data-driven modeling of spontaneous magnetoencephalographic and local field potential recordings*. PhD. Thesis, Vrije Universiteit Amsterdam (2011)
41. Hsu, D.A.: Tests for variance shift at an unknown time point. *J. R. Stat. Soc.: Ser. C: Appl. Stat.* **26**(3), 279–284 (1977)
42. Hsu, D.A.: Detecting shifts of parameter in gamma sequences with applications to stock price and air traffic flow analysis. *J. Am. Stat. Assoc.* **74**(365), 31–40 (1979)
43. Iacus, S.M.: *Simulation and inference for stochastic differential equations*. Springer Series in Statistics. Springer, New York (2008). With **R** examples
44. Iacus, S.M.: *SDE: simulation and inference for Stochastic differential equations* (2016). <https://CRAN.R-project.org/package=sde>. R package version 2.0.15
45. Jacquier, E., Polson, N.G., Rossi, P.E.: Bayesian analysis of stochastic volatility models. *J. Bus. Econ. Stat.* **12**(4), 371–389 (1994)
46. Jacquier, E., Polson, N.G., Rossi, P.E.: Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *J. Econom.* **122**(1), 185–212 (2004)
47. Karatzas, I., Shreve, S.E.: *Brownian motion and stochastic calculus*, Graduate Texts in Mathematics, vol. 113, 2nd edn. Springer, New York (1991)

48. Küchler, U., Neumann, K., Sørensen, M., Steller, A.: Stock returns and hyperbolic distributions. *Math. Comput. Model.* **29**(10), 1–15 (1999)
49. Martin, R., Ouyang, C., Domagni, F.: ‘Purposely misspecified’ posterior inference on the volatility of a jump diffusion process. *Statist. Probab. Lett.* **134**(Supplement C), 106–113 (2018)
50. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E.: Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**(6), 1087–1092 (1953)
51. Müller, P., Mitra, R.: Bayesian nonparametric inference—why and how. *Bayesian Anal.* **8**(2), 269–302 (2013)
52. Musiela, M., Rutkowski, M.: *Martingale methods in financial modelling*. Stochastic Modelling and Applied Probability, vol. 36, 2nd edn. Springer, Berlin (2005)
53. Mykland, P.A.: A Gaussian calculus for inference from high frequency data. *Ann. Finance* **8**(2–3), 235–258 (2012)
54. Nickl, R., Söhl, J.: Nonparametric Bayesian posterior contraction rates for discretely observed scalar diffusions. *Ann. Stat.* **45**(4), 1664–1693 (2017)
55. Papaspiliopoulos, O., Pokern, Y., Roberts, G.O., Stuart, A.M.: Nonparametric estimation of diffusions: a differential equations approach. *Biometrika* **99**(3), 511–531 (2012)
56. Petrone, S., Rousseau, J., Scricciolo, C.: Bayes and empirical Bayes: do they merge? *Biometrika* **101**(2), 285–302 (2014)
57. Pokern, Y., Stuart, A.M., van Zanten, J.H.: Posterior consistency via precision operators for Bayesian nonparametric drift estimation in SDEs. *Stoch. Process. Appl.* **123**(2), 603–628 (2013)
58. R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna (2017)
59. Robbins, H.: An empirical Bayes approach to statistics. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955*, vol. I, pp. 157–163. University of California Press, Berkeley (1956)
60. Robbins, H.: The empirical Bayes approach to statistical decision problems. *Ann. Math. Stat.* **35**, 1–20 (1964)
61. Robert, C.P., Casella, G.: *Monte Carlo Statistical Methods*, 2nd edn. Springer Texts in Statistics. Springer, New York (2004)
62. Roberts, G.O., Stramer, O.: On inference for partially observed nonlinear diffusion models using the Metropolis–Hastings algorithm. *Biometrika* **88**(3), 603–621 (2001)
63. Ruttor, A., Batz, P., Opper, M.: Approximate Gaussian process inference for the drift function in stochastic differential equations. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) *Proceedings of the 26th International Conference on Neural Information Processing Systems*, pp. 2040–2048. Curran Associates, New York (2013)
64. Sherlock, C., Fearnhead, P., Roberts, G.O.: The random walk Metropolis: linking theory and practice through a case study. *Stat. Sci.* **25**(2), 172–190 (2010)
65. Silverman, B.W.: *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. Chapman & Hall, London (1986)
66. Soulier, P.: Nonparametric estimation of the diffusion coefficient of a diffusion process. *Stoch. Anal. Appl.* **16**(1), 185–200 (1998)
67. Tanner, M.A., Wong, W.H.: The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **82**(398), 528–550 (1987). With discussion and with a reply by the authors
68. Tierney, L.: Markov chains for exploring posterior distributions. *Ann. Stat.* **22**(4), 1701–1762 (1994). With discussion and a rejoinder by the author
69. van der Meulen, F., Schauer, M.: Bayesian estimation of discretely observed multi-dimensional diffusion processes using guided proposals. *Electron. J. Stat.* **11**(1), 2358–2396 (2017)
70. van der Meulen, F., van Zanten, H.: Consistent nonparametric Bayesian inference for discretely observed scalar diffusions. *Bernoulli* **19**(1), 44–63 (2013)
71. van der Meulen, F., Schauer, M., van Zanten, H.: Reversible jump MCMC for nonparametric drift estimation for diffusion processes. *Comput. Stat. Data Anal.* **71**, 615–632 (2014)

72. van Zanten, H.: Nonparametric Bayesian methods for one-dimensional diffusion models. *Math. Biosci.* **243**(2), 215–222 (2013)
73. Virtanen, T., Cemgil, A.T., Godsill, S.: Bayesian extensions to non-negative matrix factorisation for audio signal modelling. In: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1825–1828. IEEE, Las Vegas (2008)
74. Wasserman, L.: *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer, New York (2006)
75. Wilkinson, D.J.: Metropolis Hastings MCMC when the proposal and target have differing support. <https://darrenjw.wordpress.com/2012/06/04/metropolis-hastings-mcmc-when-the-proposal-and-target-have-differing-support/> (2012). Accessed 23 December 2017
76. Wong, E., Hajek, B.: *Stochastic processes in engineering systems*. Springer Texts in Electrical Engineering. Springer, New York (1985)
77. Yang, T.Y., Kuo, L.: Bayesian binary segmentation procedure for a Poisson process with multiple changepoints. *J. Comput. Graph. Stat.* **10**(4), 772–785 (2001)

The Exact Asymptotics for Hitting Probability of a Remote Orthant by a Multivariate Lévy Process: The Cramér Case



Konstantin Borovkov and Zbigniew Palmowski

Abstract For a multivariate Lévy process satisfying the Cramér moment condition and having a drift vector with at least one negative component, we derive the exact asymptotics of the probability of ever hitting the positive orthant that is being translated to infinity along a fixed vector with positive components. This problem is motivated by the multivariate ruin problem introduced in Avram et al. (Ann Appl Probab 18:2421–2449, 2008) in the two-dimensional case. Our solution relies on the analysis from Pan and Borovkov (Preprint. arXiv:1708.09605, 2017) for multivariate random walks and an appropriate time discretization.

1 Introduction

In this note we consider the following large deviation problem for continuous time processes with independent increments that was motivated by the multivariate simultaneous ruin problem introduced in [1]. Let $\{X(t)\}_{t \geq 0}$ be a d -dimensional ($d \geq 2$) right-continuous Lévy process with $X(0) = 0$. One is interested in finding the precise asymptotics for the hitting probability of the orthant sG as $s \rightarrow \infty$, where

$$G := \mathbf{g} + Q^+$$

K. Borovkov (✉)

School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC, Australia
e-mail: borovkov@unimelb.edu.au

Z. Palmowski

Dept. of Applied Mathematics, Wrocław University of Science and Technology, Wrocław, Poland

for some fixed

$$\mathbf{g} \in Q^+, \quad Q^+ := \{\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d : x_j > 0, 1 \leq j \leq d\}.$$

Clearly, $sG = s\mathbf{g} + Q^+$, which is just the positive orthant translated by $s\mathbf{g}$.

We solve this problem under appropriate Cramér moment assumptions and further conditions on the process X and vertex \mathbf{g} that, roughly speaking, ensure that the “most likely place” for X to hit sG when s is large is in vicinity of the “corner point” $s\mathbf{g}$. More specifically, we show that the precise asymptotics of the hitting probability of sG are given by the following expression: letting

$$\tau(V) := \inf\{t \geq 0 : X(t) \in V\}$$

be the first hitting time of the set $V \subset \mathbb{R}^d$ by the process X , one has

$$\mathbf{P}(\tau(sG) < \infty) = A_0 s^{-(d-1)/2} e^{-sD(G)} (1 + o(1)) \quad \text{as } s \rightarrow \infty, \quad (1)$$

where the “adjustment coefficient” $D(G)$ is the value of the second rate function (see (9) below) for the distribution of $X(1)$ on the set G and the constant $A_0 \in (0, \infty)$ can be computed explicitly.

The asymptotics (1) extend a number of known results. The main body of literature on the topic of precise asymptotics for boundary crossing large deviation probabilities in the multivariate case concerns the random walk theory, see [3–5] and references therein for an overview of the relevant results. The crude logarithmic asymptotics in the multivariate case was also derived independently in [6].

The entrance probability to a remote set for Lévy processes was analyzed later, usually under some specific assumptions on the structure of these processes. For example, paper [1] dealt with the two-dimensional reserve process of the form

$$X(t) = (X_1(t), X_2(t)) = (c_1, c_2) \sum_{i=1}^{N(t)} C_i - (p_1, p_2)t, \quad t \geq 0, \quad (2)$$

where $c_i, p_i > 0, i = 1, 2$, are constants, $\{C_i\}_{i \geq 1}$ is a sequence of i.i.d. claim sizes, and $N(t)$ is an independent of the claim sizes Poisson process. That model admits the following interpretation: the components of the process $s\mathbf{g} - X$ describe the dynamics of the reserves of two insurance companies that start with the initial reserves sg_1 and sg_2 , respectively, and then divide between them both claims and premia in some pre-specified proportions. In that case, $\mathbf{P}(\tau(sG) < \infty)$ corresponds to the simultaneous ruin probability of the two companies. The main result of the present paper generalizes the assertion of Theorem 5 of [1] to the case of general Lévy processes. One may also wish to mention here the relevant papers [2, 7].

2 The Main Result

To state the main result, we will need some notations. For brevity, denote by ξ a random vector such that

$$\xi \stackrel{d}{=} \mathbf{X}(1). \tag{3}$$

Our first condition on \mathbf{X} is stated as follows.

C₁ *The distribution of ξ is non-lattice and there is no hyperplane $H = \{\mathbf{x} : \langle \mathbf{a}, \mathbf{x} \rangle = c\} \subset \mathbb{R}^d$ such that $\mathbf{P}(\xi \in H) = 1$.*

That condition can clearly be re-stated in terms of the covariance matrix of the Brownian component and spectral measure of \mathbf{X} , although such re-statement will not make it more compact nor transparent.

Next denote by

$$K(\boldsymbol{\lambda}) := \ln \mathbf{E} e^{\langle \boldsymbol{\lambda}, \xi \rangle}, \quad \boldsymbol{\lambda} \in \mathbb{R}^d, \tag{4}$$

the cumulant function of ξ and let

$$\Theta_\psi := \{\boldsymbol{\lambda} \in \mathbb{R}^d : K(\boldsymbol{\lambda}) < \infty\}$$

be the set on which the moment generating function of ξ is finite. We will need the following Cramér moment condition on \mathbf{X} :

C₂ *Θ_ψ contains a non-empty open set.*

The first rate function $\Lambda(\boldsymbol{\alpha})$ for the random vector ξ is defined as the Legendre transform of the cumulant function K :

$$\Lambda(\boldsymbol{\alpha}) := \sup_{\boldsymbol{\lambda} \in \Theta_\psi} (\langle \boldsymbol{\alpha}, \boldsymbol{\lambda} \rangle - K(\boldsymbol{\lambda})), \quad \boldsymbol{\alpha} \in \mathbb{R}^d. \tag{5}$$

The probabilistic interpretation of the first rate function is given by the following relation (see e.g. [5]): for any $\boldsymbol{\alpha} \in \mathbb{R}^d$,

$$\Lambda(\boldsymbol{\alpha}) = - \lim_{\varepsilon \rightarrow 0} \lim_{n \rightarrow \infty} \frac{1}{n} \ln \mathbf{P} \left(\frac{\mathbf{X}(n)}{n} \in U_\varepsilon(\boldsymbol{\alpha}) \right), \tag{6}$$

where $U_\varepsilon(\boldsymbol{\alpha})$ is the ε -neighborhood of $\boldsymbol{\alpha}$. Accordingly, for a set $B \subset \mathbb{R}^d$, any point $\boldsymbol{\alpha} \in B$ such that

$$\Lambda(\boldsymbol{\alpha}) = \Lambda(B) := \inf_{\mathbf{v} \in B} \Lambda(\mathbf{v}) \tag{7}$$

is called a most probable point (MPP) of the set B (cf. relation (11) in [8]). If such a point α is unique for a given set B , we denote it by

$$\alpha[B] := \arg \min_{v \in B} \Lambda(v). \quad (8)$$

Now recall the definition of the second rate function D that was introduced and studied in [4]: letting $D_u(v) := u \Lambda(v/u)$ for $v \in \mathbb{R}^d$, one sets

$$D(v) := \inf_{u>0} D_u(v), \quad v \in \mathbb{R}^d, \quad D(B) := \inf_{v \in B} D(v), \quad B \subset \mathbb{R}^d \quad (9)$$

(see also [8]). Further, we put

$$r_B := \arg \min_{r>0} D_{1/r}(B). \quad (10)$$

Recall the probabilistic meaning of the function D and the value r_B . While the first rate function Λ specifies the main term in the asymptotics of the probabilities for the random walk values $X(n)$ to be inside “remote sets” (roughly speaking, $\Lambda(B)$ equals the RHS of (6) with the neighbourhood of α in it replaced with B), the second rate function D does that for the probabilities of *ever hitting* “remote sets” by the whole random walk trajectory $\{X(n)\}_{n \geq 0}$, the meaning of r_B being that $1/r_B$ gives (after appropriate scaling) the “most probable time” for the walk to hit the respective remote set. For more detail, we refer the interested reader to [4, 8].

Define the Cramér range Ω_Λ for ξ as follows:

$$\Omega_\Lambda := \{ \alpha = \text{grad } K(\lambda) : \lambda \in \text{int}(\Theta_\psi) \},$$

where the cumulant function $K(\lambda)$ of ξ was defined in (4) and $\text{int}(B)$ stands for the interior of the set B . In words, the set Ω_Λ consists of all the vectors that can be obtained as the expectations of the Cramér transforms of the law of ξ , i.e. the distributions of the form $e^{\langle \lambda, x \rangle - K(\lambda)} \mathbf{P}(\xi \in dx)$, for parameter values $\lambda \in \text{int}(\Theta_\psi)$.

For $\alpha \in \mathbb{R}^d$, denote by $\lambda(\alpha)$ the vector λ at which the upper bound in (5) is attained (when such a vector exists, in which case it is always unique):

$$\Lambda(\alpha) = \langle \alpha, \lambda(\alpha) \rangle - K(\lambda(\alpha)).$$

For $r > 0$, assuming that $\alpha[rG] \in \Omega_\Lambda$, introduce the vector

$$N(r) := \text{grad } \Lambda(\alpha) \Big|_{\alpha=\alpha[rG]} = \lambda(\alpha[rG]), \quad (11)$$

which is a normal to the level surface of Λ at the point $\alpha[rG]$ (see e.g. (22) in [8]).

The last condition that we will need to state our main result depends on the parameter $r > 0$ and is formulated as follows:

$C_3(r)$ One has

$$\Lambda(rG) = \Lambda(r\mathbf{g}), \quad r\mathbf{g} \in \Omega_\Lambda, \quad N(r) \in Q^+, \quad \langle \mathbf{E}\xi, N(r) \rangle < 0.$$

The first part of condition $[C_3(r)]$ means that the vertex $r\mathbf{g}$ is an MPP for the set rG . Note that under the second part of the condition, this MPP $r\mathbf{g}$ for rG is unique (e.g., by Lemma 1 in [8]). Since $N(r)$ always belongs to the closure of Q^+ , the third part of condition $[C_3(r)]$ just excludes the case when the normal $N(r)$ to the level surface of Λ at the point $r\mathbf{g}$ belongs to the boundary of the set rG .

Theorem 1 *Let conditions $[C_1]$, $[C_2]$ and $[C_3(rG)]$ be met. Then the asymptotic relation (1) holds true, where $D(G)$ is the value of the second rate function (9) on G and the constant $A_0 \in (0, \infty)$ can be computed explicitly.*

The value of the constant $A_0 \in (0, \infty)$ is given by the limit as $\delta \rightarrow 0$ of the expressions given by formula (68) in [8] for the distribution of $\xi \stackrel{d}{=} X(\delta)$. When proving the theorem below, we demonstrate that that limit does exist and is finite and positive.

Proof For a $\delta > 0$, consider the embedded random walk $\{X(n\delta)\}_{n \in \mathbb{N}}$ and, for a set $V \subset \mathbb{R}^d$, denote the first time that random walk hits that set V by

$$\eta_\delta(V) := \inf\{n \in \mathbb{N} : X(n\delta) \in V\}.$$

First observe that, on the one hand, for any $\delta > 0$, one clearly has

$$\mathbf{P}(\tau(sG) < \infty) \geq \mathbf{P}(\eta_\delta(sG) < \infty). \tag{12}$$

On the other hand, assuming without loss of generality that $\min_{1 \leq j \leq d} g_j \geq 1$ and setting $I(s) := (\tau(sG), \tau(sG) + \delta] \subset \mathbb{R}$ on the event $\{\tau(sG) < \infty\}$, we have, for any $\varepsilon > 0$,

$$\begin{aligned} \mathbf{P}(\eta_\delta((s - \varepsilon)G) < \infty) &\geq \mathbf{P}\left(\tau(sG) < \infty, \sup_{t \in I(s)} \|X(t) - X(\tau(sG))\| \leq \varepsilon\right) \\ &= \mathbf{P}(\tau(sG) < \infty) \mathbf{P}\left(\sup_{t \in I(s)} \|X(t) - X(\tau(sG))\| \leq \varepsilon\right) \\ &= \mathbf{P}(\tau(sG) < \infty) \mathbf{P}\left(\sup_{t \in (0, \delta]} \|X(t)\| \leq \varepsilon\right), \end{aligned} \tag{13}$$

where the last two relations follow from the strong Markov property and homogeneity of X .

Now take an arbitrary small $\varepsilon > 0$. As the process X is right-continuous, there exists a $\delta(\varepsilon) > 0$ such that

$$\mathbf{P}\left(\sup_{t \in (0, \delta(\varepsilon)]} \|X(t)\| \leq \varepsilon\right) > (1 + \varepsilon)^{-1},$$

which, together with (13), yields the inequality

$$\mathbf{P}(\tau(sG) < \infty) \leq (1 + \varepsilon)\mathbf{P}(\eta_\delta((s - \varepsilon)G) < \infty). \tag{14}$$

The precise asymptotics of the probability on the RHS of (12) were obtained in [8]. It is given in terms of the second rate function $D^{[\delta]}$ for the distribution of the jumps $X(n\delta) - X((n - 1)\delta) \stackrel{d}{=} X(\delta)$ in the random walk $\{X(n\delta)\}_{n \geq 0}$. Recalling the well-known fact that the cumulant of $X(\delta)$ is given by δK , we see that the first rate function $\Lambda^{[\delta]}$ for $X(\delta)$ equals

$$\begin{aligned} \Lambda^{[\delta]}(\alpha) &= \sup_{\lambda \in \Theta_\psi} (\langle \alpha, \lambda \rangle - \delta K(\lambda)) \\ &= \delta \sup_{\lambda \in \Theta_\psi} (\langle \alpha/\delta, \lambda \rangle - K(\lambda)) = \delta \Lambda(\alpha/\delta), \quad \alpha \in \mathbb{R}^d \end{aligned}$$

(cf. (5)). Therefore the second rate function (see (9)) $D^{[\delta]}$ for $X(\delta)$ is

$$D^{[\delta]}(\mathbf{v}) := \inf_{u > 0} u \Lambda^{[\delta]}(\mathbf{v}/u) = \inf_{u > 0} (u\delta) \Lambda(\alpha/(u\delta)) = D(\mathbf{v}), \quad \mathbf{v} \in \mathbb{R}^d.$$

That is, the second rate function for the random walk $\{X(n\delta)\}$ is the same for all $\delta > 0$, which makes perfect sense as one would expect the same asymptotics for the probabilities $\mathbf{P}(\eta_\delta(sG) < \infty)$ for different δ . Hence the respective value $r_G^{[\delta]}$ (see (10)) can easily be seen to be given by δr_G .

Therefore, applying Theorem 1 in [8] to the random walk $\{X(n\delta)\}$ and using notation $A^{[\delta]}$ for the constant A appearing in that theorem for the said random walk, we conclude that, for any $\delta \in (0, \delta(\varepsilon)]$, as $s \rightarrow \infty$,

$$\mathbf{P}(\eta_\delta(sG) < \infty) = A^{[\delta]} s^{-(d-1)/2} e^{-sD(G)} (1 + o(1)), \tag{15}$$

and likewise

$$\mathbf{P}(\eta_\delta((s - \varepsilon)G) < \infty) = A^{[\delta]} (s - \varepsilon)^{-(d-1)/2} e^{-(s-\varepsilon)D(G)} (1 + o(1)). \tag{16}$$

Now from (12) and (14) we see that, as $s \rightarrow \infty$,

$$A^{[\delta]}(1 + o(1)) \leq R(s) := \frac{\mathbf{P}(\tau(sG) < \infty)}{s^{-(d-1)/2} e^{-sD(G)}} \leq \frac{A^{[\delta]}(1 + \varepsilon)e^{\varepsilon D(G)}}{(1 - \varepsilon/s)^{(d-1)/2}} (1 + o(1)).$$

Therefore, setting $\underline{R} := \liminf_{s \rightarrow \infty} R(s)$, $\overline{R} := \limsup_{s \rightarrow \infty} R(s)$, we have

$$A^{[\delta]} \leq \underline{R} \leq \overline{R} \leq (1 + \varepsilon)e^{\varepsilon D(G)} A^{[\delta]}$$

for any $\delta \in (0, \delta(\varepsilon)]$, and hence

$$\limsup_{\delta \rightarrow 0} A^{[\delta]} \leq \underline{R} \leq \overline{R} \leq (1 + \varepsilon)e^{\varepsilon D(G)} \liminf_{\delta \rightarrow 0} A^{[\delta]}.$$

As $\varepsilon > 0$ is arbitrary small, we conclude that there exists $\lim_{\delta \rightarrow 0} A^{[\delta]} =: A_0 \in (0, \infty)$. Therefore there also exists

$$\lim_{s \rightarrow \infty} R(s) = A_0.$$

The theorem is proved. □

Acknowledgements The authors are grateful to the international Mathematical Research Institute MATRIX for hosting and supporting the Mathematics of Risk program during which they obtained the result presented in this note. This work was partially supported by Polish National Science Centre Grant No. 2015/17/B/ST1/01102 (2016–2019) and the ARC Discovery grant DP150102758.

The authors are also grateful to Enkelejd Hashorva who pointed at a bug in the original version of the note.

References

1. Avram, F., Palmowski, Z., Pistorius, M.R.: Exit problem of a two-dimensional risk process from the quadrant: exact and asymptotic results. *Ann. Appl. Probab.* **18**, 2421–2449 (2008)
2. Bertoin, J., Doney, R.A.: Cramér’s estimate for Lévy processes. *Stat. Probab. Lett.* **21**, 363–365 (1994)
3. Borovkov, A.A.: *Probability Theory*, 2nd edn. Springer, London (2013)
4. Borovkov, A.A., Mogulskiĭ, A.A.: The second rate function and the asymptotic problems of renewal and hitting the boundary for multidimensional random walks. *Sib. Math. J.* **37**, 745–782 (1996)
5. Borovkov, A.A., Mogulskiĭ, A.A.: Limit theorems in the boundary hitting problem for a multidimensional random walk. *Sib. Math. J.* **42**, 245–270 (2001)
6. Collamore, J.F.: Hitting probabilities and large deviations. *Ann. Probab.* **24**, 2065–2078 (1996)
7. Palmowski, Z., Pistorius, M.: Cramér asymptotics for finite time first passage probabilities of general Lévy processes. *Stat. Probab. Lett.* **79**, 1752–1758 (2009)
8. Pan, Y., Borovkov, K.: The exact asymptotics of the large deviation probabilities in the multivariate boundary crossing problem. Preprint. arXiv:1708.09605v2 (2017)

Parisian Excursion Below a Fixed Level from the Last Record Maximum of Lévy Insurance Risk Process



Budhi A. Surya

Abstract This paper presents some new results on Parisian ruin under Lévy insurance risk process, where ruin occurs when the process has gone below a fixed level from the last record maximum, also known as the high-water mark or drawdown, for a fixed consecutive periods of time. The law of ruin-time and the position at ruin is given in terms of their joint Laplace transforms. Identities are presented semi-explicitly in terms of the scale function and the law of the Lévy process. They are established using recent developments on fluctuation theory of drawdown of spectrally negative Lévy process. In contrast to the Parisian ruin of Lévy process below a fixed level, ruin under drawdown occurs in finite time with probability one.

1 Introduction

Let $X = \{X_t : t \geq 0\}$ be a spectrally negative Lévy process defined on filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t : t \geq 0\}, \mathbb{P})$, where \mathcal{F}_t is the natural filtration of X satisfying the usual assumptions of right-continuity and completeness. We denote by $\{\mathbb{P}_x, x \in \mathbb{R}\}$ the family of probability measure corresponding to a translation of X s.t. $X_0 = x$, with $\mathbb{P} = \mathbb{P}_0$, and define $\bar{X}_t = \sup_{0 \leq s \leq t} X_s$ the running maximum of X up to time t . The Lévy-Itô sample paths decomposition of the Lévy process is given by

$$X_t = \mu t + \sigma B_t + \int_0^t \int_{\{x < -1\}} x v(dx, ds) + \int_0^t \int_{\{-1 \leq x < 0\}} x(v(dx, ds) - \Pi(dx)ds), \quad (1)$$

B. A. Surya (✉)

School of Mathematics and Statistics, Victoria University of Wellington, Wellington, New Zealand

e-mail: budhi.surya@vuw.ac.nz

where $\mu \in \mathbb{R}$, $\sigma \geq 0$ and $(B_t)_{t \geq 0}$ is standard Brownian motion, whilst $\nu(dx, dt)$ denotes the Poisson random measure associated with the jumps process $\Delta X_t := X_t - X_{t-}$ of X . This Poisson random measure has compensator given by $\Pi(dx)dt$, where Π is the Lévy measure satisfying the integrability condition:

$$\int_{-\infty}^0 (1 \wedge x^2)\Pi(dx) < \infty. \tag{2}$$

We refer to Chap. 2 of [12] for more details on paths decomposition of X .

Due to the absence of positive jumps, it is therefore sensible to define

$$\psi(\lambda) = \frac{1}{t} \log \mathbb{E}\{e^{\lambda X_t}\} = \mu\lambda + \frac{1}{2}\sigma^2\lambda^2 + \int_{(-\infty, 0)} (e^{\lambda x} - 1 - \lambda x \mathbf{1}_{\{x > -1\}})\Pi(dx), \tag{3}$$

which is analytic on $(\text{Im}(\lambda) \leq 0)$. It is easily shown that ψ is zero at the origin, tends to infinity at infinity and is strictly convex. We denote by $\Phi : [0, \infty) \rightarrow [0, \infty)$ the right continuous inverse of the Laplace exponent $\psi(\lambda)$, so that

$$\Phi(\theta) = \sup\{p > 0 : \psi(p) = \theta\} \quad \text{and} \quad \psi(\Phi(\theta)) = \theta \quad \text{for all } \theta \geq 0.$$

It is worth mentioning that under the Esscher transform \mathbb{P}^ν defined by

$$\frac{d\mathbb{P}^\nu}{d\mathbb{P}} \Big|_{\mathcal{F}_t} = e^{\nu X_t - \psi(\nu)t} \quad \text{for all } \nu \geq 0, \tag{4}$$

the Lévy process (X, \mathbb{P}^ν) is still a spectrally negative Lévy process. The Laplace exponent of X under the new measure \mathbb{P}^ν has changed to $\psi_\nu(\lambda)$ given by

$$\psi_\nu(\lambda) = \psi(\lambda + \nu) - \psi(\nu), \quad \text{for } \lambda \geq -\nu. \tag{5}$$

Subsequently, we define by $\Phi_\nu(\theta)$ the largest root of equation $\psi_\nu(\lambda) = \theta$ satisfying

$$\Phi_\nu(\theta) = \Phi(\theta + \psi(\nu)) - \nu.$$

Furthermore, assume that from some reference point of time in the past X has achieved maximum $y > 0$. Define drawdown process $Y = \{Y_t : t \geq 0\}$ of X by

$$Y_t = \overline{X}_t \vee y - X_t, \tag{6}$$

under measure $\mathbb{P}_{y,x}$. Notice that we altered slightly our notation for the probability measure $\mathbb{P}_{y,x}$ to denote the law of X under which at time zero X has current maximum $y \geq x$ and position $x \in \mathbb{R}$. We simply write $\mathbb{P}|_y := \mathbb{P}_{y,0}$ the law of Y under which $Y_0 = y$, and use the notation \mathbb{E}_x , $\mathbb{E}_{y,x}$ and $\mathbb{E}|_y$ to define the corresponding expectation operator to the above probability measures.

Subsequently, we denote by $\mathbb{E}_{y,x}^v$ the expectation under $\mathbb{P}_{y,x}^v$ by which the Lévy process X has the Laplace exponent $\psi_v(\lambda)$ (5). Recall that since X is a Lévy process, it follows that Y is strong Markov.

In recent developments, some results regarding excursion below a (fixed) default level, say zero, of the Lévy process X with fixed duration (Parisian ruin) have been obtained and applied in finance and insurance (e.g. option pricing, corporate finance, optimal dividend, etc). We refer among others to Chesney et al. [6], Francois and Morellec [9], Broadie et al. [4], Dassios and Wu [8], Loeffen et al. [16, 17], Czarna and Palmowski [7] and Landriault et al. [15] and the literature therein for further discussions. In these papers, the excursion takes effect from the first time $T_0^- = \inf\{t > 0 : X_t < 0\}$ the process X has gone below zero under measure \mathbb{P}_x , and default is announced at the first time $\tau_r = \inf\{t > r : (t - \sup\{s < t : X_s > 0\}) > r\}$ the Lévy process has gone below zero for $r > 0$ consecutive periods of time.

In the past decades attention has been paid to find risk protection mechanism against certain financial assets’ outperformance over their last record maximum, also referred to as high-water mark or drawdown, which in practice may affect towards fund managers’ compensation. See, among others, Agarwal et al. [1] and Goetzmann et al. [10] for details. Such risk may be protected against using an insurance contract. In their recent works, Zhang et al. [21], Palmowski and Tumilewicz [19] discussed fair valuation and design of such insurance contract.

Motivated by the above works, we consider a Parisian ruin problem, where ruin occurs when the Lévy risk process X has gone below a fixed level $a > 0$ from its last record maximum (running maximum) $\bar{X}_t \vee y$ for a fixed consecutive periods of time $r \geq 0$. This excursion takes effects from the first time $\tau_a^+ = \inf\{t > 0 : \bar{X}_t \vee y - a > X_t\}$ the process under $\mathbb{P}_{y,x}$ has gone below a fixed level $a > 0$ from the last record maximum $\bar{X}_t \vee y$. Equivalently, this stopping time can be written in terms of the first passage above level $a > 0$ of the drawdown process Y as $\tau_a^+ = \inf\{t > 0 : Y_t > a\}$. Ruin is declared at the first time the process Y has undertaken an excursion above level a for r consecutive periods of time before getting down again below a , i.e.,

$$\tau_r = \inf\{t > r : (t - g_t) \geq r\} \quad \text{with} \quad g_t = \sup\{0 \leq s \leq t : Y_s \leq a\}. \tag{7}$$

Working with the stopping time τ_r (7), we consider the Laplace transforms

$$\mathbb{E}_{y,x}\{e^{-u\tau_r} \mathbf{1}_{\{\tau_r < \infty\}}\} \quad \text{and} \quad \mathbb{E}_{y,x}\{e^{-u\tau_r + vX_{\tau_r}} \mathbf{1}_{\{\tau_r < \infty\}}\}, \tag{8}$$

for $u, v, r \geq 0$ and $y \geq x$. The first quantity gives the law of the ruin time τ_r , whereas the second describes the joint law of the ruin time τ_r and the position at ruin X_{τ_r} .

The rest of this paper is organized as follows. Section 2 presents the main results of this paper. Some preliminary results are presented in Sect. 3. Section 4 discusses the proofs of the main results. Section 5 concludes this paper.

2 Main Results

The results are expressed in terms of the scale function $W^{(u)}(x)$ of X defined by

$$\int_0^\infty e^{-\lambda x} W^{(u)}(x) dx = \frac{1}{\psi(\lambda) - u}, \quad \text{for } \lambda > \Phi(u), \tag{9}$$

with $W^{(u)}(x) = 0$ for $x < 0$. We refer to $W_v^{(u)}$ the scale function under \mathbb{P}^v . Following (9), it is straightforward to check under the new measure \mathbb{P}^v that

$$W_v^{(u)}(x) = e^{-vx} W^{(u+\psi(v))}(x), \tag{10}$$

for all u and v such that $u \geq -\psi(v)$ and $\psi(v) < \infty$. To see this, take Laplace transforms on both sides. We will also use the notation $\overline{W}_v^{(u)}(x)$ to denote $\int_0^x W_v^{(u)}(y) dy$.

It is known following [14] and [5] that, for any $u \geq 0$, the u -scale function $W^{(u)}$ is $C^1(0, \infty)$ if the Lévy measure Π does not have atoms and is $C^2(0, \infty)$ if $\sigma > 0$. For further details on spectrally negative Lévy process, we refer to Chap. 6 of Bertoin [3] and Chap. 8 of Kyprianou [12]. Some examples of Lévy processes for which $W^{(u)}$ are available in explicit form are given by Kuznetsov et al. [11]. In any case, it can be computed by numerically inverting (9), see e.g. Surya [20].

In the sequel below, we will use the notation $\Omega_\epsilon^{(u)}(x, t)$ defined by

$$\Omega_\epsilon^{(u)}(x, t) = \int_\epsilon^\infty W^{(u)}(z + x - \epsilon) \frac{z}{t} \mathbb{P}\{X_t \in dz\}, \quad \text{for } \epsilon \geq 0,$$

and define its partial derivative w.r.t x , $\frac{\partial}{\partial x} \Omega_\epsilon^{(u)}(x, t)$, by $\Lambda_\epsilon^{(u)}(x, t)$, i.e.,

$$\Lambda_\epsilon^{(u)}(x, t) = \int_\epsilon^\infty W^{(u)'}(z + x - \epsilon) \frac{z}{t} \mathbb{P}\{X_t \in dz\}.$$

For convenience, we write $\Omega^{(u)}(x, t) = \Omega_0^{(u)}(x, t)$ and $\Lambda^{(u)}(x, t) = \Lambda_0^{(u)}(x, t)$.

We denote by $\Omega_v^{(u)}$ the role of $\Omega^{(u)}$ under change of measure \mathbb{P}^v , i.e.,

$$\Omega_v^{(u)}(x, t) := \int_0^\infty W_v^{(u)}(z + x) \frac{z}{t} \mathbb{P}^v\{X_t \in dz\}, \tag{11}$$

similarly defined for $\Lambda_v^{(u)}(x, t)$. Using (4), we can rewrite $\Omega_v^{(u)}(x, t)$ as follows

$$\Omega_v^{(u)}(x, t) = e^{-vx} e^{-\psi(v)t} \Omega^{(u+\psi(v))}(x, t). \tag{12}$$

The main result concerning the Laplace transform (8) is given below.

Theorem 1 Define $z = y - x$, with $y \geq x$. For $a > 0$ and $u, r \geq 0$, the Laplace transform of τ_r is given by

$$\mathbb{E}_{y,x} \left\{ e^{-u\tau_r} \mathbf{1}_{\{\tau_r < \infty\}} \right\} = e^{-ur} \left\{ 1 + u \left[\overline{W}^{(u)}(a - z) - \frac{\Omega^{(u)}(a - z, r)}{\Lambda^{(u)}(a, r)} W^{(u)}(a) + \int_0^r \left(\Omega^{(u)}(a - z, t) - \frac{\Omega^{(u)}(a - z, r)}{\Lambda^{(u)}(a, r)} \Lambda^{(u)}(a, t) \right) dt \right] \right\}. \tag{13}$$

By inserting $u = 0$ in (13), we see that in contrary to the Parisian ruin probability under the Lévy process X , see e.g. [16], we have the following result.

Corollary 1 For $y \geq x$ and $r \geq 0$, $\mathbb{P}_{y,x} \{ \tau_r < \infty \} = 1$.

Following the result of Theorem 1 and applying Esscher transform of measure, the joint law of ruin-time τ_r and the position at ruin X_{τ_r} is given below.

Proposition 1 Define $z = y - x$, with $y \geq x$, and $p = u - \psi(v)$, with $u \geq 0$ and v such that $\psi(v) < \infty$. For $a > 0$ and $r \geq 0$, the joint Laplace transform of τ_r and X_{τ_r} is given by

$$\mathbb{E}_{y,x} \left\{ e^{-u\tau_r + vX_{\tau_r}} \mathbf{1}_{\{\tau_r < \infty\}} \right\} = e^{-pr} e^{vx} \left\{ 1 + p \left[\overline{W}_v^{(p)}(a - z) - \frac{\Omega_v^{(p)}(a - z, r)}{\Lambda_v^{(p)}(a, r)} W_v^{(p)}(a) + \int_0^r \left(\Omega_v^{(p)}(a - z, t) - \frac{\Omega_v^{(p)}(a - z, r)}{\Lambda_v^{(p)}(a, r)} \Lambda_v^{(p)}(a, t) \right) dt \right] \right\}. \tag{14}$$

3 Preliminaries

Before we prove the main results, we devote this section to some preliminary results required to establish (13)–(14); in particular, Theorem 1 on the Laplace transform of τ_r . By spatial homogeneity of the sample paths of X , we establish Theorem 1 under the measure $\mathbb{P}|_y$. To begin with, we define for $a > 0$ stopping times:

$$\tau_a^+ = \inf\{t > 0 : Y_t > a\} \quad \text{and} \quad \tau_a^- = \inf\{t > 0 : Y_t < a\} \quad \text{under } \mathbb{P}|_y. \tag{15}$$

Due to the absence of positive jumps, we have by the strong Markov property of X that τ_a^- can equivalently be rewritten as $\tau_a^- = \inf\{t > 0 : Y_t \leq a\}$ and that

$$\mathbb{E}|_y \left\{ e^{-\theta\tau_a^-} \right\} = e^{-\Phi(\theta)(y-a)}. \tag{16}$$

This is due to the fact that $\tau_a^- < \tau_{\{0\}}$ a.s., with $\tau_{\{0\}} = \inf\{t > 0 : Y_t = 0\}$, and that $\{Y_t, t \leq \tau_{\{0\}}, \mathbb{P}_{y,x}\} = \{-X_t, t \leq T_0^+, \mathbb{P}_{-z}\}$, with $z = y - x$, where

$T_a^+ = \inf\{t > 0 : X_t \geq a\}$, $a \geq 0$. We refer to Avram et al. [2] and Mijatović and Pistorius [18].

In the derivation of the main results (13)–(14), we will also frequently apply Kendall’s identity (see e.g. Corollary VII.3 in [3]), which relates the distribution $\mathbb{P}\{X_t \in dx\}$ of a spectrally negative Lévy process X to the distribution $\mathbb{P}\{T_x^+ \in dt\}$ of its first passage time T_x^+ above $x > 0$ under \mathbb{P} . This identity is given by

$$t\mathbb{P}\{T_x^+ \in dt\}dx = x\mathbb{P}\{X_t \in dx\}dt. \tag{17}$$

To establish our main results, we need to recall the following identities.

Lemma 1 Define $s = y - x$, with $y \geq x$. For $a > 0$, $u \geq 0$ and v such that $\psi(v) < \infty$, the joint Laplace transform of τ_a^+ and $Y_{\tau_a^+}$ is given by

$$\begin{aligned} \mathbb{E}_{y,x}\{e^{-u\tau_a^+ - vY_{\tau_a^+}} \mathbf{1}_{\{\tau_a^+ < \infty\}}\} &= (\psi(v) - u)e^{-vs} \int_{a-s}^{\infty} e^{-vz} W^{(u)}(z) dz \\ &+ \frac{W^{(u)}(a-s)}{W^{(u)'(a)}} \left[(\psi(v) - u)e^{-va} W^{(u)}(a) - v(\psi(v) - u) \int_a^{\infty} e^{-vz} W^{(u)}(z) dz \right]. \end{aligned} \tag{18}$$

The identity (18) is due to Theorem 1 in Avram et al. [2] taking account of (9)–(10).

Corollary 2 Define $s = y - x$, with $y \geq x$. For $a > 0$ and $u, \theta \geq 0$,

$$\begin{aligned} \mathbb{E}_{y,x}\{e^{-u\tau_a^+ - \Phi(\theta)Y_{\tau_a^+}} \mathbf{1}_{\{\tau_a^+ < \infty\}}\} &= (\theta - u)e^{-\Phi(\theta)s} \int_{a-s}^{\infty} e^{-\Phi(\theta)z} W^{(u)}(z) dz \\ &- \frac{W^{(u)}(a-s)}{W^{(u)'(a)}} \left[(u - \theta)e^{-\Phi(\theta)a} W^{(u)}(a) - (u - \theta)\Phi(\theta) \int_a^{\infty} e^{-\Phi(\theta)z} W^{(u)}(z) dz \right]. \end{aligned} \tag{19}$$

Proof The result follows from inserting $v = \Phi(\theta)$ in Eq. (18) and taking account that $\psi(\Phi(\theta)) = \theta$, and $\int_0^x e^{-vz} W^{(u)}(z) dz = \frac{1}{(\psi(v)-u)} - \int_x^{\infty} e^{-vz} W^{(u)}(z) dz$. \square

Along with Lemma 1 and Corollary 2, the three results below are used when applying inverse Laplace transforms to get the main results (13)–(14).

Lemma 2 For a given $\theta > 0$ and α such that $\alpha < \Phi(\theta)$, we have for $y \in \mathbb{R}$,

$$\int_0^{\infty} e^{-\theta t} e^{-\alpha y} \int_y^{\infty} e^{\alpha z} \frac{z}{t} \mathbb{P}\{X_t \in dz\} dt = \frac{e^{-\Phi(\theta)y}}{(\Phi(\theta) - \alpha)}. \tag{20}$$

$$\int_0^{\infty} e^{-\theta t} e^{-\alpha y} \int_y^{\infty} e^{\alpha z} \int_0^t \frac{z}{u} \mathbb{P}\{X_u \in dz\} dudt = \frac{e^{-\Phi(\theta)y}}{\theta(\Phi(\theta) - \alpha)}. \tag{21}$$

$$\int_0^{\infty} W^{(u)}(z) \frac{z}{t} \mathbb{P}\{X_t \in dz\} = e^{ut}, \quad \text{for } u \geq 0 \text{ and } t > 0. \tag{22}$$

The results above are slightly generalizations of those given in [16] and can be proved in similar fashion of [16] using Kendall's identity (17) and Tonelli.

4 Proof of the Main Results

4.1 Proof of Theorem 1

The proof is established for the case where X has paths of bounded and unbounded variation. To deal with unbounded variation case, we will use a limiting argument similar to the one employed in [16, 17] and adjust the ruin time (7) accordingly. For this reason, we introduce for $\epsilon \geq 0$ the stopping time τ_r^ϵ defined by

$$\tau_r^\epsilon = \inf\{t > r : (t - g_t^\epsilon) \geq r\} \text{ with } g_t^\epsilon := \sup\{s < t : Y_s \leq a - \epsilon\}.$$

This stopping time represents the first time that the Lévy insurance risk process X has spent a fixed $r > 0$ units of time consecutively below pre-specified level $a > 0$ from its running maximum $\bar{X}_t \vee y$ ending before X getting back up again to a level $a - \epsilon \geq 0$ below the running maximum. Note that $\tau_r = \tau_r^0$.

By spatial homogeneity of X , the proof is given under measure $\mathbb{P}_{|y}$ by which X starts at point zero and has current maximum y . We have for any $y > a$ that

$$\mathbb{E}_{|y}\{e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}}\} = e^{-ur} \mathbb{P}_{|y}\{\tau_{a-\epsilon}^- > r\} + \mathbb{E}_{|y}\{e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty, \tau_{a-\epsilon}^- \leq r\}}\}.$$

By the strong Markov property of the drawdown process Y (6), the second expectation can be worked out using tower property of conditional expectation,

$$\begin{aligned} \mathbb{E}_{|y}\{e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty, \tau_{a-\epsilon}^- \leq r\}}\} &= \mathbb{E}_{|y}\left\{\mathbb{E}\left\{e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty, \tau_{a-\epsilon}^- \leq r\}} \mid \mathcal{F}_{\tau_{a-\epsilon}^-}\right\}\right\} \\ &= \mathbb{E}_{|y}\left\{e^{-u\tau_{a-\epsilon}^-} \mathbf{1}_{\{\tau_{a-\epsilon}^- \leq r\}} \mathbb{E}_{|Y_{\tau_{a-\epsilon}^-}}\left\{e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}}\right\}\right\} \\ &= \mathbb{E}_{|y}\left\{e^{-u\tau_{a-\epsilon}^-} \mathbf{1}_{\{\tau_{a-\epsilon}^- \leq r\}}\right\} \mathbb{E}_{|a-\epsilon}\left\{e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}}\right\}, \end{aligned}$$

where the last equality is due to the absence of positive jumps of X . Hence,

$$\begin{aligned} \mathbb{E}_{|y}\{e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}}\} &= e^{-ur} (1 - \mathbb{P}_{|y}\{\tau_{a-\epsilon}^- \leq r\}) \\ &\quad + \mathbb{E}_{|y}\{e^{-u\tau_{a-\epsilon}^-} \mathbf{1}_{\{\tau_{a-\epsilon}^- \leq r\}}\} \mathbb{E}_{|a-\epsilon}\{e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}}\}. \end{aligned} \tag{23}$$

Following the above, for $y \leq a$ we have by strong Markov property of Y that

$$\begin{aligned}
 \mathbb{E}_{|y} \left\{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \right\} &= \mathbb{E}_{|y} \left\{ \mathbb{E} \left\{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \mid \mathcal{F}_{\tau_a^+} \right\} \right\} \\
 &= \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \left(e^{-ur} (1 - \mathbb{P}_{|Y_{\tau_a^+}} \{ \tau_{a-\epsilon}^- \leq r \}) \right) \right\} \\
 &\quad + \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{E}_{|Y_{\tau_a^+}} \left\{ e^{-u\tau_{a-\epsilon}^-} \mathbf{1}_{\{\tau_{a-\epsilon}^- \leq r\}} \right\} \right\} \mathbb{E}_{|a-\epsilon} \left\{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \right\} \\
 &= e^{-ur} \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \right\} - e^{-ur} \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{P}_{|Y_{\tau_a^+}} \{ \tau_{a-\epsilon}^- \leq r \} \right\} \\
 &\quad + \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{E}_{|Y_{\tau_a^+}} \left\{ e^{-u\tau_{a-\epsilon}^-} \mathbf{1}_{\{\tau_{a-\epsilon}^- \leq r\}} \right\} \right\} \mathbb{E}_{|a-\epsilon} \left\{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \right\}.
 \end{aligned} \tag{24}$$

The first expectation in the last equality of (24) can be worked out in terms of the scale function $W^{(u)}(x)$ using identity (18), whereas the second and the third expectations are given by the following propositions. To establish the results, we denote throughout by \mathbf{e}_θ exponential random time with parameter θ , independent of X .

Proposition 2 For given $u, r, \epsilon \geq 0$ and $a > 0$, we have for any $y \geq 0$ that

$$\begin{aligned}
 \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{P}_{|Y_{\tau_a^+}} \{ \tau_{a-\epsilon}^- \leq r \} \right\} &= \Omega_\epsilon^{(u)}(a - y, r) - u \int_0^r \Omega_\epsilon^{(u)}(a - y, t) dt \\
 &\quad - \frac{W^{(u)}(a - y)}{W^{(u)'(a)}} \left(\Lambda_\epsilon^{(u)}(a, r) - u \int_0^r \Lambda_\epsilon^{(u)}(a, t) dt \right).
 \end{aligned} \tag{25}$$

Proof On recalling (16), we have by Tonelli, Lemma 1 and Corollary 2,

$$\begin{aligned}
 \int_0^\infty dr e^{-\theta r} \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{P}_{|Y_{\tau_a^+}} \{ \tau_{a-\epsilon}^- \leq r \} \right\} \\
 &= \frac{1}{\theta} \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{P}_{|Y_{\tau_a^+}} \{ \mathbf{e}_\theta \geq \tau_{a-\epsilon}^- \} \right\} \\
 &= \frac{1}{\theta} e^{\Phi(\theta)(a-\epsilon)} \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+ - \Phi(\theta)Y_{\tau_a^+}} \mathbf{1}_{\{\tau_a^+ < \infty\}} \right\}.
 \end{aligned} \tag{26}$$

Furthermore, observe following the result of Corollary 2 that for $\theta > u$ we have

$$\begin{aligned}
 \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+ - \Phi(\theta)Y_{\tau_a^+}} \mathbf{1}_{\{\tau_a^+ < \infty\}} \right\} &= \frac{(\theta - u)}{\Phi(\theta)} e^{-\Phi(\theta)a} W^{(u)}(a - y) \\
 &\quad + \frac{(\theta - u)}{\Phi(\theta)} e^{-\Phi(\theta)a} \int_0^\infty e^{-\Phi(\theta)z} W^{(u)'(z + a - y)} dz \\
 &\quad - \frac{W^{(u)}(a - y)}{W^{(u)'(a)}} \left[(\theta - u) e^{-\Phi(\theta)a} \int_0^\infty e^{-\Phi(\theta)z} W^{(u)'(z + a)} dz \right].
 \end{aligned}$$

Define $\Gamma(x, r) = \int_x^\infty \frac{z}{r} \mathbb{P}\{X_r \in dz\}$. Following the above, we have from (26) that

$$\begin{aligned} & \int_0^\infty dr e^{-\theta r} \mathbb{E}_y \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{P}_{Y_{\tau_a^+}} \left\{ \tau_{a-\epsilon}^- \leq r \right\} \right\} \\ &= \frac{1}{\theta} \mathbb{E}_y \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{P}_{Y_{\tau_a^+}} \left\{ \mathbf{e}\theta \geq \tau_{a-\epsilon}^- \right\} \right\} \\ &= \frac{(\theta - u)}{\theta \Phi(\theta)} e^{-\Phi(\theta)\epsilon} W^{(u)}(a - y) \\ &\quad + \frac{(\theta - u)}{\theta \Phi(\theta)} e^{-\Phi(\theta)\epsilon} \int_0^\infty e^{-\Phi(\theta)z} W^{(u)'}(z + a - y) dz \\ &\quad - \frac{W^{(u)}(a - y)}{\theta W^{(u)'}(a)} \left[(\theta - u) e^{-\Phi(\theta)\epsilon} \int_0^\infty e^{-\Phi(\theta)z} W^{(u)'}(z + a) dz \right]. \end{aligned}$$

Next, recall following (20)–(21), (16) and the Kendall’s identity (17) that

$$\begin{aligned} & \left(\frac{1}{\Phi(\theta)} - \frac{u}{\theta \Phi(\theta)} \right) e^{-\Phi(\theta)x} = \int_0^\infty dr e^{-\theta r} \left(\Gamma(x, r) - u \int_0^r \Gamma(x, t) dt \right) \\ & \int_0^\infty \left(1 - \frac{u}{\theta} \right) e^{-\Phi(\theta)(z+\epsilon)} W^{(u)'}(z + a) dz = \int_0^\infty dr e^{-\theta r} \left(\Lambda_\epsilon^{(u)}(a, r) - u \int_0^r \Lambda_\epsilon^{(u)}(a, t) dt \right). \end{aligned}$$

Moreover, by applying integration by part we have after some calculations that

$$\int_0^\infty dz W^{(u)'}(z + x) \Gamma(z + \epsilon, t) = \Omega_\epsilon^{(u)}(x, t) - W^{(u)}(x) \Gamma(\epsilon, t). \tag{27}$$

The claim in (25) is established following the above and by Tonelli and Laplace inversion (noting that both sides of (26) is right-continuous in r) to (26). \square

Proposition 3 For given $u, r, \epsilon \geq 0$ and $a > 0$, we have for any $y \geq 0$ that

$$\begin{aligned} & \mathbb{E}_y \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{E}_{Y_{\tau_a^+}} \left\{ e^{-u\tau_{a-\epsilon}^-} \mathbf{1}_{\{\tau_{a-\epsilon}^- \leq r\}} \right\} \right\} \\ &= e^{-ur} \left(\Omega_\epsilon^{(u)}(a - y, r) - \frac{W^{(u)}(a - y)}{W^{(u)'}(a)} \Lambda_\epsilon^{(u)}(a, r) \right). \end{aligned} \tag{28}$$

Proof On recalling (16), we have by Tonelli, Lemma 1 and Corollary 2,

$$\begin{aligned} & \int_0^\infty dr e^{-\theta r} \mathbb{E}_y \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{E}_{Y_{\tau_a^+}} \left\{ e^{-u\tau_{a-\epsilon}^-} \mathbf{1}_{\{\tau_{a-\epsilon}^- \leq r\}} \right\} \right\} \\ &= \frac{1}{\theta} \mathbb{E}_y \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{E}_{Y_{\tau_a^+}} \left\{ e^{-u\tau_{a-\epsilon}^-} \mathbf{1}_{\{\mathbf{e}\theta \geq \tau_{a-\epsilon}^- \}} \right\} \right\} \\ &= \frac{1}{\theta} e^{\Phi(\theta+u)(a-\epsilon)} \mathbb{E}_y \left\{ e^{-u\tau_a^+ - \Phi(\theta+u)Y_{\tau_a^+}} \mathbf{1}_{\{\tau_a^+ < \infty\}} \right\}. \end{aligned} \tag{29}$$

From Corollary 2, the expectation on the right hand side is given by

$$\begin{aligned} \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+ - \Phi(\theta+u)Y_{\tau_a^+}} \mathbf{1}_{\{\tau_a^+ < \infty\}} \right\} &= \frac{\theta}{\Phi(\theta+u)} e^{-\Phi(\theta+u)a} W^{(u)}(a-y) \\ &+ \frac{\theta}{\Phi(\theta+u)} e^{-\Phi(\theta+u)a} \int_0^\infty e^{-\Phi(\theta+u)z} W^{(u)'}(z+a-y) dz \\ &- \frac{W^{(u)}(a-y)}{W^{(u)'}(a)} \theta e^{-\Phi(\theta+u)a} \int_0^\infty e^{-\Phi(\theta+u)z} W^{(u)'}(z+a) dz. \end{aligned}$$

Following the above, we have from the Laplace transform (29) that

$$\begin{aligned} \int_0^\infty dr e^{-\theta r} \mathbb{E}_{|y} \left\{ e^{-u\tau_a^+} \mathbf{1}_{\{\tau_a^+ < \infty\}} \mathbb{E}_{|Y_{\tau_a^+}} \left\{ e^{-u\tau_{a-\epsilon}^-} \mathbf{1}_{\{\tau_{a-\epsilon}^- \leq r\}} \right\} \right\} \\ = \frac{1}{\Phi(\theta+u)} e^{-\Phi(\theta+u)\epsilon} W^{(u)}(a-y) \\ + \frac{1}{\Phi(\theta+u)} e^{-\Phi(\theta+u)\epsilon} \int_0^\infty e^{-\Phi(\theta+u)z} W^{(u)'}(z+a-y) dz \quad (30) \\ - \frac{W^{(u)}(a-y)}{W^{(u)'}(a)} e^{-\Phi(\theta+u)\epsilon} \int_0^\infty e^{-\Phi(\theta+u)z} W^{(u)'}(z+a) dz. \end{aligned}$$

Moreover, following (16), we have by applying Kendall’s identity and (20)

$$\begin{aligned} \frac{1}{\Phi(\theta+u)} e^{-\Phi(\theta+u)x} &= \int_0^\infty dr e^{-\theta r} e^{-ur} \Gamma(x, r) \\ \int_0^\infty e^{-\Phi(\theta+u)(z+\epsilon)} W^{(u)'}(z+a) dz &= \int_0^\infty dr e^{-\theta r} e^{-ur} \Lambda_\epsilon^{(u)}(a, r). \end{aligned}$$

The claim (28) is justified using the above and (27) and by Tonelli and Laplace inversion of (30)—noting that both sides of (30) is right-continuous in r . \square

From the above two propositions, we have following (24) and (18) that

$$\begin{aligned} \mathbb{E}_{|y} \left\{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \right\} &= e^{-ur} \left[1 + u \overline{W}^{(u)}(a-y) - u \frac{W^{(u)}(a)}{W^{(u)'}(a)} W^{(u)}(a-y) \right] \\ &- e^{-ur} \left[\Omega_\epsilon^{(u)}(a-y, r) - u \int_0^r \Omega_\epsilon^{(u)}(a-y, t) dt \right] \quad (31) \\ &- \frac{W^{(u)}(a-y)}{W^{(u)'}(a)} \left(\Lambda_\epsilon^{(u)}(a, r) - u \int_0^r \Lambda_\epsilon^{(u)}(a, t) dt \right) \\ &+ e^{-ur} \left(\Omega_\epsilon^{(u)}(a-y, r) - \frac{W^{(u)}(a-y)}{W^{(u)'}(a)} \Lambda_\epsilon^{(u)}(a, r) \right) \mathbb{E}_{|a-\epsilon} \left\{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \right\}. \end{aligned}$$

We arrive at our claim (13) once the expectation on the right hand side is found. For this purpose, set $y = a - \epsilon$ on both sides of the above equation to get

$$\begin{aligned} & \left[1 - e^{-ur} \left(\Omega_\epsilon^{(u)}(\epsilon, r) - \frac{W^{(u)}(\epsilon)}{W^{(u)'(a)}} \Lambda_\epsilon^{(u)}(a, r) \right) \right] \mathbb{E}_{|a-\epsilon} \left\{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \right\} \\ &= e^{-ur} \left[1 + u \overline{W}^{(u)}(\epsilon) - u \frac{W^{(u)}(a)}{W^{(u)'(a)}} W^{(u)}(\epsilon) - \Omega_\epsilon^{(u)}(\epsilon, r) + u \int_0^r \Omega_\epsilon^{(u)}(\epsilon, t) dt \right. \\ & \quad \left. + \frac{W^{(u)}(\epsilon)}{W^{(u)'(a)}} \left(\Lambda_\epsilon^{(u)}(a, r) - u \int_0^r \Lambda_\epsilon^{(u)}(a, t) dt \right) \right]. \end{aligned} \quad (32)$$

However, on account of (22), we can rewrite the terms $\Omega_\epsilon^{(u)}(\epsilon, t)$ as follows

$$\Omega_\epsilon^{(u)}(\epsilon, t) = e^{ut} - \int_0^\epsilon W^{(u)}(z) \frac{z}{t} \mathbb{P}\{X_t \in dz\},$$

from which the Eq. (32) simplifies further after some calculations to

$$\begin{aligned} & \left[\int_0^\epsilon W^{(u)}(z) \frac{z}{r} \mathbb{P}\{X_r \in dz\} + \frac{W^{(u)}(\epsilon)}{W^{(u)'(a)}} \Lambda_\epsilon^{(u)}(a, r) \right] \mathbb{E}_{|a-\epsilon} \left\{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \right\} \\ &= \int_0^\epsilon W^{(u)}(z) \frac{z}{r} \mathbb{P}\{X_r \in dz\} + \frac{W^{(u)}(\epsilon)}{W^{(u)'(a)}} \Lambda_\epsilon^{(u)}(a, r) + u \left[\overline{W}^{(u)}(\epsilon) - \frac{W^{(u)}(a)}{W^{(u)'(a)}} W^{(u)}(\epsilon) \right. \\ & \quad \left. - \int_0^r dt \left(\int_0^\epsilon W^{(u)}(z) \frac{z}{t} \mathbb{P}\{X_t \in dz\} + \frac{W^{(u)}(\epsilon)}{W^{(u)'(a)}} \Lambda_\epsilon^{(u)}(a, t) \right) \right], \end{aligned}$$

or equivalently, we obtain after dividing both sides of the equation by $W^{(u)}(\epsilon)$,

$$\begin{aligned} & \mathbb{E}_{|a-\epsilon} \left\{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \right\} \\ &= 1 + u \frac{\left[\frac{\overline{W}^{(u)}(\epsilon)}{W^{(u)}(\epsilon)} - \frac{W^{(u)}(a)}{W^{(u)'(a)}} - \int_0^r dt \left(\int_0^\epsilon \frac{W^{(u)}(z)}{W^{(u)}(\epsilon)} \frac{z}{t} \mathbb{P}\{X_t \in dz\} + \frac{\Lambda_\epsilon^{(u)}(a, t)}{W^{(u)'(a)}} \right) \right]}{\left(\int_0^\epsilon \frac{W^{(u)}(z)}{W^{(u)}(\epsilon)} \frac{z}{r} \mathbb{P}\{X_r \in dz\} + \frac{\Lambda_\epsilon^{(u)}(a, r)}{W^{(u)'(a)}} \right)}. \end{aligned}$$

Using this result and putting it back in the Eq. (31), we arrive at

$$\begin{aligned}
 e^{ur} \mathbb{E}_{|y} \{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \} &= 1 + u \overline{W}^{(u)}(a - y) - u \frac{W^{(u)}(a)}{W^{(u)'(a)}} W^{(u)}(a - y) \\
 &\quad + u \int_0^r dt \left(\Omega_\epsilon^{(u)}(a - y, t) - \frac{W^{(u)}(a - y)}{W^{(u)'(a)}} \Lambda_\epsilon^{(u)}(a, t) \right) \\
 &\quad + u \frac{\left[\frac{\overline{W}^{(u)}(\epsilon)}{W^{(u)}(\epsilon)} - \frac{W^{(u)}(a)}{W^{(u)'(a)}} - \int_0^r dt \left(\int_0^\epsilon \frac{W^{(u)}(z)}{W^{(u)}(\epsilon)} \frac{z}{t} \mathbb{P}\{X_t \in dz\} + \frac{\Lambda_\epsilon^{(u)}(a, t)}{W^{(u)'(a)}} \right) \right]}{\left(\int_0^\epsilon \frac{W^{(u)}(z)}{W^{(u)}(\epsilon)} \frac{z}{r} \mathbb{P}\{X_r \in dz\} + \frac{\Lambda_\epsilon^{(u)}(a, r)}{W^{(u)'(a)}} \right)} \\
 &\quad \times \left(\Omega_\epsilon^{(u)}(a - y, r) - \frac{W^{(u)}(a - y)}{W^{(u)'(a)}} \Lambda_\epsilon^{(u)}(a, r) \right). \tag{33}
 \end{aligned}$$

We now want to compute the limit as $\epsilon \downarrow 0$ of (33). In order to do this, recall by the spatial homogeneity that $\mathbb{P}_{|y} \{ \tau_r^\epsilon \leq t \} = \mathbb{P}_{|y+\epsilon} \{ \tau_r \leq t \}$ and therefore by the right-continuity of the map $y \rightarrow \mathbb{P}_{|y} \{ \tau_r \leq t \}$, we have $\mathbb{P}_{|y} \{ \tau_r \leq t \} = \lim_{\epsilon \downarrow 0} \mathbb{P}_{|y} \{ \tau_r^\epsilon \leq t \}$. Hence, by weak convergence theorem the Laplace transform of $\mathbb{P}_{|y} \{ \tau_r^\epsilon \leq t \}$ converges as $\epsilon \downarrow 0$ to that of $\mathbb{P}_{|y} \{ \tau_r \leq t \}$, i.e., $\lim_{\epsilon \downarrow 0} \mathbb{E}_{|y} \{ e^{-u\tau_r^\epsilon} \mathbf{1}_{\{\tau_r^\epsilon < \infty\}} \} = \mathbb{E}_{|y} \{ e^{-u\tau_r} \mathbf{1}_{\{\tau_r < \infty\}} \}$.

We consider two cases: $W^{(u)}(0+) > 0$ (X has paths of bounded variation) and $W^{(u)}(0+) = 0$ (X has unbounded variation). For the case $W^{(u)}(0+) > 0$,

$$\begin{aligned}
 e^{ur} \mathbb{E}_{|y} \{ e^{-u\tau_r} \mathbf{1}_{\{\tau_r < \infty\}} \} &= 1 + u \overline{W}^{(u)}(a - y) - u \frac{W^{(u)}(a)}{W^{(u)'(a)}} W^{(u)}(a - y) \\
 &\quad + u \int_0^r dt \left(\Omega^{(u)}(a - y, t) - \frac{W^{(u)}(a - y)}{W^{(u)'(a)}} \Lambda^{(u)}(a, t) \right) \\
 &\quad - u \left(\frac{W^{(u)}(a)}{\Lambda^{(u)}(a, r)} + \int_0^r \frac{\Lambda^{(u)}(a, t)}{\Lambda^{(u)}(a, r)} dt \right) \left(\Omega^{(u)}(a - y, r) - \frac{W^{(u)}(a - y)}{W^{(u)'(a)}} \Lambda^{(u)}(a, r) \right), \tag{34}
 \end{aligned}$$

which after some further calculations simplifies to the main result (13).

For the case $W^{(u)}(0+) = 0$, we have after applying integration by parts that

$$\int_0^\epsilon \frac{W^{(u)}(z)}{W^{(u)}(\epsilon)} \frac{z}{r} \mathbb{P}\{X_r \in dz\} = \int_0^\epsilon \frac{z}{r} \mathbb{P}\{X_r \in dz\} - \int_0^\epsilon dz \frac{W^{(u)'(z)}}{W^{(u)}(\epsilon)} \int_0^z \frac{w}{r} \mathbb{P}\{X_r \in dw\}.$$

Therefore, by employing l'Hôpital rule we obtain

$$\lim_{\epsilon \downarrow 0} \int_0^\epsilon \frac{W^{(u)}(z)}{W^{(u)}(\epsilon)} \frac{z}{r} \mathbb{P}\{X_r \in dz\} = \lim_{\epsilon \downarrow 0} \frac{W^{(u)'(\epsilon)} \int_0^\epsilon \frac{z}{r} \mathbb{P}\{X_r \in dz\}}{W^{(u)'(\epsilon)}} = 0.$$

The claim is established once we show that $\lim_{\epsilon \downarrow 0} \frac{\overline{W}^{(u)}(\epsilon)}{W^{(u)}(\epsilon)} = \lim_{\epsilon \downarrow 0} \frac{W^{(u)}(\epsilon)}{W^{(u)'(\epsilon)}(\epsilon)} = 0$. This turns out to be the case when X has paths of unbounded variation since $W^{(u)'(0+)} = 2/\sigma^2$ if $\sigma \neq 0$ and is equal to ∞ if $\sigma = 0$. See for instance Lemma 4.4. in Kyprianou and Surya [13]. On account of these results, we arrive at the identity (34), which after some further calculations simplifies to the main result (13). \square

We have shown that (13) holds for $z \leq a$. We now prove that (13) holds for $z > a$. For this purpose, recall that under measure $\mathbb{P}|_y$, with $y > a$, $\tau_a^+ = 0$ a.s. On account of the fact $W^{(u)}(x) = 0$ for $x < 0$, we have from (25) and (28) for $\epsilon = 0$ and $y > a$,

$$\mathbb{P}|_y\{\tau_a^- \leq r\} = \Omega^{(u)}(a - y, r) - u \int_0^r \Omega^{(u)}(a - y, t) dt. \tag{35}$$

$$\mathbb{E}|_y\{e^{-u\tau_a^-} \mathbf{1}_{\{\tau_a^- \leq r\}}\} = e^{-ur} \Omega^{(u)}(a - y, r). \tag{36}$$

These identities can be proved by Kendall’s identity, Tonelli, (16) and Laplace inversion taking account for $x < 0$, $\int_0^\infty e^{-\theta t} \Omega^{(u)}(x, t) dt = e^{\Phi(\theta)x}/(\theta - u)$, $\theta > u$. Indeed,

$$\begin{aligned} \int_0^\infty e^{-\theta t} \Omega^{(u)}(x, t) dt &= \int_0^\infty e^{-\theta t} \int_0^\infty W^{(u)}(z + x) \frac{z}{t} \mathbb{P}\{X_t \in dz\} dt \\ &= \int_0^\infty e^{-\theta t} \int_0^\infty W^{(u)}(z + x) \mathbb{P}\{T_z^+ \in dt\} dz \\ &= \int_0^\infty W^{(u)}(z + x) \int_0^\infty e^{-\theta t} \mathbb{P}\{T_z^+ \in dt\} dz \\ &= \int_0^\infty W^{(u)}(z + x) e^{-\Phi(\theta)z} dz \\ &= e^{\Phi(\theta)x} \int_x^\infty e^{-\Phi(\theta)z} W^{(u)}(z) dz. \end{aligned} \tag{37}$$

Starting from Eq. (23) with $\epsilon = 0$, we obtain following identities (35)–(36),

$$\begin{aligned} \mathbb{E}|_y\{e^{-u\tau_r} \mathbf{1}_{\{\tau_r < \infty\}}\} &= e^{-ur} \left[1 + u \int_0^r \Omega^{(u)}(a - y, t) dt - \Omega^{(u)}(a - y, r) \right] \\ &\quad + e^{-ur} \Omega^{(u)}(a - y, r) \mathbb{E}|_a\{e^{-u\tau_r} \mathbf{1}_{\{\tau_r < \infty\}}\}. \end{aligned} \tag{38}$$

The expression for $\mathbb{E}|_a\{e^{-u\tau_r} \mathbf{1}_{\{\tau_r < \infty\}}\}$ is given by setting $z = a$ in (13):

$$\mathbb{E}|_a\{e^{-u\tau_r} \mathbf{1}_{\{\tau_r < \infty\}}\} = 1 - u \left(\frac{W^{(u)}(a)}{\Lambda^{(u)}(a, r)} + \int_0^r \frac{\Lambda^{(u)}(a, t)}{\Lambda^{(u)}(a, r)} dt \right), \tag{39}$$

where we have used in the calculation above the fact that $\Omega^{(u)}(0, t) = e^{ut}$, see Eq. (22). By inserting (38) in (37) we obtain after some further calculations that

$$\mathbb{E}_{|y} \{ e^{-u\tau_r} \mathbf{1}_{\{\tau_r < \infty\}} \} = e^{-ur} \left\{ 1 + u \left[- \frac{\Omega^{(u)}(a - y, r)}{\Lambda^{(u)}(a, r)} W^{(u)}(a) + \int_0^r \left(\Omega^{(u)}(a - y, t) - \frac{\Omega^{(u)}(a - y, r)}{\Lambda^{(u)}(a, r)} \Lambda^{(u)}(a, t) \right) dt \right] \right\},$$

which corresponds to (13) for $z > a$ showing that (13) holds for any $z \geq 0$. □

4.2 Proof of Proposition 1

Applying Esscher transform of measure (4) to the result (13), we have

$$\begin{aligned} \mathbb{E}_{y,x} \{ e^{-u\tau_r + \nu X_{\tau_r}} \mathbf{1}_{\{\tau_r < \infty\}} \} &= e^{\nu x} \mathbb{E}_{y,x} \{ e^{-p\tau_r} e^{\nu(X_{\tau_r} - x) - \psi(\nu)\tau_r} \mathbf{1}_{\{\tau_r < \infty\}} \} \\ &= e^{\nu x} \mathbb{E}_{y,x}^{\nu} \{ e^{-p\tau_r} \mathbf{1}_{\{\tau_r < \infty\}} \}, \end{aligned} \tag{39}$$

where we have defined $p = u - \psi(\nu)$. Under the new measure \mathbb{P}^{ν} ,

$$\begin{aligned} \mathbb{E}_{y,x}^{\nu} \{ e^{-p\tau_r} \mathbf{1}_{\{\tau_r < \infty\}} \} &= e^{-pr} \left\{ 1 + p \left[\overline{W}_v^{(p)}(a + x - y) - \frac{\Omega_v^{(p)}(a + x - y, r)}{\Lambda_v^{(p)}(a, r)} W_v^{(p)}(a) \right. \right. \\ &\quad \left. \left. + \int_0^r \left(\Omega_v^{(p)}(a + x - y, t) - \frac{\Omega_v^{(p)}(a + x - y, r)}{\Lambda_v^{(p)}(a, r)} \Lambda_v^{(p)}(a, t) \right) dt \right] \right\}, \end{aligned}$$

following which and the Eq. (39) our claim in (14) is established. □

5 Conclusions

We have presented some new results concerning Parisian ruin problem under Lévy insurance risk process, where ruin is announced when the risk process has gone below a certain level from the last record maximum of the process, also known as the drawdown, for a fixed consecutive period of time. They further extend the existing results on Parisian ruin below a fixed level of the risk process. Using recent developments on fluctuation and excursion theory of the drawdown of the Lévy risk process, the law of ruin-time and the position at ruin was given in terms of their joint Laplace transforms. Identities are presented semi-explicitly in terms of the scale function and the law of the Lévy process. The results can be used to calculate some quantities of interest in finance and insurance as discussed in the introduction.

Acknowledgements The author would like to thank a number of anonymous referees and associate editors for their useful suggestions and comments that improved the presentation of this paper. This paper was completed during the time the author visited the Hugo Steinhaus Center of Mathematics at Wrocław University of Science and Technology in Poland. The author acknowledges the support and hospitality provided by the Center. He thanks to Professor Zbigniew Palmowski for the invitation, and for some suggestions over the work discussed during the MATRIX Mathematics of Risk Workshop in Melbourne organized by Professors Konstantin Borovkov, Alexander Novikov and Kais Hamza to whom the author also like to thanks for the invitation. This research is financially supported by Victoria University PBRF Research Grants # 212885 and # 214168 for which the author is grateful.

References

1. Agarwal, V., Daniel, N., Naik, N.: Role of managerial incentives and discretion in hedge fund performance. *J. Financ.* **64**, 2221–2256 (2009)
2. Avram, F., Kyprianou, A.E., Pistorius, M.R.: Exit problems for spectrally negative Lévy processes and applications to (Canadized) Russian options. *Ann. Appl. Probab.* **14**, 215–238 (2004)
3. Bertoin, J.: *Lévy Processes*. Cambridge University Press, Cambridge (1996)
4. Broadie, M., Chernov, M., Sundaresan, S.: Optimal debt and equity values in the presence of Chapter 7 and Chapter 11. *J. Financ.* **LXII**, 1341–1377 (2007)
5. Chan, T., Kyprianou, A.E., Savov, M.: Smoothness of scale functions for spectrally negative Lévy processes. *Probab. Theory Rel.* **150**, 691–708 (2011)
6. Chesney, M., Jeanblanc-Picqué, M., Yor, M.: Brownian excursions and Parisian barrier options. *Adv. Appl. Probab.* **29**, 165–184 (1997)
7. Czarna, I., Palmowski, Z.: Ruin probability with Parisian delay for a spectrally negative Lévy process. *J. Appl. Probab.* **48**, 984–1002 (2011)
8. Dassios, A., Wu, S.: Perturbed Brownian motion and its application to Parisian option pricing. *Finance Stoch.* **14**, 473–494 (2010)
9. François, P., Morellec, E.: Capital structure and asset prices: some effects of bankruptcy procedures. *J. Bus.* **77**, 387–411 (2004)
10. Goetzmann, W.N., Ingersoll Jr., J.E., Ross, S.A.: High-water marks and hedge fund management contracts. *J. Financ.* **58**, 1685–1717 (2003)
11. Kusnetzov, A., Kyprianou, A.E., Rivero, V.: *The Theory of Scale Functions for Spectrally Negative Lévy Processes, Lévy Matters II*. Springer Lecture Notes in Mathematics. Springer, Berlin (2013)
12. Kyprianou, A.E.: *Introductory Lectures on Fluctuations of Lévy Processes with Applications*. Springer, Berlin (2006)
13. Kyprianou, A.E., Surya, B.A.: Principles of smooth and continuous fit in the determination of endogenous bankruptcy levels. *Finance Stoch.* **11**, 131–152 (2007)
14. Lambert, A.: Completely asymmetric Lévy processes confined in a finite interval. *Ann. Inst. Henri Poincaré* **2**, 251–274 (2000)
15. Landriault, D., Renaud, J-F., Zhou, X.: An insurance risk model with Parisian implementation delays. *Methodol. Comput. Appl. Probab.* **16**, 583–607 (2014)
16. Loeffen, R., Czarna, I., Palmowski, Z.: Parisian ruin probability for spectrally negative Lévy processes. *Bernoulli* **19**, 599–609 (2013)
17. Loeffen, R., Palmowski, Z., Surya, B.A.: Discounted penalty function at Parisian ruin for Lévy insurance risk process. *Insur. Math. Econ.* **83**, 190–197 (2017)
18. Mijatović, A., Pistorius, M.R.: On the drawdown of completely asymmetric Lévy process. *Stoc. Proc. Appl.* **122**, 3812–3836 (2012)

19. Palmowski, Z., Tumilewicz, J.: Pricing insurance drawdown-type contracts with underlying Lévy assets. *Insur. Math. Econ.* **79**, 1–14 (2018)
20. Surya, B.A.: Evaluating scale function of spectrally negative Lévy processes. *J. Appl. Probab.* **45**, 135–149 (2008)
21. Zhang, H., Leung, T., Hadjiliadis, O.: Stochastic modeling and fair valuation of drawdown insurance. *Insur. Math. Econ.* **53**, 840–850 (2013)

Simple Group Actions on Arc-Transitive Graphs with Prescribed Transitive Local Action



Marston Conder

Abstract This paper gives a partial answer to a question asked by Pierre-Emmanuel Caprace at the Groups St Andrews conference at Birmingham (UK) in August 2017, and investigated at the ‘Tutte Centenary Retreat’ workshop held at MATRIX in November 2017. Caprace asked if there exists a 2-transitive permutation group P such that only finitely many simple groups act arc-transitively on a connected graph X with local action P (of the stabiliser of a vertex v on the neighbourhood of v). Some evidence is given to suggest that the answer is “No”, even when ‘2-transitive’ is replaced by ‘transitive’, and then by way of illustration, a follow-up question is answered by showing that all but finitely many alternating groups have such an action on a 6-valent connected graph with vertex-stabiliser A_6 .

1 Introduction

At the Groups St Andrews conference held at Birmingham (UK) in August 2017, Pierre-Emmanuel Caprace asked if there exists a 2-transitive permutation group P such that only finitely many simple groups act arc-transitively on a connected graph X in such a way that the stabiliser (in the simple group) of a vertex v induces P on the neighbourhood of v . This question and a follow-up question about what happens when P is the alternating group A_6 were conveyed by Gabriel Verret and Michael Giudici at the ‘Tutte Centenary Retreat’ workshop held at MATRIX in November 2017. What follows is a partial answer to the main question, showing that even when ‘2-transitive’ is replaced by ‘transitive’, no such group P can exist if a certain conjecture about alternating quotients of amalgamated free products is valid, and then a full answer to the sub-question, showing that P cannot be A_6 , as well as noting that P cannot be one of a number of other permutation groups.

M. Conder (✉)

Mathematics Department, University of Auckland, Auckland, New Zealand
e-mail: m.conder@auckland.ac.nz

© Springer Nature Switzerland AG 2019

D. R. Wood et al. (eds.), 2017 *MATRIX Annals*, MATRIX Book Series 2,
https://doi.org/10.1007/978-3-030-04161-8_22

327

2 The General Question

One approach that can be taken to the general question is to consider the action of a group G on a graph X with the property that the stabiliser V in G of a vertex v of X is isomorphic to P (and induces P on the neighbourhood $X(v)$). First, let d be the degree of P (as a transitive permutation group). Then we may suppose that $d \geq 3$, since the automorphism groups of 2-valent connected arc-transitive graphs are dihedral and therefore soluble, and the question by Caprace is not relevant.

Now observe that if E is the stabiliser of an edge $e = \{v, w\}$ incident with v , and A is the stabiliser of the arc (v, w) (and hence isomorphic to a point-stabiliser in the given group P), then G is a homomorphic image of the free product $V *_A E$ with the subgroup $A = V \cap E$ amalgamated. Moreover, A has index d in V , and 2 in E .

Next, a conjecture made by Džambić and Jones [10] and supported by the author asserts that if V and E are any two finite groups with a common subgroup A with index $|V : A| \geq 3$ and index $|E : A| \geq 2$, then all but finitely many alternating groups A_n occur as homomorphic images of the amalgamated free product $V *_A E$. An even stronger version of this conjecture (believed to be true by the author) is as follows:

Conjecture 1 Let V and E be any two finite groups with a common subgroup A with index $|V : A| \geq 3$ and index $|E : A| \geq 2$, and let K be the core of A in $V *_A E$. Then all but finitely many A_n occur as the image of the amalgamated free product $V *_A E$ under some homomorphism that takes V and E to subgroups (of A_n) isomorphic to V/K and E/K respectively. In particular, if the amalgamated subgroup A is core-free in $V *_A E$, then all but finitely many A_n occur as images of $V *_A E$ under homomorphisms that are faithful on each of V and E .

It is easy to see this is stronger than the conjecture in [10], since for example any quotient of $C_2 *_C C_3 = C_2 * C_3$ (the modular group) is also a quotient of $C_4 *_C C_6$, but not vice versa. Also there is plenty of evidence in support of it. Indeed it is known to be true in many special cases (proved well before the original conjecture was made in [10]), such as those arising in the way described above from the study of finite arc-transitive and/or path-transitive 3-valent graphs [4, 8], or 7-arc-transitive 4-valent graphs [9], or similarly from the study of arc-transitive digraphs [6], chiral maps [2] or chiral polytopes [5], and even hyperbolic 3-manifolds [7].

Furthermore, if the above conjecture is valid, then the answer to Caprace's main question can be shown to be 'No', even when V is not 2-transitive:

Theorem 1 *If Conjecture 1 is valid, then for every transitive finite permutation group P , all but finitely many alternating groups A_n act arc-transitively on a connected graph X in such a way that the stabiliser in A_n of a vertex v induces P on the neighbourhood of v .*

Proof Let V , E and A be as above, with E chosen as a group containing an index 2 subgroup isomorphic to A , and consider the amalgamated free product $V *_A E$. Note that because A is a point-stabiliser in the permutation group $V (= P)$, it is core-free

in V and hence also A is core-free in $V *_A E$. Suppose further that $\theta: V *_A E \rightarrow G$ is any epimorphism to a finite non-abelian simple group G such that θ is faithful on each of V and E , and also let a be any element of the image of $E \setminus A$ in G , and let H be the θ -image of V (so that H is isomorphic to P).

Now let X be the double-coset graph $X(G, H, a)$, with vertices defined as the right cosets of H in G , with cosets Hx and Hy adjacent in X if and only if $xy^{-1} \in HaH$. This is a well-known construction attributed to Sabidussi, and described in detail in [4, 9] for example. The construction ensures that G acts as an arc-transitive group of automorphisms of the graph X (by right multiplication of cosets Hx in G), with vertex-stabiliser H acting transitively on the neighbourhood $X(H) = \{Hah : h \in H\}$ of the trivial coset H . Moreover, this action of H is equivalent to the action of V on cosets of A (by right multiplication), and hence the same as the natural action of the given permutation group P , as required.

Finally, if Conjecture 1 is valid then we can take G as the alternating group A_n for all but finitely many n , and this completes the proof. \square

3 Some Specific Cases

The same argument as used in the above proof can be applied to many specific cases where Conjecture 1 is known to be valid.

For example, this is often known to happen when the amalgamated subgroup A in $V *_A E$ is trivial. The validity of Conjecture 1 for the free products $C_3 * C_2$ and $C_k * C_2$ for all $k \geq 7$ follows from the fact that all but finitely many alternating groups are quotients of the ordinary $(2, 3, k)$ triangle group for any given such k (see [3]), and the same holds for $C_k * C_2$ for all $k \in \{4, 5, 6\}$ by the analogous properties of the $(2, k, m)$ triangle groups for $4 \leq k < m$ (see [11]).

Similarly, the fact that all but finitely many alternating groups are quotients of the extended $(2, 3, k)$ triangle (see [3]) shows that the same thing holds for $D_k *_{C_2} V_4$ for $k = 3$ and all $k \geq 7$. Also in [10] it was shown that infinitely many alternating groups occur as quotients of $A_5 *_{C_5} D_5$.

Hence, in particular, the answer to Caprace's question is 'No' when P is a cyclic or dihedral group of degree 3 or more, or the group of degree 12 induced by A_5 on cosets of a subgroup of order 5. It is fairly clear that the same answer holds for many other permutation groups besides these, and we complete this paper (and answer the sub-question mentioned earlier) by considering the case where $P = A_6$.

From now on we take V as A_6 and A as its point-stabiliser A_5 , and we choose E as $A_5 \times C_2$. Just as before, note that A is core-free in V and hence also core-free in $V *_A E$. We will show that all but finitely many alternating groups A_n occur as images of $V *_A E$ under homomorphisms that are faithful on each of V and E .

To do this, first we note that $V *_A E = A_6 *_A (A_5 \times C_2)$ is generated by three elements x , y and a with the following properties:

- x and y generate $V = A_6$ and satisfy the relations $x^2 = y^5 = (xy)^5 = (xy^2)^4 = 1$,
- y and $u = xy^{-1}xyx$ generate $A = A_5$ (and satisfy $u^2 = y^5 = (uy^2)^3 = 1$), and
- y , u and a generate $E = A_5 \times C_2$, and satisfy $a^2 = [u, a] = 1$ and $y^a = y^{-1}$.

These properties may be seen by taking $x = (3, 6)(4, 5)$ and $y = (1, 2, 3, 4, 5)$ in A_6 , with $u = (1, 4)(3, 5)$, and by viewing a as the inner automorphism of A_5 induced by conjugation by $(1, 3)(4, 5)$.

Next, we consider six particular transitive permutation representations of $V *_A E$, of degrees 1, 12, 42, 62, 21 and 31, as given below. In each case we give also the permutations induced by $u = xy^{-1}xyx$ and $w = xa$, and identify the fixed points of the subgroup E (generated by y , u and a), and call these fixed points ‘link points’, for reasons that should soon become clear.

Representation R_1 (degree 1)

$x \mapsto ()$,
 $y \mapsto ()$,
 $u \mapsto ()$,
 $a \mapsto ()$,
 $w \mapsto ()$.

Link point 1

Representation R_2 (degree 12)

$x \mapsto (1, 2)(3, 4)(9, 12)(10, 11)$,
 $y \mapsto (2, 3, 4, 5, 6)(7, 8, 9, 10, 11)$,
 $u \mapsto (2, 4)(3, 5)(7, 10)(9, 11)$,
 $a \mapsto (2, 7)(3, 11)(4, 10)(5, 9)(6, 8)$,
 $w \mapsto (1, 7, 2)(3, 10)(4, 11)(5, 9, 12)(6, 8)$.

Link points 1 and 12

Representation R_3 (degree 42)

$x \mapsto (1, 2)(3, 4)(7, 12)(8, 20)(9, 32)(10, 27)(11, 24)(13, 29)(14, 33)(15, 18)$
 $(16, 31)(17, 25)(21, 34)(23, 36)(26, 28)(30, 35)(39, 42)(40, 41)$,
 $y \mapsto (2, 3, 4, 5, 6)(7, 8, 9, 10, 11)(12, 13, 14, 15, 16)(17, 18, 19, 20, 21)$
 $(22, 23, 24, 25, 26)(27, 28, 29, 30, 31)(32, 33, 34, 35, 36)$
 $(37, 38, 39, 40, 41)$,
 $u \mapsto (2, 4)(3, 5)(7, 10)(9, 11)(12, 17)(13, 22)(14, 27)(15, 28)(16, 23)(18, 21)$
 $(19, 31)(20, 29)(24, 26)(25, 30)(32, 34)(33, 35)(37, 40)(39, 41)$,
 $a \mapsto (2, 7)(3, 11)(4, 10)(5, 9)(6, 8)(13, 16)(14, 15)(18, 21)(19, 20)(22, 23)$
 $(24, 26)(27, 28)(29, 31)(32, 37)(33, 41)(34, 40)(35, 39)(36, 38)$,
 $w \mapsto (1, 7, 12, 2)(3, 10, 28, 24)(4, 11, 26, 27)(5, 9, 37, 32)(6, 8, 19, 20)$
 $(13, 31)(14, 41, 34, 18)(15, 21, 40, 33)(16, 29)(17, 25)(22, 23, 38, 36)$
 $(30, 39, 42, 35)$.

Link points 1 and 42

Representation R_4 (degree 62)

$x \mapsto (1, 2)(3, 4)(7, 12)(8, 20)(10, 14)(11, 21)(13, 16)(15, 18)(23, 32)(24, 31)$
 $(25, 35)(26, 40)(27, 41)(28, 33)(29, 39)(36, 38)(42, 52)(43, 50)(44, 46)$
 $(45, 54)(47, 55)(48, 53)(57, 61)(60, 62),$
 $y \mapsto (2, 3, 4, 5, 6)(7, 8, 9, 10, 11)(12, 13, 14, 15, 16)(17, 18, 19, 20, 21)$
 $(22, 23, 24, 25, 26)(27, 28, 29, 30, 31)(32, 33, 34, 35, 36)$
 $(37, 38, 39, 40, 41)(42, 43, 44, 45, 46)(47, 48, 49, 50, 51)$
 $(52, 53, 54, 55, 56)(57, 58, 59, 60, 61),$
 $u \mapsto (2, 4)(3, 5)(7, 10)(9, 11)(12, 17)(13, 19)(16, 20)(18, 21)(22, 27)(23, 29)$
 $(26, 30)(28, 31)(33, 37)(34, 39)(35, 41)(38, 40)(44, 47)(45, 49)(46, 51)$
 $(48, 50)(52, 54)(53, 56)(57, 60)(58, 61),$
 $a \mapsto (2, 7)(3, 11)(4, 10)(5, 9)(6, 8)(12, 22)(13, 26)(14, 25)(15, 24)(16, 23)$
 $(17, 27)(18, 31)(19, 30)(20, 29)(21, 28)(32, 42)(33, 46)(34, 45)(35, 44)$
 $(36, 43)(37, 51)(38, 50)(39, 49)(40, 48)(41, 47)(52, 57)(53, 61)(54, 60)$
 $(55, 59)(56, 58),$
 $w \mapsto (1, 7, 22, 12, 2)(3, 10, 25, 44, 33, 21)(4, 11, 28, 46, 35, 14)(5, 9)$
 $(6, 8, 29, 49, 39, 20)(13, 23, 42, 57, 53, 40)(15, 31)$
 $(16, 26, 48, 61, 52, 32)(17, 27, 47, 59, 55, 41)(18, 24)(19, 30)$
 $(34, 45, 60, 62, 54)(36, 50)(37, 51)(38, 43)(56, 58).$

Link points 1 and 62

Representation R_5 (degree 21)

$x \mapsto (1, 2)(3, 4)(7, 12)(8, 20)(10, 14)(11, 21)(13, 16)(15, 18),$
 $y \mapsto (2, 3, 4, 5, 6)(7, 8, 9, 10, 11)(12, 13, 14, 15, 16)(17, 18, 19, 20, 21),$
 $u \mapsto (2, 4)(3, 5)(7, 10)(9, 11)(12, 17)(13, 19)(16, 20)(18, 21),$
 $a \mapsto (2, 7)(3, 11)(4, 10)(5, 9)(6, 8)(13, 16)(14, 15)(18, 21)(19, 20),$
 $w \mapsto (1, 7, 12, 2)(3, 10, 15, 21)(4, 11, 18, 14)(5, 9)(6, 8, 19, 20).$

Link point 1

Representation R_6 (degree 31)

$x \mapsto (1, 2)(3, 4)(7, 12)(8, 20)(10, 14)(11, 21)(13, 16)(15, 18)(23, 25)(24, 31)$
 $(26, 28)(27, 29),$
 $y \mapsto (2, 3, 4, 5, 6)(7, 8, 9, 10, 11)(12, 13, 14, 15, 16)(17, 18, 19, 20, 21)$
 $(22, 23, 24, 25, 26)(27, 28, 29, 30, 31),$
 $u \mapsto (2, 4)(3, 5)(7, 10)(9, 11)(12, 17)(13, 19)(16, 20)(18, 21)(22, 27)(23, 29)$
 $(26, 30)(28, 31),$
 $a \mapsto (2, 7)(3, 11)(4, 10)(5, 9)(6, 8)(12, 22)(13, 26)(14, 25)(15, 24)(16, 23)$
 $(17, 27)(18, 31)(19, 30)(20, 29)(21, 28),$

$$w \mapsto (1, 7, 22, 12, 2)(3, 10, 25, 16, 26, 21)(4, 11, 28, 13, 23, 14)(5, 9) \\ (6, 8, 29, 17, 27, 20)(15, 31)(18, 24)(19, 30).$$

Link point 1

Note that in each case, the permutations induced by x , y and u are necessarily even, since they generate a subgroup isomorphic to A_6 or the trivial group. On the other hand, the permutations induced by the involution a have 0, 5, 18, 30, 9 and 15 transpositions respectively, and hence the permutations induced by a and $w = xa$ are even in representations R_1 , R_3 and R_4 , but are odd in representations R_2 , R_5 and R_6 . Indeed, the cycle structure of the permutation induced by $w = xa$ in representations R_2 to R_6 is $2^3 3^2$, $2^3 4^9$, $2^8 5^2 6^6$, $1^3 2^1 4^4$ and $2^4 5^1 6^3$, respectively.

We will use these six representations as ‘building blocks’ for constructing transitive permutation representations of $V *_A E$ of arbitrarily large degree, by using the link points to join representations together.

To help to explain that, we observe how the image of each representation of $V *_A E$ splits into orbits of the subgroups $V = \langle x, y \rangle \cong A_6$, $E = \langle y, u, a \rangle \cong A_5 \times C_2$ and $A = V \cap E = \langle y, u \rangle \cong A_5$. For example, the image of R_3 (of degree 42) splits into three orbits of V , of lengths 6, 6 and 30, namely $\{1, 2, \dots, 6\}$, $\{7, 8, \dots, 36\}$ and $\{37, 38, \dots, 42\}$, and these in turn split into seven orbits of A , of lengths 1, 5, 5, 20, 5, 5 and 1, namely $\{1\}$, $\{2, 3, \dots, 6\}$, $\{7, 8, \dots, 11\}$, $\{12, 13, \dots, 31\}$, $\{32, 33, \dots, 36\}$, $\{37, 38, \dots, 41\}$ and $\{42\}$. Every orbit of the subgroup $E = \langle A, a \rangle$ is then either an orbit of A preserved by a , or a union of two orbits of A that are interchanged by a . For example (again), in R_3 the subgroup E has five orbits, of lengths 1, 10, 20, 10 and 1, namely $\{1\}$, $\{2, 3, \dots, 11\}$, $\{12, 13, \dots, 31\}$, $\{32, 33, \dots, 41\}$ and $\{42\}$.

This orbit decomposition is depicted for all six of our ‘building block’ representations in Fig. 1, with each small box indicating an orbit of A (and the number inside it indicating the length of that orbit), and each thin horizontal line indicating a connection between a pair of orbits of A that are interchanged by a . In particular, each small box with a ‘1’ inside it contains a link point, fixed by $E = \langle y, u, a \rangle$.

Next, if we take any two transitive permutation representations of $V *_A E$, say of degrees n_1 and n_2 , such that each representation contains at least one link point, then we can join them together to form a larger one of degree $n_1 + n_2$, by simply concatenating the permutations induced by each of x , y and a , and then adding a transposition to a that swaps the two chosen link points.

For example, we can join the first two representations together by re-labelling the single point of R_1 as ‘13’, and then adding a new transposition $(12, 13)$ to the permutation induced by a . This gives a transitive representation on 13 points, in which x , y and u induce the same permutations as given in R_2 , while a induces the involution $(2, 7)(3, 11)(4, 10)(5, 9)(6, 8)(12, 13)$ and $w = xa$ induces the permutation $(1, 7, 2)(3, 10)(4, 11)(5, 9, 13, 12)(6, 8)$.

Here, and in general when a pair of transitive representations are joined together in this way, the images of x , y and a still satisfy the same relations as in $V *_A E$, and hence (by the universal property of amalgamated products), the definition of the

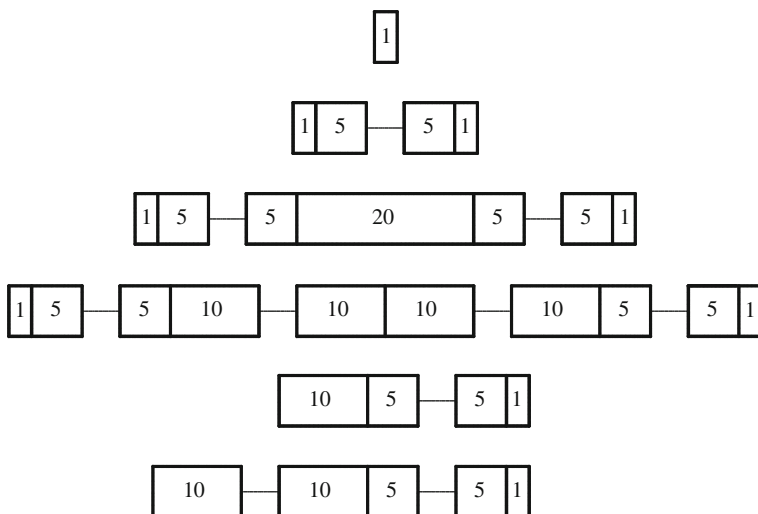


Fig. 1 Our ‘building block’ representations 1 to 6 (on 1, 12, 42, 62, 21 and 31 points respectively)

images extends to a new permutation representation of $V *_{A_5} E$. The only significant change is made to the permutation induced by a , and this simply joins two single-point orbits of $E = \langle y, u, a \rangle$ into a single two-point orbit of E . Similarly, the cycles of $w = xa$ containing the two link points are merged into a single cycle.

For another example, suppose we join together a copy of each of R_5 and R_2 by adding a new transposition that swaps the link point 1 of R_5 with link point 1 of R_2 (suitably re-labelled). Then we obtain a transitive permutation representation on $21 + 12 = 33$ points. Before the join, the permutations induced by w have cycle structures $1^3 2^1 4^4$ and $2^3 3^2$, with link point 1 of R_5 lying in a cycle of length 4 and link point 12 of R_2 lying in a cycle of length 3. The effect of the join is to merge those two cycles into a single cycle of length 7, leaving other cycles unchanged.

We have now dealt with enough properties of the building blocks and their conjunction to prove the following:

Theorem 2 *For all but finitely many positive integers n , both the alternating group A_n and the symmetric group S_n are homomorphic images of the amalgamated free product $A_6 *_{A_5} (A_5 \times C_2)$, and hence act faithfully as an arc-transitive group of automorphisms of some 6-valent graph with vertex-stabiliser isomorphic to A_6 .*

Proof For any positive integers k and m , let $n = 21 + 12k + 62m$, and observe that every odd positive integer $n \geq 395$ is expressible in this way.

Now construct a transitive permutation representation of $A_6 *_{A_5} (A_5 \times C_2)$ of odd degree n by stringing together a single copy of R_5 with k copies of R_2 , and then m copies of R_4 . Then the permutation induced by a has $9 + 6k + 31m$ transpositions (with $k + m$ of these coming from the linkages), and so the permutations induced by

a and w are even when m is odd, but odd when m is even. Indeed, the permutation induced by w has cycle structure $1^3 2^{1+3k+8m} 4^3 5^1 6^{k-1+6m} 7^1 8^1 10^{m-1}$.

The single 7-cycle comes from the linkage between the copy of R_5 and the first copy of R_2 . Also the length of every other cycle of w divides 120, so w^{120} is a single 7-cycle. Moreover, this 7-cycle contains a pair of points interchanged by x , a fixed point of y , and a pair of points interchanged by a . It follows that the image of this new representation is primitive (for otherwise there would be a block B of imprimitivity containing all 7 points of the 7-cycle, but then B would be preserved by each of x , y and a and hence by the whole group). And now by a theorem of Jordan [12, Theorem 13.9], this 7-cycle ensures that the permutations generate A_n for large $n \equiv 3 \pmod 4$ when m is odd, and S_n for large $n \equiv 1 \pmod 4$ when m is even.

Next, we can add a copy of R_1 to the final copy of R_4 , and get a transitive representation of $A_6 *_{A_5} (A_5 \times C_2)$ of even degree $n = 21 + 12k + 62m + 1$, and the same argument works, except that the parity of the permutations a and w changes, with a 5-cycle of $w = xa$ becoming another 6-cycle. In this case the permutations generate S_n with $n \equiv 0 \pmod 4$ when m is odd, and A_n with $n \equiv 2 \pmod 4$ when m is even.

Finally, we can replace the single copy of R_5 by a copy of R_6 , and insert a single copy of R_3 between the k copies of R_2 and the m copies of R_4 , and get transitive permutation representations of odd degree $n = 31 + 12k + 42 + 62m$ and even degree $n = 31 + 12k + 42 + 62m + 1$, in which the permutations induced by a and w are even if and only if m is even in the first case, and are even if and only if m is odd in the second case. The cycle structure of xa is altered by addition of a 9-cycle, plus some changes in the 1-, 2-, 4-, 6- and 8-cycles, and replacement of the single 7-cycle by a new single 7-cycle coming from the linkage between R_2 and R_3 , but the same arguments apply as earlier. In this case the induced permutations generate A_n for large $n \equiv 1 \pmod 4$ and S_n for large $n \equiv 3 \pmod 4$ when m is even, and S_n for large $n \equiv 2 \pmod 4$ and A_n for large $n \equiv 0 \pmod 4$ when m is odd.

These constructions cover all residue classes mod 4 for the degree n , for both A_n and S_n for large enough n (indeed for all $n \geq 447$), as required. □

Incidentally, we also obtain the following, because if a is any involution in $E \setminus A$, then the index 2 subgroup $S = \langle V, V^a \rangle$ in the group $V *_A E = A_6 *_{A_5} (A_5 \times C_2)$ used above is isomorphic to $A_6 *_{A_5} A_6$, and also maps onto A_n for large n . This strengthens an observation made by Peter Neumann and Cheryl Praeger at the Groups St Andrews conference that $A_6 *_{A_5} A_6$ has infinitely many alternating quotients.

Corollary 1 *All but finitely many alternating groups occur as quotients of the amalgamated free product $A_6 *_{A_5} A_6$.*

Acknowledgements The author acknowledges the help of MAGMA [1] in finding and analysing representations of $A_6 *_{A_5} (A_5 \times C_2)$ used in Sect. 3, and is grateful for support from the N.Z. Marsden Fund (via project UOA1626).

References

1. Bosma, W., Cannon, J., Playoust, C.: The Magma algebra system. I. The user language. *J. Symbolic Comput.* **24**, 235–265 (1997)
2. Bujalance, E., Conder, M.D.E., Costa A.F.: Pseudo-real Riemann surfaces and chiral regular maps. *Trans. Amer. Math. Soc.* **362**, 3365–3376 (2010)
3. Conder, M.D.E.: More on generators for alternating and symmetric groups. *Quart. J. Math. (Oxford)* **32**, 137–163 (1981)
4. Conder, M.: An infinite family of 5-arc-transitive cubic graphs. *Ars Combin.* **25A**, 95–108 (1988)
5. Conder, M., Hubbard, I., O'Reilly-Regueiro, E., Pellicer, D.: Construction of chiral 4-polytopes with an alternating or symmetric group as automorphism group. *J. Algebraic Combin.* **42**, 225–244 (2015)
6. Conder, M., Lorimer, P., Praeger, C.: Constructions for arc-transitive digraphs. *J. Aust. Math. Soc.* **59**, 61–80 (1995)
7. Conder, M., Martin, G., Torstenson, A.: Maximal symmetry groups of hyperbolic 3-manifolds. *N. Z. J. Math.* **35**, 37–62 (2006)
8. Conder, M.D.E., Praeger, C.E.: Remarks on path-transitivity in finite graphs. *European J. Combin.* **17**, 371–378 (1996)
9. Conder, M.D.E., Walker, C.G.: The infinitude of 7-arc-transitive graphs. *J. Algebra* **208**, 619–629 (1998)
10. Džambić, A., Jones, G.A.: p -adic Hurwitz groups. *J. Algebra* **379**, 179–207 (2013)
11. Everitt, B.: Alternating quotients of Fuchsian groups. *J. Algebra* **223**, 457–476 (2000)
12. Wielandt, H.: *Finite Permutation Groups*. Academic Press, New York (1964)

Biquasiprimitive Oriented Graphs of Valency Four



Nemanja Poznanović and Cheryl E. Praeger

Abstract In this short note we describe a recently initiated research programme aiming to use a normal quotient reduction to analyse finite connected, oriented graphs of valency 4, admitting a vertex- and edge-transitive group of automorphisms which preserves the edge orientation. In the first article on this topic (Al-bar et al. *Electr J Combin* 23, 2016), a subfamily of these graphs was identified as ‘basic’ in the sense that all graphs in this family are normal covers of at least one ‘basic’ member. These basic members can be further divided into three types: quasiprimitive, biquasiprimitive and cycle type. The first and third of these types have been analysed in some detail. Recently, we have begun an analysis of the basic graphs of biquasiprimitive type. We describe our approach and mention some early results. This work is on-going. It began at the Tutte Memorial MATRIX Workshop.

A graph Γ is said to be G -oriented for some subgroup $G \leq \text{Aut}(\Gamma)$, if G acts transitively on the vertices and edges of Γ , and Γ admits a G -invariant orientation of its edges. Any graph Γ admitting a group of automorphisms which acts transitively on its vertices and edges but which does not act transitively on its arcs can be viewed as a G -oriented graph. These graphs are usually said to be G -half-arc-transitive, and form a well-studied class of vertex-transitive graphs.

A G -oriented graph Γ necessarily has even valency, with exactly half of the edges incident to a vertex α being oriented away from α to one of its neighbours. Since such a graph is vertex-transitive, it follows that all of its connected components are isomorphic, and each connected component is itself a G -oriented graph. We therefore restrict our attention to the study of connected G -oriented graphs. For each even integer m , we let $\mathcal{OG}(m)$ denote the family of all graph-group pairs (Γ, G) such that Γ is a finite connected G -oriented graph of valency m .

N. Poznanović (✉)
University of Melbourne, Melbourne, VIC, Australia

C. E. Praeger
University of Western Australia, Perth, Australia
e-mail: cheryl.praeger@uwa.edu.au

Graphs contained in $\mathcal{OG}(2)$ are simply oriented cycles. The family $\mathcal{OG}(4)$ on the other hand, has been studied for several decades now, see for instance [4–6]. These papers suggest a framework for describing the structure of graphs Γ for pairs (Γ, G) in the class $\mathcal{OG}(4)$ by considering various types of quotients based on the structure of certain kinds of cycles of Γ . In this new approach we study quotients based on normal subgroups of the group G .

Normal Quotients Given a pair $(\Gamma, G) \in \mathcal{OG}(4)$ and a normal subgroup N of G , we define a new graph Γ_N as follows: the vertex set of Γ_N consists of all N -orbits on the vertices of Γ , and there is an edge between two N -orbits $\{B, C\}$ in Γ_N if and only if there is an edge of the form $\{\alpha, \beta\}$ in Γ , with $\alpha \in B$ and $\beta \in C$. The graph Γ_N is called a *G-normal-quotient* of Γ . The group G induces a subgroup G_N of automorphisms of Γ_N , namely $G_N = G/K$ for some normal subgroup K of G such that $N \leq K$. The K -orbits are the same as the N -orbits so $\Gamma_K = \Gamma_N$, although sometimes K may be strictly larger than N .

In general, the pair (Γ_N, G_N) need not lie in $\mathcal{OG}(4)$. For instance, if the normal subgroup N is transitive on the vertex set of Γ , then Γ_N consists of just a single vertex. If the graph Γ is bipartite and the two N -orbits form the bipartition, then Γ_N will be isomorphic to the complete graph on two vertices K_2 . In other cases the quotient graph Γ_N may also be a cycle graph C_r , for $r \geq 3$. These three types of quotients are defined to be *degenerate* in the sense that in each of these cases Γ_N does not have valency 4 and so $(\Gamma_N, G_N) \notin \mathcal{OG}(4)$. It turns out that these cases are the only obstacles to the pair (Γ_N, G_N) lying in $\mathcal{OG}(4)$.

Theorem 1 ([1] Theorem 1.1) *Let $(\Gamma, G) \in \mathcal{OG}(4)$ with vertex set X , and let N be a normal subgroup of G . Then G induces a permutation group G_N on the set of N -orbits in X , and either*

- (i) (Γ_N, G_N) is also in $\mathcal{OG}(4)$, Γ is a G -normal cover of Γ_N , N is semiregular on vertices, and $G_N = G/N$; or
- (ii) (Γ_N, G_N) is a degenerate pair, (i.e. Γ_N is isomorphic to K_1 , K_2 or C_r , for some $r \geq 3$).

This leads to a framework for studying the family $\mathcal{OG}(4)$ using normal quotient reduction. The first goal of this approach is to develop a theory to describe the ‘basic’ pairs in $\mathcal{OG}(4)$. A graph-group pair $(\Gamma, G) \in \mathcal{OG}(4)$ is said to be *basic* if all of its G -normal quotients relative to non-trivial normal subgroups are degenerate. Since Theorem 1 ensures that every member of $\mathcal{OG}(4)$ is a normal cover of a basic pair, the second aim of this framework is to develop a theory to describe the G -normal covers of these basic pairs. This approach has been successfully used in the study of other families of graphs with prescribed symmetry properties, see for instance [7–9].

The basic pairs may be further divided into three types. A pair $(\Gamma, G) \in \mathcal{OG}(4)$ is said to be basic of *quasiprimitive type* if all G -normal quotients Γ_N of Γ are isomorphic to K_1 . This occurs precisely when all non-trivial normal subgroups of G are transitive on the vertices of Γ . (Such a permutation group is said to be quasiprimitive.)

If the only normal quotients of a basic pair $(\Gamma, G) \in \mathcal{OG}(4)$ are the graphs K_1 or K_2 , and Γ has at least one G -normal quotient isomorphic to K_2 , then (Γ, G) is said to be basic of *biquasiprimitive type*. (The group G here is biquasiprimitive: it is not quasiprimitive but each nontrivial normal subgroup has at most two orbits.) The other basic pairs in $\mathcal{OG}(4)$ must have at least one normal quotient isomorphic to a cycle graph C_r , and these basic pairs are said to be of *cycle type*.

The basic pairs of quasiprimitive type have been analysed in [1], and further analysis was conducted on basic pairs of cycle type in [2] and [3]. Although more remains to be done to describe the structure of basic pairs of cycle type, the main focus of our work is the biquasiprimitive case.

Basic Pairs of Biquasiprimitive Type: Early Results Our current work aims to develop a theory to describe the basic pairs of biquasiprimitive type. Following the work done in [1] describing quasiprimitive basic pairs, we aim to produce similar structural results and constructions for the biquasiprimitive case. In [10] there is a group theoretic tool available for studying finite biquasiprimitive groups analogous to the O’Nan-Scott Theorem for finite primitive and quasiprimitive permutation groups. We outline our general approach below, though this work is still in progress.

Let Γ be a graph with vertex set X and suppose that $(\Gamma, G) \in \mathcal{OG}(4)$ is basic of biquasiprimitive type for some group G . Then there exists a normal subgroup N of G with exactly two orbits on X , and all normal subgroups of G have at most two orbits. It is easy to see that Γ is bipartite: since Γ is connected there is an edge joining vertices in different N -orbits, and since G normalises N and is edge-transitive, each edge joins vertices in different N -orbits. Thus the two orbits of N form a bipartition of Γ .

Let $\{\Delta, \Delta'\}$ denote the bipartition of the vertices of Γ , and let G^+ be the index 2 subgroup of G fixing Δ (and Δ') setwise. Since Γ is G -vertex-transitive it follows that G^+ is transitive on both Δ and Δ' . As we just saw, any non-trivial intransitive normal subgroup N of G must have the sets Δ, Δ' as its two orbits on X , and hence $N \leq G^+$. It can also be shown that the action of G^+ on Δ is faithful.

Consider now a minimal normal subgroup M of G^+ . If M is also normal in G then the M -orbits on X are Δ and Δ' and M is a minimal normal subgroup of G .

On the other hand, if M is not normal in G , then for any element $x \in G \setminus G^+$, we see that M^x is also a minimal normal subgroup of $(G^+)^x = G^+$, and furthermore, $M \neq M^x$ since otherwise G would normalise M . It follows from the minimality of M that $M \cap M^x = 1$ and hence that $M \times M^x$ is contained in G^+ and is normal in G . Letting $N := M \times M^x$, we see that the N -orbits on X are Δ and Δ' since N is normal in G .

In summary, we always have a normal subgroup N of G contained in G^+ with Δ and Δ' the N -orbits in X , and such that either

- (a) N is a minimal normal subgroup of G^+ and $N = T^k$ for some simple group T and $k \geq 1$; or
- (b) $N = M \times M^x$ where $x \in G \setminus G^+$, and $M = T^\ell$ is a minimal normal subgroup of G^+ with T a simple group and $\ell \geq 1$. In particular, $N \cong T^k$ with $k = 2\ell$.

For a vertex $\alpha \in \Delta$, the vertex stabilisers G_α and G_α^+ are equal and $G^+ \cong NG_\alpha$. Moreover, since the vertex stabilisers of 4-valent G -oriented graphs are 2-groups, it follows that $G^+/N \cong G_\alpha/N_\alpha$ is also a 2-group.

Hence by analysing the minimal normal subgroups of G and G^+ as above, and considering the various possibilities for the direct factors T of N , we can reduce the possibilities for basic pairs of biquasiprimitive type to several cases. In fact, our main result so far uses combinatorial arguments to bound the values of ℓ and k in cases (a) and (b) above, though this is still a work in progress.

We give an infinite family of examples of basic biquasiprimitive pairs. These graphs have order $2p^2$, with p prime, and G^+ has an elementary abelian normal subgroup. There were no analogues of these examples in the basic quasiprimitive case since the minimal normal subgroups in that case are nonabelian, [1, Theorem 1.3].

Example 1 Let p be a prime such that $p \equiv 3 \pmod{4}$, let $\Delta = \{(x, y)_0 \mid x, y \in C_p\}$ and $\Delta' = \{(x, y)_1 \mid x, y \in C_p\}$, two copies of the additive group $N = C_p^2$, and let $X = \Delta \cup \Delta'$. Define $\delta \in \text{Sym}(X)$ by $(x, y)_\varepsilon^\delta = (y, -x)_{1-\varepsilon}$, for $x, y \in C_p$ and $\varepsilon \in \{0, 1\}$, and let $G := N \rtimes \langle \delta \rangle$. Note that δ has order 4 and normalises N . Also for $\alpha = (0, 0)_0$, $G_\alpha = N_\alpha = \langle \delta^2 \rangle \cong C_2$.

Define the G -oriented graph Γ to have vertex set X and, for each $x, y \in C_p$, edges oriented from $(x, y)_0$ to $(x, y \pm 1)_1$ and from $(x \pm 1, y)_1$ to $(x, y)_0$.

Then Γ has valency 4, G is vertex- and edge-transitive, and G preserves the edge orientation. Thus $(\Gamma, G) \in \mathcal{OG}(4)$. Also Γ is bipartite, and G is biquasiprimitive (verifying the latter property uses the fact that $p \equiv 3 \pmod{4}$). Hence (Γ, G) is basic of biquasiprimitive type.

Our goal is to refine our restrictions on k and ℓ to such an extent that we can give constructions of families of examples for all possible values of these parameters. Example 1 defines the graphs as BiCayley graphs, and other BiCayley examples arise naturally in our context. However, in the case where G^+ has no normal subgroup which is regular on the two G^+ orbits Δ and Δ' , different constructions will be required.

As noted above, these results develop the work initiated in [1] and developed further in [2, 3].

Acknowledgements Both authors are grateful for the opportunity to participate in the Tutte Memorial MATRIX retreat which gave them to chance to commence work on this problem. The authors also thank Georgina Liversidge for some useful discussions at the retreat. The first author acknowledges support of a Research Training Program Scholarship at the University of Melbourne. The second author is grateful for project funding from the Deanship of Scientific Research, King Abdulaziz University (grant no. HiCi/H1433/363-1) which provided the opportunity to focus on this research problem.

References

1. Al-bar, J.A., Al-kenani, A.N., Muthana, N.M., Praeger, C.E., Spiga, P.: Finite edge-transitive oriented graphs of valency four: a global approach. *Electron. J. Combin.* **23**(1), #P1.10 (2016). arXiv: 1507.02170
2. Al-Bar, J.A., Al-Kenani, A.N., Muthana, N.M., Praeger, C.E.: Finite edge-transitive oriented graphs of valency four with cyclic normal quotients. *J. Algebraic Combin.* **46**(1), 109–133 (2017)
3. Al-bar, J.A., Al-kenani, A.N., Muthana, N.M., Praeger, C.E.: A normal quotient analysis for some families of oriented four-valent graphs. *Ars Mat. Contemp.* **12**(2), 361–381 (2017)
4. Marušič, D.: Half-transitive group actions on finite graphs of valency 4. *J. Combin. Theory B* **73**(1), 41–76 (1998)
5. Marušič, D., Praeger, C.E.: Tetravalent graphs admitting half-transitive group actions: alternating cycles. *J. Combin. Theory B* **75**(2), 188–205 (1999)
6. Marušič, D., Šparl, P.: On quartic half-arc-transitive metacirculants. *J. Algebraic Combin.* **28**, 365–395 (2008)
7. Morris, J., Praeger, C.E., Spiga, P.: Strongly regular edge-transitive graphs. *Ars Mat. Contemp.* **2**(2), 137–155 (2009)
8. Praeger, C.E.: An O’Nan-Scott theorem for finite quasiprimitive permutation groups and an application to 2-arc transitive graphs. *J. Lond. Math. Soc.* **2**(2), 227–239 (1993)
9. Praeger, C.E.: Finite normal edge-transitive Cayley graphs. *Bull. Aust. Math. Soc.* **60**(2), 207–220 (1999)
10. Praeger, C.E.: Finite transitive permutation groups and bipartite vertex-transitive graphs. III. *J. Math.* **47**(1–2), 461–475 (2003)

The Contributions of W.T. Tutte to Matroid Theory



Graham Farr and James Oxley

Abstract Bill Tutte was born on May 14, 1917 in Newmarket, England. In 1935, he began studying at Trinity College, Cambridge reading natural sciences specializing in chemistry. After completing a master's degree in chemistry in 1940, he was recruited to work at Bletchley Park as one of an elite group of codebreakers that included Alan Turing. While there, Tutte performed “one of the greatest intellectual feats of the Second World War.” Returning to Cambridge in 1945, he completed a Ph.D. in mathematics in 1948. Thereafter, he worked in Canada, first in Toronto and then as a founding member of the Department of Combinatorics and Optimization at the University of Waterloo. His contributions to graph theory alone mark him as arguably the twentieth century's leading researcher in that subject. He also made groundbreaking contributions to matroid theory including proving the first excluded-minor theorems for matroids, one of which generalized Kuratowski's Theorem. He extended Menger's Theorem to matroids and laid the foundations for structural matroid theory. In addition, he introduced the Tutte polynomial for graphs and extended it and some relatives to matroids. This paper will highlight some of his many contributions focusing particularly on those to matroid theory.

1 Introduction

The task of summarizing Bill Tutte's mathematical contributions in a short paper is an impossible one. There are too many, they are too deep, and their implications are too far-reaching. This paper will discuss certain of these contributions giving particular emphasis to his work in matroid theory and the way in which that work links to graph theory. The terminology used here will follow Oxley [16].

G. Farr

Faculty of Information Technology, Monash University, Clayton, VIC, Australia
e-mail: Graham.Farr@monash.edu

J. Oxley (✉)

Mathematics Department, Louisiana State University, Baton Rouge, LA, USA
e-mail: oxley@math.lsu.edu

This paper will attempt to give insight into the thoughts and motivations that guided Tutte's mathematical endeavours. To do this, we shall quote extensively from three sources. Dan Younger, Tutte's long-time colleague and friend at the University of Waterloo, wrote the paper *William Thomas Tutte 14 May 1917–2 May 2002* [53] in the *Biographical Memoirs of Fellows of the Royal Society*, and that paper includes many quotes from Tutte that are reproduced here. In 1999, Tutte presented the Richard Rado Lecture *The Coming of the Matroids* at the British Combinatorial Conference in Canterbury. We will also quote from Tutte's write-up of that lecture in the conference proceedings [47]. Finally, we draw on commentaries by Tutte on his own papers that appear in *Selected Papers of W.T. Tutte I, II* [45, 46], published in 1979 to mark Tutte's 60th birthday.

These *Selected Papers* were edited by D. McCarthy and R. G. Stanton. Ralph Stanton was a noted mathematician who had been the first Dean of Graduate Studies at the University of Waterloo and who recruited Tutte to Waterloo from Toronto in 1962. Stanton's foreword to the *Selected Papers* provides a context for the magnitude of Tutte's achievements:

Not too many people are privileged to practically create a subject, but there have been several this century. Albert Einstein created Relativity . . . Similarly, modern Statistics owes its existence to Sir Ronald Fisher's exceptionally brilliant and creative work. And I think that Bill Tutte's place in Graph Theory is exactly like that of Einstein in Relativity and that of Fisher in Statistics. He has been both a great creative artist and a great developer.

Bill Tutte was born on May 14, 1917 in Newmarket, England. His family moved several times when he was young but they returned to the Newmarket area, to the village of Cheveley, when Bill was about seven. Bill attended the local school. In May, 1927 and again a year later, he won a scholarship to the Cambridge and County High School for Boys, some eighteen miles from his home. The first time he won, his parents judged that it was too far for their son to travel and he was kept home. A year later his parents permitted him to attend the school despite the long daily commute each way, by bike and by train [53, p. 287]. In the high school library, Bill came across Rouse Ball's book *Mathematical Recreations and Essays* [1], first published in 1892. That book included discussions of chess-board recreations, map colouring problems, and unicursal problems (including Euler tours and Hamiltonian cycles). Some parts of his chemistry classes were [45, p. 1] pure graph theory and in his physics classes, he learned about electrical circuits and Kirchhoff's Laws. Tutte wrote [45, p. 1],

When I became an undergraduate at Trinity College, Cambridge, I already possessed much elementary graph-theoretical knowledge though I do not think I had this knowledge well-organized at the time.

In 1935, Tutte began studying at Trinity College, Cambridge. He read natural sciences, specializing in chemistry. From the beginning, he attended lectures of the Trinity Mathematical Society. Three other members of that Society, all of whom were first-year mathematics students, were R. Leonard Brooks, Cedric A.B. Smith, and Arthur H. Stone. This group had various names [47, p. 4] including 'The Important Members, The Four Horsemen, The Gang of Four.' They became fast

friends spending many hours discussing mathematical problems. Tutte wrote [53, p. 288],

As time went on, I yielded more and more to the seductions of Mathematics.

Tutte's first paper [22], in chemistry, was published in 1939 in the prestigious scientific journal *Nature*. His first mathematical paper, *The dissection of rectangles into squares*, was published with Brooks, Smith, and Stone [3] in 1940 in the *Duke Mathematical Journal*. Their motivating problem was to divide a square into a finite number of unequal squares. In 1939, Sprague [21] from Berlin published a solution to this problem just as The Four were in the final stages of preparing their paper in which, ingeniously, they converted the original problem into one for electrical networks. Writing later about The Four's paper, Tutte said [45, p. 3],

I value the paper not so much for its ostensible geometrical results, which Sprague largely anticipated, as for its graph-theoretical methods and observations.

Tutte went on to note [45, p. 4] that, in this paper,

two streams of graph theory from my early studies came together, Kirchhoff's Laws from my Physics lessons, and planar graphs from Rouse Ball's account of the Four Colour Problem.'

Tutte wrote a very readable account of this work in Martin Gardner's *Mathematical Games* column in *Scientific American* in November, 1958, and that account is now available online [32].

The Four's paper is remarkable not only for its solution to the squaring-the-square problem and its beautiful graph-theoretic ideas, but also for the extent to which it contains the seeds of Tutte's later work. In it, we find planarity, duality, flows, numbers of spanning trees, a deletion-contraction relation, symmetry, and above all the powerful application of linear algebra to graph theory.¹

After completing his chemistry degree in 1938, Tutte worked as a postgraduate student in physical chemistry at Cambridge's famous Cavendish Laboratory completing a master's degree in 1940. Tutte's work in chemistry [53, p. 288]

convinced him that he would not succeed as an experimenter. He asked his tutor, Patrick Duff, to arrange his transfer from natural sciences to mathematics. This transfer took place at the end of 1940.

Tutte later wrote [47, p. 4],

I left Cambridge in 1941 with the idea that graph theory could be reduced to abstract algebra but that it might not be the conventional kind of algebra.

¹Incidentally, it may also be regarded as Tutte's first paper on graph drawing. In that field, too, he is regarded as a pioneer, mostly because of his 1963 paper 'How to draw a graph' [35]. But it is still worth noting the graph-drawing aspect of his very first mathematics paper: the squared rectangles are a type of simultaneous "visibility drawing" of a planar graph and its dual.

Like so many of the brightest minds in Britain at the time, Tutte was recruited as a codebreaker and worked at the Bletchley Park Research Station—now famous, but then top secret—from 1941 till 1945. He wrote [47, p. 5],

at Bletchley I was learning an odd new kind of linear algebra.

Narrating the 2011 BBC documentary, *Code-Breakers: Bletchley Park's Lost Heroes* [2], the actress Keeley Hawes says,

This is Bletchley Park. In 1939, it became the wartime headquarters of MI6. If you know anything, about what happened here, it will be that a man named Alan Turing broke the German Naval code known as 'Enigma' and saved the nation; and he did. But that's only half the story.

Then Captain Jerry Roberts, who had been a Senior Cryptographer at the Park during the war, speaks:

There were three heroes of Bletchley Park. The first was Alan Turing; the second was Bill Tutte, who broke the Tunny system, a quite amazing feat; and the third was Tommy Flowers who, with no guidelines, built the first computer ever.

Tutte's work at Bletchley Park was truly profound. The problem he faced was to break into communications encoded by an *unknown* cypher machine, codenamed Tunny by the British. (Its real name was Lorenz SZ40.) This machine was much more secure and complex than the famous Enigma machine, reflecting its use at the highest levels of the Nazi regime including by Hitler himself. Furthermore, although the British knew the architecture of the Enigma machine, the design of the Tunny machine was a complete mystery to them. The problem Tutte faced was thus far harder than the Enigma problem which Turing is justly celebrated for solving. Tutte's first problem was the *diagnosis* of the Tunny machine (that is, determining how it worked), just from collected cyphertext; only then could he and his colleagues move on to *cryptanalysis*. Tutte made the crucial breakthrough in diagnosis, an astonishing achievement. He then went on to develop cryptanalysis algorithms. These were very computationally intensive. The Colossus cryptanalytic computers, designed by Tommy Flowers, were built to implement Tutte's algorithms and performed service of incalculable value for the remainder of the war.

The University of Waterloo's magazine for Spring, 2015 has an article *Keeping Secrets* about this work in which one reads,

According to Bletchley Park's historians, General Dwight D. Eisenhower himself described Tutte's work as one of the greatest intellectual feats of the Second World War.

Some details of this work can be found in [7, 12, 53]. Tutte's own account of these efforts appear in [47, 48].

As a consequence of Tutte's top-secret code-breaking work at Bletchley Park, he was elected a Fellow of Trinity College in 1942. He wrote [53, p. 291] of this,

It seemed to me that the election might be criticized as a breach of security, but no harm came of it.

In 1945, after the war, Tutte returned to Cambridge for a Ph.D. in mathematics, supervised by Shaun Wylie, with whom Tutte had worked at Bletchley. Tutte completed his Ph.D. thesis, *An algebraic theory of graphs*, in 1948 despite Wylie's advice [53, p. 291] to

drop graph theory and take up something respectable, such as differential equations.

Tutte's thesis, which was xi + 417 pages, was an extraordinary accomplishment with the ideas in it forming the basis for much of his work for the next two decades. We discuss it in more detail in Sect. 8. He wrote [53, p. 291] of his decision to stick with graph theory,

If one assumes that graph theory was my *métier*, it was just as well that I had the prestige of a Fellow of Trinity.

2 Tutte's Doctoral Research

Tutte's first year of doctoral research was remarkably productive. He submitted six papers during the period from November 1945 to December 1946, including four that became classics of the field, though most of them bore no relation to his Ph.D. thesis.

In the first of these classic papers [24], Tutte found a 46-vertex counterexample to an 1884 conjecture of Tait [23] that every cubic planar graph is Hamiltonian.

Tutte's paper *A ring in graph theory* is his first paper on the Tutte polynomial and one of his most profound. His polynomial is not given explicitly in any of its usual forms, and its presence is somewhat obscured by some technical details and the use of multivariate polynomials to develop much of the theory. But the main ingredients of Tutte-polynomial theory are all there. We return to it shortly, in Sect. 2.1.

His third classic paper [27] studied symmetry in cubic graphs. An *s-arc* in a graph is a walk with s edges in which consecutive edges are always distinct. Apart from this constraint, vertices and edges may occur repeatedly. Note that a walk and its reverse are considered to be different. A graph G is *s-arc-transitive* if it has at least one *s-arc* and, for any two *s-arcs*, there is an automorphism of G that maps one to the other. This is a very strong symmetry property indeed. Tutte showed that there are no *s-arc-transitive* cubic graphs with $s > 5$, gave an inequality relating girth to s , and characterized graphs where the inequality comes as close to equality as possible for a given girth; these are the *g-cages*, a finite family of graphs, the most complex being his 8-cage. This paper became enormously influential in the theory of symmetric graphs.

The fourth of these groundbreaking papers [25] proved the characterization of when a graph has a 1-factor, or perfect matching. This theorem is now a staple of most introductory courses on graph theory.

2.1 ‘A Ring in Graph Theory’

The starting point and driving principle of this paper is the observation that certain functions on graphs obey *deletion-contraction relations*. As an example of such a function, Tutte considered the *complexity* $C(G)$ of a connected graph G , this being the number of spanning trees of G . When The Four were working on the problem of partitioning a rectangle into unequal squares, they observed that complexity obeys the following recursion.

Lemma 1 *In a graph G , let e be an edge that is neither a loop or a cut edge. Then*

$$C(G) = C(G \setminus e) + C(G/e).$$

Proof Partition the set of spanning trees of G into

- (i) those not using e ; and
- (ii) those using e .

There are $C(G \setminus e)$ spanning trees in (i); and the spanning trees in (ii) match up with the spanning trees of G/e .

Tutte wrote [45, p. 51],

I wondered if complexity, or tree number, could be characterized by the above identity alone and decided that it could not.

His paper considered the following.

Problem 1 What isomorphism-invariant functions W of graphs satisfy

$$W(G) = W(G \setminus e) + W(G/e)$$

for all non-loop edges e of G ?

He called such a function, taking values in an abelian group, a *W-function*. A *W-function* is a *V-function* if

$$W(G_1 \cup G_2) = W(G_1)W(G_2)$$

for all disjoint graphs G_1 and G_2 , where W now takes values in a commutative ring with unity.

For a graph G , let $P(G; \lambda)$ denote the number of proper λ -colourings of G . Tutte noted that $(-1)^{|V(G)|}$ times $P(G; \lambda)$ is an example of a *V-function*. This is an immediate consequence of the following lemma, which was first proved by Foster, in “Note added in proof” in [51, p. 718].

Lemma 2 *For a non-loop edge e of a graph G ,*

$$P(G; \lambda) = P(G \setminus e; \lambda) - P(G/e; \lambda).$$

Proof Let e have distinct endpoints u and v . Partition the proper k -colourings of $G \setminus e$ into

- (i) those in which u and v have different colours; and
- (ii) those in which u and v have the same colour.

In (i), we are counting the number of proper k -colourings of G ; while (ii) corresponds to the number of proper k -colourings of G/e .

Tutte's insightful breakthrough here was to focus on the two recursions:

1. $W(G) = W(G \setminus e) + W(G/e)$; and
2. $W(G_1 \cup G_2) = W(G_1)W(G_2)$.

Many readers will recognize here the origins of the Tutte polynomial. There are technical differences between the multivariate polynomials in this paper and the more familiar polynomials of Whitney and Tutte, which we will define in Sect. 3. But some simple adjustments—such as substitutions to give bivariate specializations, and dividing by $x^{k(G)}$ or $(x-1)^{k(G)}$ —reveal both the Whitney rank generating function and Tutte polynomial, albeit in period costume. The relationship between these two polynomials, which is just a coordinate translation of one step in each direction, is subsumed by a more general result in the paper. In fact, most of the main ingredients of Tutte-polynomial theory are here, with deletion-contraction relations at the core. The details of this paper are discussed in [9].

3 Graph Polynomials

In 1954, Tutte published [29] *A contribution to the theory of chromatic polynomials*. By then, he was at the University of Toronto having been recruited there in 1948 by H.S.M. Coxeter, another famous graduate of Trinity College, Cambridge. In this paper, Tutte introduced what he called the *dichromate* of a graph, this now being known as the *Tutte polynomial* of the graph. The dichromate is a two-variable polynomial not to be confused with another two-variable polynomial Tutte labelled the *dichromatic polynomial* of a graph. The latter is now known as the *Whitney rank-generating function* of the graph. Welsh [50, p. 44] draws attention to the rather confused history of these polynomials and their nomenclature. This history is clarified in [8, 9] where Tutte [49, p. 8] is quoted concerning the use of the name ‘Tutte polynomial’ as saying,

This may be unfair to Hassler Whitney who knew and used analogous coefficients without bothering to affix them to two variables.

Tutte cites Whitney's 1932 paper [51] as his source. Formally, let G be a graph with edge set E . For a subset X of E , let $G[X]$ be the subgraph of G induced by X , and let $r(X)$, the *rank* of X , be the difference between the number of vertices and the number of connected components of $G[X]$. The *Whitney rank-generating*

function $R(G; x, y)$ of G is

$$R(G; x, y) = \sum_{X \subseteq E} x^{r(E)-r(X)} y^{|X|-r(X)}.$$

The *Tutte polynomial* $T(G; x, y)$ is the translation of $R(G; x, y)$ defined by

$$T(G; x, y) = R(G; x - 1, y - 1).$$

In particular, when G is connected, $T(G; 1, 1)$ is the complexity of G , that is, its number of spanning trees. When G has $k(G)$ components, $P(G; \lambda)$, the number of proper λ -colourings of G is $\lambda^{k(G)}(-1)^{r(E)}T(G; 1 - \lambda, 0)$. Another important evaluation of the Tutte polynomial involves flows.

To define a flow in a graph G , first assign directions to every edge of G . A *nowhere-zero k -flow* assigns a flow value $f(e)$ from $\mathbb{Z}_k - \{0\}$ to every edge e of G such that, at every vertex v , the sum of the flows on the edges directed into v equals the sum of the flows on the edges directed out from v . Intuitively, Kirchhoff's Current Law holds at each vertex of G . For example, G has a nowhere-zero 2-flow if and only if every vertex has even degree. When G is connected, this is, of course, equivalent to G being Eulerian.

It is straightforward to show that if G has a nowhere-zero k -flow, then G has no cut edges. Moreover, a plane graph G without cut edges has a nowhere-zero k -flow if and only if its dual G^* is k -colourable.

Let A be an additive abelian group. A *nowhere-zero A -flow* takes flow values from $A - \{0\}$ such that, at every vertex, the flow into the vertex equals the flow out from that vertex. Thus a nowhere-zero k -flow is just a nowhere-zero \mathbb{Z}_k -flow. Remarkably, Tutte [29] showed that the number of nowhere A -flows on a graph depends only on the cardinality of A .

Proposition 1 (Tutte, 1954) *For $n \geq 2$, let A be an abelian group with n elements and G be a graph without cut edges. Then the number of nowhere-zero A -flows on G equals the number of nowhere-zero n -flows on G .*

Tutte [29] made two striking conjectures about flows.

Conjecture 1 (Tutte, 1954) There is a fixed number t such that every graph without cut edges has a nowhere-zero t -flow.

This conjecture was not settled for over 20 years until Jaeger [14] proved the following.

Theorem 1 (Jaeger, 1976) *Every graph without cut edges has a nowhere-zero 8-flow.*

Tutte's second flow conjecture is even more elusive and still remains open.

Conjecture 2 (Tutte, 1954) Every graph without cut edges has a nowhere-zero 5-flow.

The best partial result towards this 5-Flow Conjecture was proved by Seymour [19]. Just as Jaeger’s proof relied on the fact that 8 is 2 cubed, Seymour’s proof relies on 6 being the product of 3 and 2.

Theorem 2 (Seymour, 1981) *Every graph without cut edges has a nowhere-zero 6-flow.*

4 Matroids

Before discussing Tutte’s contributions to matroid theory, we briefly introduce matroids to readers unfamiliar with them.

Let A be a matrix having E as its set of column labels. Let \mathcal{I} be the collection of subsets X of E such that X labels a linearly independent set of columns. The pair (E, \mathcal{I}) is an example of a matroid M with the members of the set \mathcal{I} being its *independent sets*. We denote this matroid by $M[A]$. In general, (E, \mathcal{I}) is a *matroid* M with *ground set* E if \mathcal{I} is a non-empty hereditary collection of subsets of the finite set E with the property that, whenever X and Y are in \mathcal{I} and $|X| > |Y|$, there is an element x of $X - Y$ such that $Y \cup \{x\} \in \mathcal{I}$. Subsets of E that are not in \mathcal{I} are *dependent* and the minimal dependent sets are the *circuits* of M . Evidently, M is uniquely determined by its collection of circuits. If G is a graph, there is a matroid $M(G)$ having $E(G)$ as its ground set and the set of edge sets of cycles of G as its set of circuits. The matroid $M(G)$ is the *cycle matroid* of G .

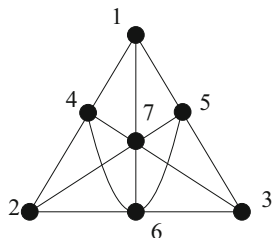
For a field \mathbb{F} , a matroid M is \mathbb{F} -*representable* if there is a matrix A over \mathbb{F} such that $M = M[A]$. A $GF(2)$ -representable matroid is called *binary*. For example, over $GF(2)$, let

$$A = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} \end{matrix}.$$

Then $M[A]$ is a matroid with ground set $\{1, 2, \dots, 7\}$ whose circuits include $\{4, 5, 6\}$ since, over $GF(2)$, the three corresponding vectors are linearly dependent although any two of them are linearly independent. This matroid is usually called the *Fano matroid* and is denoted by F_7 . A geometric representation of this matroid is shown in Fig. 1. In such a picture, three collinear points form a circuit as do four coplanar points of which no three are collinear. The dual, F_7^* , of the Fano matroid is the matroid $M[A^*]$ where

$$A^* = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}.$$

Fig. 1 The Fano matroid, F_7



In general, if the n -element matroid $M = M[I_r|D]$, its dual M^* is $M[D^T|I_{n-r}]$. More generally, suppose M is a matroid on the set E having \mathcal{B} as its set of maximal independent sets (*bases*). The collection $\{E - B : B \in \mathcal{B}\}$ can be shown to be the set of bases of a matroid on E ; this matroid M^* is the *dual* of M .

The deletion of the element 1 from F_7 is the matroid of the matrix that is obtained from A by deleting the first column. The reader may wish to check that this deletion is actually equal to $M(K_4)$ where $\{2, 3, \dots, 7\}$ is the edge set of K_4 . The contraction of 1 from F_7 is the matroid of the matrix that is obtained from A by deleting the first row and the first column. This contraction is the cycle matroid of the *doubled triangle* graph, obtained from a triangle with edge set $\{2, 6, 3\}$ by adding 4, 7, and 5 in parallel with 2, 6, and 3, respectively. In general, for a matroid M , the *deletion* of the element e from M is the matroid $M \setminus e$ having ground set $E - \{e\}$ and set of independent sets $\{I \in \mathcal{I} : e \notin I\}$. Moreover, provided $\{e\}$ is independent, the *contraction* M/e of e from M is the matroid with ground set $E - \{e\}$ and set of independent sets $\{I' \subseteq E - \{e\} : I' \cup \{e\} \in \mathcal{I}\}$. When $\{e\}$ is dependent, we define M/e to be $M \setminus e$. A *minor* of M is any matroid that can be obtained from M by a sequence of deletions and contractions. As partially outlined above, every minor of an \mathbb{F} -representable matroid is \mathbb{F} -representable. This means that the class of \mathbb{F} -representable matroids can be characterized by the matroids that are themselves not \mathbb{F} -representable but for which every minor is \mathbb{F} -representable. These minor-minimal matroids that are not \mathbb{F} -representable are the *excluded minors* for the class of \mathbb{F} -representable matroids.

Tutte wrote [53, p. 292] that his Ph.D. thesis

attempted to reduce Graph Theory to Linear Algebra. It showed that many graph-theoretical results could be generalized to algebraic theorems about structures I called ‘chain-groups’. Essentially, I was discussing a theory of matrices in which elementary operations could be applied to rows but not columns.

As Dan Younger noted in his wonderful memoir of Tutte [53, p. 292]:

This is matroid theory.

His chain-groups, called *nets* in his thesis, are essentially row spaces of representative matrices of representable matroids. In a sense, they may be regarded as *represented* matroids. But it would be pedantic to make much of the difference between these and *representable* matroids.

In essence, then, Tutte developed a theory of representable matroids as generalizations of graphs. Some of his work is valid for arbitrary matroids, in that some definitions and arguments only use matroid ideas (such as rank) in a way that does not depend on representability. But the thesis does not mention arbitrary matroids, and does not cite Whitney's seminal 1935 paper on matroids [52].

5 The Excluded-Minor Theorems

In a commentary on one of his matroid papers, Tutte wrote [46, p. 497],

If a theorem about graphs can be stated in terms of edges and circuits only it probably exemplifies a more general theorem about matroids.

The application of this principle is evident in much of Tutte's work and has guided the efforts of a number of other researchers in matroid theory. Two of the most well-known graphs are K_5 and $K_{3,3}$, the latter being the *three-houses-three-utilities graph*. These graphs are forever linked by their appearance in Kuratowski's famous characterizations [15] of planar graphs in terms of excluded (topological) minors. Tutte introduced the operation of contraction for matroids and also the notion of a minor of a matroid. In a very productive period in the late 1950s, Tutte published three important papers that included excluded-minor characterizations of various classes of matroids. This section will discuss these theorems.

Looking back on his thesis, Tutte wrote [47, p. 6],

I went on happily developing a theory of chain-groups and their elementary chains, these latter of course being defined by minimal supports. The method was to select theorems about graphs and try to generalize them to chain-groups. This was not too difficult for theorems expressible in terms of circuits. But theorems about 1-factors imposed problems. As I look back on this episode I am grieved to recall that I still did not appreciate the work of Whitney [on matroids]. Yet these chain-groups were half-way to matroids and their minimal supports were Whitney's matroid circuits.

Later in the same paper, Tutte wrote [47, p. 7],

By 1958 ... I had learned to appreciate matroids. I put the work in my thesis into matroid terminology and generalized from chain-groups to matroids. ... Then from the thesis-theorems I got the now well-known excluded minor conditions for a binary matroid to be regular and for a regular matroid to be graphic.

The uniform matroid $U_{2,4}$, which geometrically corresponds to four collinear points, is the matroid $M[A]$ where A is the real matrix

$$\begin{array}{cccc} 1 & 2 & 3 & 4 \\ \left(\begin{array}{cccc} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & -1 \end{array} \right). \end{array}$$

It is straightforward to see that $U_{2,4}$ is not binary. Tutte's first excluded-minor theorem [31], which is relatively straightforward to prove, establishes that $U_{2,4}$ is the unique excluded minor for the class of binary matroids.

Theorem 3 (Tutte, 1958) *A matroid is binary if and only if it has no $U_{2,4}$ -minor.*

A real matrix A is *totally unimodular* if the determinant of every square submatrix of A is in $\{0, 1, -1\}$. A matroid M is *regular* if there is a totally unimodular matrix A such that $M = M[A]$. As an example, take a graph G and arbitrarily orient its edges. Then take the vertex-edge incidence matrix A of this directed graph. In this real matrix, each non-zero column has one 1 and one -1 . By a result of Poincaré [17], A is totally unimodular. The matroid $M[A]$ of this matrix can be shown to be equal to the cycle matroid $M(G)$ of G . In general, a matroid is graphic if equals the cycle matroid of some graph. Thus every graphic matroid is regular. Tutte [31] proved the following.

Lemma 3 *A matroid M is regular if and only if M is \mathbb{F} -representable for all fields \mathbb{F} .*

Tutte's second excluded-minor characterization [31] is significantly more difficult than his first.

Theorem 4 (Tutte, 1958) *A matroid is regular if and only if it has none of $U_{2,4}$, F_7 , or F_7^* as a minor.*

The last theorem was proved in two papers in the *Transactions of the American Mathematical Society* called *A homotopy theorem for matroids I, II*. In a 1959 paper *Matroids and graphs* in the same journal, Tutte [33] characterized graphic matroids in terms of excluded minors. For a graph G , the dual of its cycle matroid $M(G)$ is denoted by $M^*(G)$. Recognizing the link between cycles in a plane graph and bonds in the dual graph, the reader may not be surprised to learn that the circuits of $M^*(G)$ coincide with the bonds in G . One attractive feature of $M^*(G)$ is that it is defined whether or not G is planar. Thus, although non-planar graphs do not have graphic duals, the cycle matroids of such graphs do have matroid duals.

Theorem 5 (Tutte, 1959) *A regular matroid is graphic if and only if it has neither $M^*(K_{3,3})$ nor $M^*(K_5)$ as a minor.*

Tutte wrote [47, p. 8] of this theorem that it

was guided, in the usual vague graph-to-matroid way, by Kuratowski's Theorem and my favourite proof thereof.

6 Higher Connectivity for Matroids

Whitney [52] had introduced the notion of a *non-separable* matroid as one with the property that, for every two distinct elements, there is a circuit containing both. Such a matroid is now more commonly called *connected*. A loopless graph G has the

property that every two edges lie in a cycle if and only if G is 2-connected, provided G has at least three vertices and has no isolated vertices. Given the importance of higher connectivity for graphs, it was natural to seek a matroid analogue. Tutte [40] did this. One feature of Tutte’s definition was the desire for the connectivity of a matroid and its dual to be equal. Of course, a 3-connected graph cannot have a bond of size at most two. Dually, Tutte felt that a 3-connected graph should have no cycles of size at most two; in other words, it should be simple. Tutte began his work in this area by proving the following result for graphs [34].

Theorem 6 (Tutte, 1961) *A 3-connected simple graph G has an edge e such that $G \setminus e$ or G/e is 3-connected and simple unless G is a wheel.*

Five years later, in the paper *Connectivity in matroids*, Tutte [40] generalized this theorem to matroids. Indeed, it is in his commentary [46, p. 487] on this paper that Tutte made the statement about generalizing graph results to matroids quoted at the beginning of Sect. 5. Let M be a matroid with ground set E . For a subset X of E , the rank $r(X)$ of X is the cardinality of the largest independent set that is contained in X . Earlier, we defined the rank of a set of edges in a graph G . That rank is precisely the rank of X in the cycle matroid of G .

Tutte defined the matroid M to be 2-connected if

$$r(X) + r(E - X) - r(M) \geq 1$$

for all $X \subseteq E$ with $|X|, |E - X| \geq 1$. He then defined a 2-connected matroid M to be 3-connected if

$$r(X) + r(E - X) - r(M) \geq 2$$

for all $X \subseteq E$ with $|X|, |E - X| \geq 2$.

The following result is elementary.

Proposition 2 *Let G be a graph with at least four vertices. Then*

- (i) $M(G)$ is 2-connected if and only if G is 2-connected and loopless; and
- (ii) $M(G)$ is 3-connected if and only if G is 3-connected and simple.

In the cycle matroid $M(\mathscr{W}_r)$ of the r -spoked wheel \mathscr{W}_r , the rim R is a cycle whose complement is a bond. The set R has the same size as the bases of $M(\mathscr{W}_r)$, that is, as the spanning trees of \mathscr{W}_r . Indeed, Tutte defined a new matroid \mathscr{W}^r , the rank- r whirl, on the set of edges of \mathscr{W}_r having as its bases all of the bases of $M(\mathscr{W}_r)$ together with the set R .

Theorem 7 (Tutte, 1966) *A 3-connected matroid M has an element e such that $M \setminus e$ or M/e is 3-connected unless M has rank at least three and is a whirl or the cycle matroid of a wheel.*

In 1980, Seymour [18] generalized this theorem by proving the following.

Theorem 8 (Seymour, 1980) *Let M and N be 3-connected matroids such that N is a proper minor of M . Then M has an element e such that $M \setminus e$ or M/e is 3-connected having a minor isomorphic to N unless M is a wheel or a whirl.*

7 The First Conference on Matroids

In 1964, Jack Edmonds was working at the National Bureau of Standards in Washington. He and his colleagues there organized the first conference on matroids. Tutte gave a series of *Lectures on Matroids* [36]. These appeared in the conference proceedings, which were published in the *Journal of Research of the National Bureau of Standards* in 1965. Tutte wrote [47, p. 8] about that 1964 meeting,

To me that was the year of the Coming of the Matroids. Then and there the theory of matroids was proclaimed to the mathematical world. And outside the halls of lecture there arose the repeated cry: ‘What the hell is a matroid?’

The 1965 *Journal of Research of the National Bureau of Standards* included Tutte’s paper, *Menger’s Theorem for matroids* [37]. That important paper was largely ignored for about 35 years until, in 2002, Geelen et al. [10] recognized its utility. The theorem has been used extensively since then.

For disjoint sets X and Y in a matroid M , define the *connectivity* between X and Y by

$$\kappa_M(X, Y) = \min\{r(S) + r(E - S) - r(M) : X \subseteq S \subseteq E - Y\}.$$

Theorem 9 (Tutte, 1965) *Let X and Y be disjoint sets in a matroid M . Then $\kappa_M(X, Y)$ is the maximum value of $\kappa_N(X, Y)$ over all minors N of M with ground set $X \cup Y$.*

Subsequently, Geelen et al. [11, Theorem 4.2] proved that this maximum could be restricted to minors N of M with $E(N) = X \cup Y$ such that $N|X = M|X$ and $N|Y = M|Y$. As an example, let $\{1, 2, 3\}$ and $\{4, 5, 6\}$ be the disjoint triangles in a triangular prism graph P . Then, by contracting the three edges of P that are not in triangles, we get a doubled triangle with edge set $\{1, 2, \dots, 6\}$. A consequence of Theorem 9 is that, in an arbitrary 3-connected binary matroid M , if $\{1, 2, 3\}$ and $\{4, 5, 6\}$ are disjoint 3-element circuits, then M has a minor on $\{1, 2, \dots, 6\}$ consisting of the cycle matroid of a doubled triangle.

8 Tutte’s Ph.D. Thesis

So far, we have mostly described Tutte’s *published* work on matroids. But many of his discoveries were made much earlier and were included in his remarkable Ph.D. thesis, completed in 1948 [28]. In this section, we discuss some particulars of the thesis.

In reading the thesis, it must be borne in mind that Tutte's viewpoint for matroids is dual to the usual one, so that, for example, his "circuits" in nets generalize bonds (or minimal edge cuts) of graphs, and a matroid is "graphic" if its dual is graphic in the sense defined above. Similarly, the terminology for deletion and contraction aligns with standard usage for graphs but is swapped around for nets; see the discussion in [9]. There is also much nonstandard terminology, for example, "codendroids" for bases, "dendroids" for cobases, and "cyclic elements" for blocks in graphs and components in matroids.

In Chapter III of the thesis, Tutte presents his extension of Menger's Theorem to matroids although it is not until Chapter VII that he deduces Menger's Theorem for graphs from his generalization. His paper 'Menger's theorem for matroids' was not published until 1965 [37].

Chapter IV introduces regular matroids, under the name "simple nets", approaching them from an unusual direction. Tutte then shows that a matroid of rank r on n elements is regular (according to his definition) if and only if it has an $r \times n$ representative matrix over \mathbb{Z} such that the determinant of every $r \times r$ submatrix is in $\{0, 1, -1\}$. It is routine to show that this condition is equivalent to total unimodularity of the matrix. Parts of this chapter were published and extended in [30].

Chapter V, the shortest in the thesis, is about his polynomials. It is the only chapter of the thesis that contains results he published before the thesis was completed in 1948. Its results are generalizations of a subset of those in 'A ring in graph theory' (published in 1947) [26]. Whereas [26] is restricted to graphs, this chapter of the thesis introduces polynomials for representable matroids. Instead of the V -functions of [26], we now have *chromatic functions*, which are called *Tutte invariants* or *Tutte-Grothendieck invariants* by later writers. Tutte's definition of chromatic functions only needs deletion, contraction, and the notion of a matroid component. He then extends the Whitney rank generating function to matroids, which only needs a rank function. Thus these definitions make no real use of representability, and it is reasonable to regard them as the first extension to matroids of any polynomials in the Tutte-Whitney family. It would be another 20 years until Crapo [4] formally defined the Tutte polynomial for matroids.

Tutte gives a recipe theorem for the (matroidal) Whitney rank generating function and defines, without name, the (matroidal) Tutte polynomial. An appropriate evaluation gives the number of bases, generalizing his observation for the number of spanning trees of a graph in [26]. Other evaluations give a representable-matroid analogue of counting q -colourings in a graph. Care is need with Tutte's terminology in this chapter, as discussed in [9].

Chapter VI concerns connectivity in binary matroids, extending to them the notion of a 2-separation of a graph. For graphs, some of the theory appears in [39, Ch. 11].

Having worked entirely at the level of representable matroids for Chapters II–VI, Tutte establishes the relationship with graphs in Chapter VII. He develops the theory of cycle matroids and cocycle matroids and applies the theory of the previous chapters to them. Graphic matroids and their duals are shown to be regular. The

Tutte polynomial evaluations of Chapter V are specialized to counting colourings and spanning trees. The theory of Chapter VI is applied to 2-connected graphs.

Chapters VIII–IX, occupying 140 pages, give Tutte’s excluded-minor characterization of (the duals of) graphic matroids among binary matroids. The four excluded minors are called “gnarls” and he calls his result the “gnarl theorem”. It has been the foundation and inspiration of matroid structure theory ever since, and is a fitting climax for one of the greatest doctoral theses of twentieth century mathematics.

9 The Move Away from Matroids

Although Tutte did publish some matroid papers after 1966, these later papers were in conference proceedings [43, 44] reiterating results from earlier journal papers or were supplements to earlier papers [41, 42]. Tutte’s 1966 paper *On the algebraic theory of graph colorings* [38] proposed a conjecture for binary matroids now called *Tutte’s Tangential 2-Block Conjecture*, which can be viewed as an analogue of Hadwiger’s Conjecture. The same paper included [38, p. 22] the following conjecture on 4-flows, now known as ‘*Tutte’s 4-Flow Conjecture*’. This conjecture remains open in general.

Conjecture 3 A graph without cut edges or nowhere-zero 4-flows has a Petersen-graph minor.

For cubic graphs, the last conjecture is equivalent to the assertion that every cubic graph without a cut edge or a Petersen-graph minor is 3-edge-colourable. A proof of this has been announced by Robertson, Sanders, Seymour, and Thomas. It appears in a series of papers including [6], which provides details of the other papers.

In 1981, Seymour [20] reduced Tutte’s Tangential 2-Block Conjecture to the 4-Flow Conjecture by using his decomposition theorem for regular matroids [18].

By 1967, Tutte had essentially stopped publishing new results in matroid theory. Why? Looking back on his homotopy theorem for matroids, the excluded-minor characterization of regular matroids noted above (Theorem 4), Tutte wrote [47, p. 8],

One aspect of this work rather upset me. I had valued matroids as generalizations of graphs. All graph theory, I had supposed would be derivable from matroid theory and so there would be no need to do independent graph theory any more. Yet what was this homotopy theorem, with its plucking of bits of circuit across elementary configurations, but a result in pure graph theory? Was I reducing matroid theory to graph theory in an attempt to do the opposite? Perhaps it was this jolt that diverted me from matroids back to graphs.

10 Tutte’s Contributions

Tutte’s contributions to mathematics were immense. MathSciNet credits him with 160 publications. As of May 17, 2018, MathSciNet also lists 3656 citations for his papers although it should be noted that this source primarily constructs its list for

the years 2000 onwards. From 1967, he was the Editor-in-Chief of the *Journal of Combinatorial Theory*.

Under his leadership the journal flourished. It became such a desirable place to publish that in time it was partitioned into two, series A and B, with Tutte retaining the leadership of the latter until his retirement as professor from the University of Waterloo in 1985 [53, pp. 294–95]

To this day, that journal remains preeminent in combinatorics. Tutte was a founding member of the Department of Combinatorics and Optimization at the University of Waterloo, and he had eight Ph.D. students most notably Ron Mullin and Neil Robertson.

In 2012, British Prime Minister David Cameron wrote a letter to Tutte's niece Jeanne Youlden [53, p. 286] expressing the gratitude of the United Kingdom for Tutte's codebreaking work. Cameron wrote [53, p. 286],

We should never forget how lucky we were to have men like Professor Tutte in our darkest hour and the extent to which their work not only helped protect Britain itself but also shorten the war by an estimated two years, saving countless lives.

One aspect of Tutte's creative work has yet to be touched on here. The Four invented a mathematical poetess named 'Blanche Descartes'. Any one of them could add works under her name but Tutte was believed to be the primary contributor.

The Four carefully refused to admit Blanche was their creation. Visiting Tutte's office in 1968, [Tutte's fifth Ph.D. student Arthur] Hobbs had the following conversation with him:

Hobbs: "Sir, I notice you have two copies of that proceedings. I wonder if I could buy your extra copy?"

Tutte: "Oh, no, I couldn't sell that. It belongs to Blanche Descartes." [13, p. 4]

At the conference banquet celebrating Tutte's eightieth birthday, he recited the following poem written by Ms Descartes especially for the occasion [5]. The second author, on requesting a copy of the poem from Professor Tutte, was handed the original handwritten version.

The Three Houses Problem

In central Spain in mainly rain
Three houses stood upon the plain.

The houses of our mystery
To which from realms of industry
Came pipes and wires to light and heat
And other pipes with water sweet.

The owners said, "Where these things cross
Burn, leak or short, we'll suffer loss
So let a graphman living near
Plan each from each to keep them clear."

Tell them, graphman, come in vain,
 They'll bear the cross that must remain
 Explain the planeness of the plain.

Blanche Descartes

Acknowledgements Part of this paper was presented by the second author at the Tutte Centenary Retreat (<https://www.matrix-inst.org.au/events/tutte-centenary-retreat/>) held at the MATHEMATICAL Research Institute (MATRIx), Creswick, Victoria, Australia, 26 Nov. to 2 Dec. 2017. The authors gratefully acknowledge the support of MATRIx for this retreat.

References

1. Ball, W.W.R.: *Mathematical Recreations and Essays*, 1st edn. Macmillan, London (1892). Revised and updated by H.S.M. Coxeter, 13th edn. Dover, New York (1987)
2. BBC 2011: *Code-breakers: Bletchley Park's lost heroes*. Documentary (October 2011). Producer/Director Julian Carey
3. Brooks, R.L., Smith, C.A.B., Stone, A.H., Tutte, W.T.: The dissection of rectangles into squares. *Duke Math. J.* **7**, 312–340 (1940)
4. Crapo, H.H.: The Tutte polynomial. *Aequationes Math.* **3**, 211–229 (1969)
5. Descartes, B.: The three houses problem. Recited by W.T. Tutte. Tutte Eightieth Birthday Conference Dinner, University of Waterloo, 1997
6. Edwards, K., Sanders, D.P., Seymour, P., Thomas, R.: Three-edge colouring doublecross cubic graphs. *J. Combin. Theory Ser. B* **119**, 66–95 (2016)
7. Farr, G.: Remembering Bill Tutte: another brilliant codebreaker from World War II, 12 May 2017. <http://theconversation.com/remembering-bill-tutte-another-brilliant-codebreaker-from-world-war-ii-77556>
8. Farr, G.E.: Tutte-Whitney polynomials: some history and generalizations. In: Grimmett, G., McDiarmid, C. (eds.) *Combinatorics, Complexity, and Chance*, pp. 28–52. Oxford University Press, Oxford (2007)
9. Farr, G.E., The history of Tutte-Whitney polynomials, with commentary on the classics. In: Ellis-Monaghan, J., Moffatt, I. (eds.) *Handbook of the Tutte Polynomial*. CRC Press, Boca Raton (to appear)
10. Geelen, J.F., Gerards, A.M.H., Whittle, G.: Branch-width and well-quasi-ordering in matroids and graphs. *J. Combin. Theory Ser. B* **84**, 270–290 (2002)
11. Geelen, J., Gerards, B., Whittle, G.: Excluding a planar graph from GF(q)-representable matroids. *J. Combin. Theory Ser. B* **97**, 971–998 (2007)
12. Harper, N.: *Keeping secrets*. University of Waterloo Magazine, Spring, 2015. <https://uwaterloo.ca/magazine/spring-2015/features/keeping-secrets>
13. Hobbs, A.M., Oxley, J.G.: William T. Tutte, 1917–2001. *Not. Am. Math. Soc.* **51**, 320–330 (2004)
14. Jaeger, F.: On nowhere-zero flows in multigraphs. In: Nash-Williams, C.St.J.A., Sheehan, J. (eds.) *Proceedings of the Fifth British Combinatorial Conference*, pp. 373–378. *Congressus Numerantium*, No. XV. Utilitas Mathematica, Winnipeg (1976)
15. Kuratowski, K.: Sur le problème des courbes gauches en topologie. *Fund. Math.* **15**, 271–283 (1930)
16. Oxley, J.: *Matroid Theory*. 2nd edn. Oxford University Press, New York (2011)
17. Poincaré, H.: Second complément à l'analysis situs. *Proc. London Math. Soc.* **32**, 277–308 (1900)
18. Seymour, P.D.: Decomposition of regular matroids. *J. Combin. Theory Ser. B* **28**, 305–359 (1980)

19. Seymour, P.D.: Nowhere-zero 6-flows. *J. Combin. Theory Ser. B* **30**, 130–135 (1981)
20. Seymour, P.D.: On Tutte's extension of the four-colour problem. *J. Combin. Theory Ser. B* **31**, 82–94 (1981)
21. Sprague, R.: Beispiel einer Zerlegung des Quadrats in lauter verschiedene Quadrate. *Math. Z.* **45**, 607–608 (1939)
22. Sutherland, G.B.B.M., Tutte, W.T.: Absorption of polymolecular films in the infra-red. *Nature* **144**, 707 (1939)
23. Tait, P.G.: Listing's topologie. *Philos. Mag.*, 5th Series **17**, 30–46 (1884)
24. Tutte, W.T.: On Hamiltonian circuits. *J. Lond. Math. Soc.* **21**, 98–101 (1946)
25. Tutte, W.T.: The factorization of linear graphs. *J. Lond. Math. Soc.* **22**, 107–111 (1947)
26. Tutte, W.T.: A ring in graph theory. *Proc. Camb. Philol. Soc.* **43**, 26–40 (1947)
27. Tutte, W.T.: A family of cubical graphs. *Proc. Camb. Philol. Soc.* **43**, 459–474 (1947)
28. Tutte, W.T.: An algebraic theory of graphs. Ph.D. thesis, Cambridge University (1948). <https://billtuttememorial.org.uk/links/>
29. Tutte, W.T.: A contribution to the theory of chromatic polynomials. *Canad. J. Math.* **6**, 80–91 (1954)
30. Tutte, W.T.: A class of Abelian groups. *Canad. J. Math.* **8**, 13–28 (1956)
31. Tutte, W.T.: A homotopy theorem for matroids. I, II. *Trans. Am. Math. Soc.* **88**, 144–174 (1958)
32. Tutte, W.T.: Squaring the square. In: *Scientific American, Mathematical Games Column* (November 1958). Reprinted in Gardner, M.: *More Mathematical Puzzles and Diversions*. G. Bell and Sons, London (1963) http://www.squaring.net/history_theory/brooks_smith_stone_tutte_II.html
33. Tutte, W.T.: Matroids and graphs. *Trans. Am. Math. Soc.* **90**, 527–552 (1959)
34. Tutte, W.T.: A theory of 3-connected graphs. *Nederl. Akad. Wetensch. Proc. Ser. A* **64**, 441–455 (1961)
35. Tutte, W.T.: How to draw a graph. *Proc. Lond. Math. Soc.* (3) **13**, 743–767 (1963)
36. Tutte, W.T.: Lectures on matroids. *J. Res. Nat. Bur. Stand. Sect. B* **69B**, 1–47 (1965)
37. Tutte, W.T.: Menger's theorem for matroids. *J. Res. Nat. Bur. Stand. Sect. B* **69B**, 49–53 (1965)
38. Tutte, W.T.: On the algebraic theory of graph colorings. *J. Combin. Theory* **1**, 15–50 (1966)
39. Tutte, W.T.: *Connectivity in graphs*. University of Toronto Press, Toronto (1966)
40. Tutte, W.T.: *Connectivity in matroids*. *Canad. J. Math.* **18**, 1301–1324 (1966)
41. Tutte, W.T.: A correction to: "On the algebraic theory of graph colorings". *J. Combin. Theory* **3**, 102 (1967)
42. Tutte, W.T.: On even matroids. *J. Res. Nat. Bur. Stand. Sect. B* **71B**, 213–214 (1967)
43. Tutte, W.T.: Projective geometry and the 4-color problem. In: Tutte, W.T. (ed.) *Recent Progress in Combinatorics*, pp. 199–207. Academic Press, New York (1969)
44. Tutte, W.T.: A geometrical version of the four color problem. In: Bose, R.C., Dowling, T.A. (eds.) *Combinatorial Mathematics and Its Applications*, pp. 553–560. University of North Carolina Press, Chapel Hill (1969)
45. Tutte, W.T.: *Selected Papers of W. T. Tutte/Edited by D. McCarthy, R. G. Stanton, vol. I*. Charles Babbage Research Centre, Winnipeg (1979)
46. Tutte, W.T.: *Selected Papers of W. T. Tutte/Edited by D. McCarthy, R. G. Stanton, vol. II*. Charles Babbage Research Centre, Winnipeg (1979)
47. Tutte, W.T.: The coming of the matroids. In: Lamb, J.D., Preece, D.A. (eds.) *Surveys in Combinatorics*, pp. 3–14. Cambridge University Press, Cambridge (1999)
48. Tutte, W.T.: FISH and I. In: Joyner, D. (ed.) *Coding Theory and Cryptography*, pp. 9–17. Springer, Berlin (2000)
49. Tutte, W.T.: Graph-polynomials. *Adv. Appl. Math.* **32**, 5–9 (2004)
50. Welsh, D.J.A.: *Complexity: Knots, Colourings and Counting*. Cambridge University Press, Cambridge (1993)
51. Whitney, H.: The coloring of graphs. *Ann. Math. (2)* **33**, 688–718 (1932)
52. Whitney, H.: On the abstract properties of linear dependence. *Am. J. Math.* **57**, 509–533 (1935)
53. Younger, D.H.: William Thomas Tutte, 14 May 1917–2 May 2002. *Biogr. Mem. Fellows Roy. Soc.* **58**, 283–297 (2012)

Cluster Decorated Geometric Crystals, Generalized Geometric RSK-Correspondences, and Donaldson-Thomas Transformations



Gleb Koshevoy

Abstract For a simply connected, connected, semisimple complex algebraic group G , we define two geometric crystals on the \mathcal{A} -cluster variety of double Bruhat cell $B_- \cap Bw_0B$. These crystals are related by the $*$ duality. We define the graded Donaldson-Thomas correspondence as the crystal bijection between these crystals. We show that this correspondence is equal to the composition of the cluster chamber Ansatz, the inverse generalized geometric RSK-correspondence, and transposed twist map due to Berenstein and Zelevinsky.

1 Introduction

For reductive split algebraic groups, Berenstein and Kazhdan [1] defined decorated geometric crystals. One of important feature of such a crystal is a *decoration function*. For double Bruhat cells, in relation to mirror symmetry, this decoration function have been appeared in [10] as a pullback of a Landau-Ginzburg potential defined in the cluster setup in [12] with respect to a proper map of \mathcal{A} -cluster variety to \mathcal{X} -cluster variety on the double Bruhat cells, for the Langlands dual groups.

We follow the recipes of [7, 9, 15], and endow the \mathcal{A} -cluster variety of double Bruhat cell $G^{w_0, e} := B_- \cap Bw_0B$ with two geometric crystals for Langlands dual group G^\vee , related by the $*$ duality. The Kashiwara crystal admits a duality operation $*$. One may regard the above $*$ duality as a geometric lift of the Kashiwara $*$ duality. The decoration function for the $*$ dual geometric crystal can be regarded as the pullback of the Landau-Ginzburg potential for the cluster algebra which is obtained by reversing all directions of edges in quivers of that considered in [10]. In this paper we will consider the case of simply-laced groups.

There are two actions of Cartan torus H on $G^{w_0, e}$ from the left and from the right. Under the action H from the left, we regard $G^{w_0, e}$ as $H \times B_-^{w_0}$, where $B_-^{w_0} := B_- \cap Nw_0N$ is the reduced Bruhat cell. Berenstein and Kazhdan endowed such a

G. Koshevoy (✉)

IITP RAS, MCCME, and Interdisciplinary Scientific Center J.-V. Poncelet, Moscow, Russia

reduced cell with decorated geometric crystal structure. Under the action H from the right, we regard $G^{w_0, e}$ as $N_-^{w_0} \times H$, where $N_-^{w_0} := N_- \cap Bw_0B$ is also a reduced cell. We endow $N_-^{w_0}$ with $*$ dual decorated geometric crystal structure. For the former crystal we let the frozen variables $\Delta_{w_0\omega_i, \omega_i}$, $i \in I$, be fixed (I denotes the set of vertices of the Dynkin diagram), while for the $*$ dual crystal we let be fixed another half of frozen variables $\Delta_{\omega_i, \omega_i}$, $i \in I$.

In order to obtain combinatorial crystal from geometric one, we have to consider toric charts of a positive structure and corresponding tropicalization [1].

There are several positive structures on $G^{w_0, e}$.

For one of such structures we use toric charts which constitute the Berenstein-Zelevinsky positive structures of $B_-^{w_0}$, BZ-variety, see [1] and [18]. The graded cones corresponding to the charts of such a positive structure are defined by tropicalization of the Berenstein-Kazhdan decoration function, and the cones turn out to be polyhedral realizations of the (graded) Kashiwara crystal due to Nakashima-Zelevinsky [19]. Specifically, such charts and cones correspond to the same reduced decomposition $\mathbf{i} \in R(w_0)$ of the longest element w_0 of the Weyl group. For $\mathbf{i} \in R(w_0)$, we denote such a cone $gr \mathcal{N} \mathcal{L}_{\mathbf{i}}$.

Another positive structure is related to the Lusztig variety on $N_-^{w_0}$ [4, 15]. The charts of this variety are also defined for reduced decompositions of $R(w_0)$. For $\mathbf{i} \in R(w_0)$, the tropicalization with respect to the $*$ dual potential gives polyhedral realization of the combinatorial crystal with vertices being lattice vertices of the Kashiwara $*$ dual Lusztig graded cone, $gr \mathcal{L}_{\mathbf{i}}^*$.

One more positive structure is related to the \mathcal{A} -cluster variety. Specifically, we consider only a part of cluster toric charts of \mathcal{A} -cluster variety which correspond to the reduced decompositions of $R(w_0)$. We consider two families of positive charts, for one we let to be fixed frozen $\Delta_{w_0\omega_i, \omega_i}$, $i \in I$, we call fixing frozen *specialization*, and for another we make specialization at the frozen $\Delta_{\omega_i, \omega_i}$, $i \in I$. For $\mathbf{i} \in R(w_0)$, tropicalization with respect to the corresponding chart of the former one and Landau-Ginzburg potential provides us with the polyhedral realization of the Kashiwara crystal being unimodular isomorphic to the graded Lusztig cone $gr \mathcal{L}_{\mathbf{i}}$ and tropicalization with respect to the latter one and $*$ dual LG potential gives us the polyhedral realization of the Kashiwara $*$ dual being unimodular isomorphic to the graded Littelmann cone $gr \mathcal{A}_{\mathbf{i}}$, see [10].

We provide birational positive mappings between these positive structures. For that we use birational automorphisms tori $(\mathbb{C}^*)^{l(w_0)}$ called the generalized geometric RSK-correspondence, gRSK, and its inversion (Sect. 5), two mappings from cluster tori localized at frozen coordinates called Chamber Ansatz [10] and transposed twist map of [3].

We define the *graded Donaldson-Thomas transformation* as the map which, for each reduced decomposition $\mathbf{i} \in R(w_0)$, makes the following diagram with the positive structures on $G^{w_0, e}$ commutative and tropical graded DT-transformation as that wrt the tropicalization.

$$\begin{array}{ccc}
 \boxed{\text{Lusztig-variety} \times H} & \xrightarrow{\eta_{w_0, e}^T} & \boxed{H \times \text{BZ-variety}} \\
 \uparrow \alpha & & \uparrow \beta \\
 \boxed{\text{cluster charts specialized at}} & \xrightarrow{\mathcal{DT}} & \boxed{\text{cluster charts specialized at}} \\
 \{\Delta_{\omega_i, \omega_i}\}_{i \in I} & & \{\Delta_{w_0 \omega_i, \omega_i}\}_{i \in I}
 \end{array} \tag{1}$$

where $\eta_{w_0, e}^T$ is transposition of the twist map defined in [3, Definition 4.1], α is the composition of CA^- and inverse gRSK, β is the composition of CA^+ and gRSK, where CA^+ denotes the tuples maps $grCA_{\mathbf{i}}$, $\mathbf{i} \in R(w_0)$, and CA^- denotes the inverse of $grNA_{\mathbf{i}}$, defined in [10, Definition 6.1 and 7.1], for details see Sects. 5 and 6. The latter mappings are motivated by the Chamber Ansatz [4] for the Lusztig- and Berenstein-Zelevinsky-parametrizations of $N_-^{w_0}$ and $B_-^{w_0}$, respectively.

The twist $\eta_{w_0, e}^T$ is a crystal bijection sending $N_- \cap Bw_0B \times H$ to $H \times B_- \cap Nw_0N$. Since all vertical maps are crystal isomorphism we get that the graded Donaldson-Thomas transformation is an isomorphism of geometric cluster crystals.

In other words, the graded Donaldson-Thomas transformation is the composition of five maps, the inverse generalized geometric RSK and CA^- , sending cluster variety specialized at the half of frozen $\Delta_{\omega_i, \omega_i}$, $i \in I$, to the graded Lusztig variety, both endowed with $*$ -dual geometric crystal structure, then transposed twist map which sends the Lusztig variety to the Berenstein-Zelevinsky variety, where the latter is endowed with the geometric crystal as in [1, 18], and finally the inverse generalized geometric RSK and inverse CA^+ sending BZ-variety to the cluster variety specialized at another half of frozen $\Delta_{\omega_i, \omega_i}$, $i \in I$.

The tropical graded DT-transformation is as in [10]. Specifically, tropicalization of the above diagram leads to the definition of a *tropical DT-transformation* which makes the following diagram commutative

$$\begin{array}{ccc}
 \boxed{gr\mathcal{L}_{\mathbf{i}}^*} & \xrightarrow{\text{tropical BZ-twist}} & \boxed{gr\mathcal{N}\mathcal{L}_{\mathbf{i}}} \\
 \uparrow \text{tropical inverse RSK} & & \uparrow \text{tropical RSK} \\
 \boxed{gr\mathcal{S}_{\mathbf{i}}} & \xrightarrow{\text{tropical } \mathcal{DT}} & \boxed{gr\mathcal{L}_{\mathbf{i}}}
 \end{array} \tag{2}$$

The mappings between the SW-NE and NW-SE corners of diagram (1) are geometric lifting of the Kashiwara $*$ dual crystal isomorphisms between corresponding corners of the diagram (2).

Thus we may regard $*$ duality (Kashiwara $*$ duality) as the composition of the transposed twist, the inverse generalized geometric RSK correspondence, and inverse CA^+ (tropical twist, tropical gRSK, and tropical inverse CA^+).

Goncharov and Shen [11] conjectured that the Donaldson-Thomas transformation, defined for \mathcal{X} -cluster variety, is the twist $\eta_{w_0, e}$ under specialization at all frozen variables. This conjecture is proven in [20].

Our graded Donaldson-Thomas transformation is the composition of transposed twist and maps α and β^{-1} .

2 Preliminary and Notations

2.1 Simply-Connected Algebraic Groups

For a simply connected, connected, semisimple complex algebraic group G , let B and B_- be Borel subgroup and its opposite, N and N^- be unipotent radicals, a maximal torus $H = B \cap B_-$, and $W = NormG(H)/H$ be the Weyl group. Let $A = (a_{ij})_{i, j \in I}$ be the Cartan matrix, cardinality of I , $|I|$, equals the rank of G . The Weyl group W is canonically identified with the Coxeter group generated by the involutions $s_1, \dots, s_{|I|}$, subject to the relations $(s_i s_j)^{d_{ij}} = e$, $d_{ij} = 0, 3, 4, 6$ if $a_{ij} = 0, 1, 2, 3$, respectively. A reduced decomposition of $w \in W$ is a word $\mathbf{i} = (i_1 \cdots i_l)$ in the alphabet I , such that $w = s_{i_1} \cdots s_{i_l}$ gives a factorization of smallest length. The length l is called the length of w and denoted by $l(w)$. For $w \in W$, the set of all reduced decompositions is denoted by $R(w)$. We denote by w_0 the element of maximal length in W . Any two reduced decompositions $\mathbf{i}, \mathbf{i}' \in R(w)$ are related by the Artin relations. For simply-laced cases, Artin relations are 2-moves and 3-moves. Specifically, a reduced word $\mathbf{j} = (j_1, \dots, j_l)$ is defined to be obtained from $\mathbf{i} = (i_1, \dots, i_l)$ by a 2-move at position $k \in [l - 1]$ if $i_\ell = j_\ell$ for all $\ell \notin \{k, k + 1\}$, $(i_{k+1}, i_k) = (j_k, j_{k+1})$ and $a_{i_k, i_{k+1}} = 0$.

A reduced word \mathbf{j} is defined to be obtained from \mathbf{i} by a 3-move at position $k \in [l - 1]$ if $i_\ell = j_\ell$ for all $\ell \notin \{k - 1, k, k + 1\}$, $j_{k-1} = j_{k+1} = i_k$, $j_k = i_{k-1} = i_{k+1}$ and $a_{i_k, i_{k+1}} = -1$.

Let \mathfrak{g} be the Lie algebra of G , and \mathfrak{h} the Cartan subalgebra. Let $\{\alpha_1, \dots, \alpha_{|I|}\} \subset \mathfrak{h}^*$ be simple roots for which the corresponding root subgroups are contained in N . For $i \in I$, let ϕ_i be the homomorphism $SL_2 \rightarrow G$ corresponding to the i th simple root of G . Given $i \in I$ define

$$x_i(t) = \phi_i \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \quad y_i(t) = \phi_i \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}, \quad \bar{s}_i = \phi_i \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix},$$

and

$$x_{-i}(p) = \phi_i \begin{pmatrix} \frac{1}{p} & 0 \\ 1 & p \end{pmatrix}.$$

$$\alpha_i^\vee(c) = \phi_i \begin{pmatrix} c & 0 \\ 0 & c^{-1} \end{pmatrix}, \quad \alpha_i(\mathbf{h}) = \frac{h_i}{h_{i+1}}, \quad i \in I.$$

2.2 Cluster Seeds Associated to Reduced Decompositions

Recall that, for a dominant weight $\lambda : H \rightarrow \mathbb{G}_m$, the *principal minor* $\Delta_\lambda : G \rightarrow \mathbb{A}^1$ is the function defined on the open subset $N^-HN \subset G$ by

$$\Delta_\lambda(u^- \mathbf{h} u^+) := \lambda(\mathbf{h}) \quad u^- \in N^-, \mathbf{h} \in H, u^+ \in N.$$

Let γ, δ be extremal weights such that $\gamma = w_1\lambda, \delta = w_2\lambda$ for some $w_1, w_2 \in W, \lambda \in P^+$. The *generalized minor* associated to γ and δ is

$$\Delta_{\gamma,\delta}(g) := \Delta_\lambda(\bar{w}_1^{-1} g \bar{w}_2), \quad g \in G,$$

where \bar{w} is a lift of W into $Norm_G H$ using $\bar{s}_i, i \in I$

The base affine space G/N is the partial compactification of the *open double Bruhat cell*

$$G^{w_0,e} := Bw_0B \cap B_-$$

obtained by allowing the generalized minors $\Delta_{\omega_a, \omega_a}$ and $\Delta_{w_0\omega_a, \omega_a}$ to vanish.

Here we need a small part of cluster seeds of the \mathcal{A} -variety. Namely, for a reduced decomposition $\mathbf{i} \in R(w_0)$, we consider the corresponding seed $\mathcal{S}(\mathbf{i})$ follow [5]. The vertices of the quiver $Q(\mathbf{i})$ are labeled by the fundamental weights $\omega_i, i \in I$, and $\mathbf{i}|_{\leq k} \omega_{i_k}, k \in [l(w_0)], \mathbf{i}|_{\leq k}$ denotes the subword \mathbf{i} of the first k letters.

The frozen vertices are labeled by the fundamental weights $\omega_i, i \in I$, and $w_0\omega_i, i \in I$.

The cluster variables of $\mathcal{S}(\mathbf{i})$ are the generalized minors $\Delta_{\mathbf{i}|_{\leq k} \omega_{i_k}, \omega_{i_k}}$ attached to vertices labeled by $\mathbf{i}|_{\leq k} \omega_{i_k}$.

Follow [5] we associate to every reduced word \mathbf{i} a seed $\Sigma(\mathbf{i})$. The set of edges quiver $\Gamma_{\mathbf{i}}$ is described as follows. For $k \in [-n]$ we set $i_k = -k$. For $k \in [l(w_0)]$ we denote by $k^+ = k_{\mathbf{i}}^+$ the smallest ℓ such that $k < \ell$ and $i_\ell = i_k$. If no such ℓ exists, we set $k^+ = l(w_0) + 1$. For $k \in [l(w_0)]$, we further let k^- be the largest index ℓ with that $\ell < k$ and $i_\ell = i_k$.

There is an edge connecting v_k and v_ℓ with $k < \ell$ if at least one of the two vertices is mutable and one of the following conditions is satisfied:

1. $\ell = k^+$,
2. $\ell < k^+ < \ell^+$, $c_{k,\ell} < 0$ and $k, \ell \in [N]$.

Edges of type (1) are called *horizontal* and are directed from k to ℓ . Edges of type (2) are called *inclined* and are directed from ℓ to k .

We need the following fact. Let $\mathbf{j} \in R(w_0)$ be obtained from \mathbf{i} by a 3-move in position k . Then the transposition $(k, k+1)$ is an isomorphism of quivers $\Gamma_{\mathbf{j}} \simeq \mu_k \Gamma_{\mathbf{i}}$, where μ_k is a mutation at the vertex labeled by $\mathbf{i}|_{\leq k} \omega_{i_k}$.

The new variables is obtained by the \mathcal{A} -cluster mutation

$$\mu_k A_\ell = \begin{cases} \frac{\prod_{\ell: (\ell,k) \in \Gamma(\mathbf{j})} A_\ell}{A_k} + \frac{\prod_{m: (k,m) \in \Gamma(\mathbf{j})} A_m}{A_k} & \text{if } \ell = k, \\ A_\ell & \text{else,} \end{cases}$$

For reduced seeds corresponding to reduced words, such cluster mutation take the form of Plücker relations between generalized minors.

3 Cluster Geometric Crystals

On the \mathcal{A} -cluster variety $G^{w_0,e}$, we define two geometric crystals (for Langlands dual group) related by the Kashiwara $*$ -involution.

3.1 Geometric Crystal for \mathcal{A} -Variety Specialized at $\Delta_{w_0 \omega_i, \omega_i}$'s

We consider simply-laced case and define the main ingredients of the geometric crystal on the \mathcal{A} -cluster variety G^r obtained of $G^{w_0,e}$ by the specialization at the frozen variables $\Delta_{w_0 \omega_i, \omega_i}$, $i \in I$.

For $k \in I$, we denote by \mathbf{i}_k a reduced decomposition which starts with s_k , we call such a reduced decomposition *optimal from the head for k* .

For such an optimal reduced decomposition \mathbf{i}_k , we consider the corresponding seed $\mathcal{S}(\mathbf{i}_k)$.

We define the crystal actions $f_k(c, \dots) : \mathbb{C}^* \times G^{w_0,e} \rightarrow G^{w_0,e}$, $k \in I$, by specifying it on the variables of the seed $\mathcal{S}(\mathbf{i}_k)$. Namely, we set

$$f_k(c, \cdot) : \Delta_{\omega_k, \omega_k} \rightarrow c \Delta_{\omega_k, \omega_k}, \tag{3}$$

and $f_k(c, \cdot)$ does not change other generalized minors labeling nodes of $\mathcal{S}(\mathbf{i}(k))$.

In order to get the action of another crystal operation $f_j(c, \cdot)$ on variables of this seed, firstly, we have to express the cluster variables of $\mathcal{S}(\mathbf{i}(k))$ as Laurent

polynomials of cluster variables $\mathcal{S}(\mathbf{i}(l))$, secondly we have to apply $f_l(c, \cdot)$ to variables of these Laurent polynomials, and then to express such obtained Laurent polynomials in the variables of $\mathcal{S}(\mathbf{i}(k))$.

Because of refinement of the Laurent phenomenon for cluster algebras [8], that claims that frozen variables do not appear in denominators of Laurent polynomials expressing a cluster variable of one seed in the variables of another, we get that the crystal operations take the form of Laurent polynomials, indeed.

Note that the frozen variables $\Delta_{w_0\omega_i, \omega_i}$, $i \in I$, do not change under any of such crystal actions. This is a reason to specialize the cluster algebra at these frozen variables.

We take the potential $\Phi_{BK} : G^{w_0, e} \rightarrow \mathbb{C}$, as the decoration function due to Berenstein and Kazhdan [1]

$$\Phi_{BK}(M) = \sum_{i \in I} \frac{\Delta_{w_0\omega_i, s_i\omega_i}(M)}{\Delta_{w_0\omega_i, \omega_i}(M)} + \sum_{i \in I} \frac{\Delta_{w_0s_i\omega_i, \omega_i}(M)}{\Delta_{w_0\omega_i, \omega_i}(M)}, \quad M \in G^{w_0, e}. \tag{4}$$

For a group G with simply-laced Lie algebra, it follows from [10], that, for each k and any reduced decomposition $\mathbf{i}(k)$ optimal for k , we have

$$\begin{aligned} & \Phi_{BK}(f_k(c, M_{\mathbf{i}(k)})) - \Phi_{BK}(M_{\mathbf{i}(k)}) \\ &= (c - 1) \frac{\Delta_{\omega_k, \omega_k}(M_{\mathbf{i}(k)})}{\Delta_{s_k\omega_k, \omega_k}(M_{\mathbf{i}(k)})} + \left(\frac{1}{c} - 1\right) \frac{\Delta_{\omega_{k-1}, \omega_{k-1}}(M_{\mathbf{i}(k)}) \Delta_{\omega_{k+1}, \omega_{k+1}}(M_{\mathbf{i}(k)})}{\Delta_{s_k\omega_k, \omega_k}(M_{\mathbf{i}(k)}) \Delta_{\omega_k, \omega_k}(M_{\mathbf{i}(k)})}. \end{aligned} \tag{5}$$

where $M_{\mathbf{i}(k)}$ a toric chart of the \mathcal{A} cluster variety $G^{w_0, e}$ written in cluster variables of $\mathcal{S}(\mathbf{i}(k))$.

For SL_n , this means the following. We consider matrix elements of $M \in G^{w_0, e}$, as Laurent polynomials which express $\frac{\Delta_{s_{i-1} \dots s_j \omega_j, \omega_j}}{\Delta_{\omega_{j-1}, \omega_{j-1}}}$, in variables of the cluster seed $\mathcal{S}(\mathbf{i}(k))$. $M_{\mathbf{i}(k)}$ denotes such a representation of matrix elements.

Because of that if we consider a point of the \mathcal{A} -variety, that is a collection of tuples, related by the cluster mutations, then each tuple of the collection defines the same matrix.

Because of Positivity Theorem [14], these matrix elements are Laurent polynomials with non-negative coefficients.

We define the functions φ , ε and γ being geometric lifting of the Kashiwara functions as follows.

For the seed $\mathcal{S}(\mathbf{i}(k))$, we set

$$\begin{aligned} \varphi_k(M_{\mathbf{i}(k)}) &= \frac{\Delta_{s_k\omega_k, \omega_k}(M_{\mathbf{i}(k)})}{\Delta_{\omega_k, \omega_k}(M_{\mathbf{i}(k)})}, \\ \varepsilon_k(M_{\mathbf{i}(k)}) &= \frac{\Delta_{s_k\omega_k, \omega_k}(M_{\mathbf{i}(k)}) \Delta_{\omega_k, \omega_k}(M_{\mathbf{i}(k)})}{\Delta_{\omega_{k-1}, \omega_{k-1}}(M_{\mathbf{i}(k)}) \Delta_{\omega_{k+1}, \omega_{k+1}}(M_{\mathbf{i}(k)})}, \text{ and} \\ \alpha_k(\gamma(M_{\mathbf{i}(k)})) &:= \frac{\varphi_k}{\varepsilon_k}. \end{aligned} \tag{6}$$

Thus we have,

$$\alpha_k(\gamma(M_{\mathbf{i}(k)})) = \frac{\Delta_{\omega_k, \omega_k}(M_{\mathbf{i}(k)})^2}{\Delta_{\omega_{k-1}, \omega_{k-1}}(M_{\mathbf{i}(k)})\Delta_{\omega_{k+1}, \omega_{k+1}}(M_{\mathbf{i}(k)})}.$$

Because of the above formula and that all seeds before action of operations f_k 's has the same frozen variables, we get that γ does not depend on a seed and

$$\alpha_k(\gamma(M)) = \frac{\Delta_{\omega_k, \omega_k}(M)^2}{\Delta_{\omega_{k-1}, \omega_{k-1}}(M)\Delta_{\omega_{k+1}, \omega_{k+1}}(M)}, \quad M \in G^{w_0, e}. \quad (7)$$

Note that we can also regard functions φ and ε independently of cluster seeds,

$$\begin{aligned} \varphi_k(M) &= \frac{\Delta_{s_k \omega_k, \omega_k}(M)}{\Delta_{\omega_k, \omega_k}(M)}, \\ \varepsilon_k(M) &= \frac{\Delta_{s_k \omega_k, \omega_k}(M)\Delta_{\omega_k, \omega_k}(M)}{\Delta_{\omega_{k-1}, \omega_{k-1}}(M)\Delta_{\omega_{k+1}, \omega_{k+1}}(M)}. \end{aligned}$$

From the refined Laurent phenomenon, we get that, for any cluster seed, the functions φ and ε are Laurent polynomials in variables of that seed.

3.2 SL_3

For example for SL_3 and a cluster seed, corresponding to a reduced word 121, let us denote the cluster variables $t_1 = \Delta_{\omega_1, \omega_1}$, $t_2 = \Delta_{s_1 \omega_1, \omega_1}$, $t_3 = \Delta_{s_2 s_1 \omega_1, \omega_1}$, $t_{12} = \Delta_{\omega_2, \omega_2}$, $t_{23} = \Delta_{w_0 \omega_2, \omega_2}$.

Then elements of the corresponding cluster chart are matrices of the form

$$M_{121} := \begin{pmatrix} t_1 & 0 & 0 \\ t_2 & \frac{t_{12}}{t_1} & 0 \\ t_3 & \frac{t_1 t_{23} + t_3 t_{12}}{t_1 t_2} & \frac{1}{t_{12}} \end{pmatrix}$$

and, since the 121 is optimal for s_1 , the action $f_1(c, \cdot)$ is

$$f_1(c, M_{121}) = \begin{pmatrix} ct_1 & 0 & 0 \\ t_2 & \frac{t_{12}}{ct_1} & 0 \\ t_3 & \frac{ct_1 t_{23} + t_3 t_{12}}{ct_1 t_2} & \frac{1}{t_{12}} \end{pmatrix}$$

Then the potential Φ_{BK} computed in variables of this cluster chart is

$$\Phi_{BK}^{121} = \frac{t_{12}}{t_1 t_2} + \frac{t_2}{t_{12} t_{23}} + \frac{t_{23}}{t_2 t_3} + \frac{t_1}{t_2} + \frac{t_{12} t_3}{t_{23} t_2} + \frac{t_2}{t_3}.$$

The functions $\varphi_1(M_{121}) = \frac{t_2}{t_1}$, $\varepsilon_1(M_{121}) = \frac{t_1 t_2}{t_{12}}$ and $\gamma_1 = \frac{t_1^2}{t_{12}}$.

In order to have the action $f_2(c, \cdot)$ and to compute $\varphi_2(M_{121})$ and $\varepsilon_2(M_{121})$, we have to represent M_{121} in the cluster coordinates of the chart for the reduced decomposition 212. For 212, we get

$$M_{212} := \begin{pmatrix} t_1 & 0 & 0 \\ \frac{t_1 t_{23} + t_3 t_{12}}{t_{13}} & \frac{t_{12}}{t_1} & 0 \\ t_3 & \frac{t_{13}}{t_1} & \frac{1}{t_{12}} \end{pmatrix},$$

where $t_{13} = \Delta_{s_2 \omega_2, \omega_2}$ and there due to the Plücker we have $t_{13} t_2 = t_1 t_{23} + t_3 t_{12}$, where $t_1, t_2, t_3, t_{12}, \dots, t_{23}$ are as above, and $t_1, t_3, t_3, t_{12}, \dots, t_{23}$ are the cluster variables of $\mathcal{S}(212)$.

Then the action $f_2(c, \cdot)$ at the chart M_{212} , corresponding to the seed $\mathcal{S}(212)$, takes the form

$$f_2(c, M_{212}) := \begin{pmatrix} t_1 & 0 & 0 \\ \frac{t_1 t_{23} + c t_3 t_{12}}{t_{13}} & c \frac{t_{12}}{t_1} & 0 \\ t_3 & \frac{t_{13}}{t_1} & \frac{1}{c t_{12}} \end{pmatrix},$$

and, hence, $f_2(c, \cdot)$ acts in the chart for 121 as follows

$$f_2(c, M_{121}) = \begin{pmatrix} t_1 & 0 & 0 \\ t_2 + \frac{(c-1)t_3 t_{12} t_2}{t_1 t_{23} + t_3 t_{12}} & c \frac{t_{12}}{t_1} & 0 \\ t_3 & \frac{t_1 t_{23} + t_3 t_{12}}{t_1 t_2} & \frac{1}{c t_{12}} \end{pmatrix}.$$

Hence we get $\varphi_2(M_{121}) = \frac{t_1 t_{23} + t_3 t_{12}}{t_{12} t_2}$, $\varepsilon_2(M_{121}) = \frac{t_{12}(t_1 t_{23} + t_3 t_{12})}{t_{12}}$, and $\gamma_2(M) = \frac{t_{12}^2}{t_1}$.

Note the potential Φ_{BK} computed in variables of cluster chart for 212 is

$$\Phi_{BK}^{212} = \frac{t_1}{t_{12} t_{13}} + \frac{t_{13}}{t_1 t_2} + \frac{t_3}{t_{13} t_{23}} + \frac{t_1}{t_3} + \frac{t_1 t_{23}}{t_3 t_{13}} + \frac{t_{12}}{t_{13}}.$$

3.3 Cluster Charts and Geometric Crystals

Denote by $G^r(\mathbf{h})$ a leaf of $G^{w_0, e}$ obtained by fixing $\Delta_{w_0 \omega_i, \omega_i} =: h_i, i \in I$. Namely, for a cluster seed, corresponding to a reduced word $\mathbf{i} \in R(w_0)$, we regard matrix $M_{\mathbf{i}}$ as product

$$M_{\mathbf{i}} = \alpha_1^\vee \left(\frac{1}{\Delta_{w_0 \omega_{|I|}} } \right) \alpha_2^\vee \left(\frac{1}{\Delta_{w_0 \omega_{|I|-1}, \omega_{|I|-1}} } \right) \cdots \alpha_{|I|-1}^\vee \left(\frac{1}{\Delta_{w_0 \omega_1, \omega_1}} \right) \tilde{M}_{\mathbf{i}},$$

where $\tilde{M}_{\mathbf{i}} \in G^r(\mathbf{1})$.

Theorem 1 For any $\mathbf{h} \in H$, the crystal operations defined by the rule (3), the decoration function defined by (4), the functions $\varphi, \varepsilon, \gamma$ defined by (6), define a geometric crystal on the \mathcal{A} -cluster variety $G^r(\mathbf{h})$, in the sense of [1].

Proof We have to verify that the crystal operations satisfy the Verma relations and proper behavior of the above functions under the crystal actions.

- a) The claim that the crystal actions $f_k, k \in I$, satisfy the Verma relations (quantum Yang-Baxter equation¹)

$$f_k^a(f_{k'}^{ab}(f_k^b)) = f_{k'}^b(f_k^{ab}(f_{k'}^a))$$

if k and k' are joined by an edge in the Dynkin diagram and commute elsewhere,

can be reduced to that claim for the case of SL_3 (see [7]). The later case is straightforward computation.

- b) The relation of the decoration function Φ_{BK} and the Kashiwara functions φ and ε , takes the form

$$\Phi_{BK}(f_k(c, M)) - \Phi_{BK}(M) = \frac{c - 1}{\varphi_k(M)} + \frac{1/c - 1}{\varepsilon_k(M)}. \tag{8}$$

For simply-laced groups, $k \in I$ and a seed $\mathcal{S}(\mathbf{i}(k))$, $\mathbf{i}(k) \in R(w_0)$ and is optimal for k , (8) follows from [9, 10].

The claim that relations between ε, φ and γ fulfill the requirements of [1] also follows from the above claim. □

4 * Dual Geometric Crystal

We define the * dual geometric crystal on \mathcal{A} -variety G^l , obtained from $G^{w_0 \cdot e}$ by the specialization at the frozen variables $\Delta_{\omega_i, \omega_i}, i \in I$.

The Kashiwara crystal admits a duality operation * (see, for example, [13]), and one may regard such * dual geometric crystal as a geometrization the Kashiwara duality.

Namely, for $a \in I$, a reduced decomposition \mathbf{i}^a is *optimal from the tail for a*, if i_a is the last element of \mathbf{i}^a . One can consider \mathbf{i}^a as reversed $\mathbf{i}(a)$ with $s_{w_0(j)}$ replacing s_j .

Let us consider the corresponding seed $\mathcal{S}(\mathbf{i}^a)$.

¹Integrable system related to this quantum Yang-Baxter equation is Toda lattice, and we will come to this issue in another paper.

For such a reduced decomposition \mathbf{i}^a , we define the action of $f_{w_0(a)}^*(c, \cdot)$ on the variables of the seed $\mathcal{S}(\mathbf{i}^a)$ by acting only on frozen variable

$$\Delta_{w_0\omega_a, \omega_a} \rightarrow c \Delta_{w_0\omega_a, \omega_a},$$

of this seed and does not changing other cluster variables of $\mathcal{S}(\mathbf{i}^a)$.

To define $f_{w_0(a)}^*(c, \dots)$ in another seed \mathcal{S} , we have to mutate from \mathcal{S} to $\mathcal{S}(\mathbf{i}^a)$, than apply $f_{w_0(a)}(c, \cdot)$ on $\mathcal{S}(\mathbf{i}^a)$, and than mutate back to \mathcal{S} .

Remark that the frozen variables $\Delta_{\omega_i, \omega_i}$, $i \in I$, do not change under all such crystal actions. Because of that we make specialization at these frozen.

To define all functions for a geometric crystal, we firstly define the decoration function

$$\Psi_{*K}(M) := \sum_{i \in I} \frac{\Delta_{\omega_i, s_i \omega_i}(M)}{\Delta_{\omega_i, \omega_i}(M)} + \sum_{i \in I} \frac{\Delta_{w_0\omega_i, s_i \omega_i}(M)}{\Delta_{w_0\omega_i, \omega_i}(M)}, \quad M \in G^{w_0, e}. \tag{9}$$

Then, in the seed $\mathcal{S}(\mathbf{i}^a)$ we get the following functions

$$\varphi_{w_0(a)}^*(M_{\mathbf{i}^a}) = \frac{\Delta_{w_0s_a\omega_a, \omega_a}(M_{\mathbf{i}^a})}{\Delta_{w_0\omega_a, \omega_a}(M_{\mathbf{i}^a})}, \tag{10}$$

$$\varepsilon_{w_0(a)}^*(M_{\mathbf{i}^a}) = \frac{\Delta_{w_0s_a\omega_a, \omega_a}(X) \Delta_{w_0\omega_a, \omega_a}(M_{\mathbf{i}^a})}{\Delta_{w_0\omega_{a-1}, \omega_{a-1}}(M_{\mathbf{i}^a}) \Delta_{w_0\omega_{a+1}, \omega_{a+1}}(M_{\mathbf{i}^a})} \tag{11}$$

$$\alpha_{w_0(k)}(\gamma^*(M)) = \frac{\Delta_{w_0\omega_k, \omega_k}(M)^2}{\Delta_{w_0\omega_{k-1}, \omega_{k-1}}(M) \Delta_{w_0\omega_{k+1}, \omega_{k+1}}(M)}, \quad M \in X. \tag{12}$$

Note that γ^* is the ‘highest weight’ for the Kashiwara geometric crystal with the potential Φ_{BK} .

For SL_n and $\mathbf{i} \in R(w_0)$, we have the following relations, which shows symmetry of weights and highest weights on the language of geometric crystals,

$$\alpha_k(\gamma(M_{\mathbf{i}}))\alpha_k(\gamma^*(M_{\mathbf{i}})) = \prod_{\rho \in T(k)} t_m^{\text{sign} \rho \cdot \chi(m, \rho)} \prod_{\rho \in T(k+1)} t_m^{\text{sign} \rho \cdot \chi(m, \rho)}, \tag{13}$$

where $T(k)$ is a train track colored by k in the rhombus tiling for \mathbf{i} , $\chi(m, \rho)$ is the delta function of positive roots labeled by m -th cluster variable and the tile ρ , and $\mathbf{t} := CA^+(\mathcal{S}(\mathbf{i}))$ (for details see [9]). Note that symmetry between γ and γ^* breaks when we choose from what side the Cartan torus acts on B_- . Another relations between weights and highest weights is

$$\alpha_k(\gamma(M_{\mathbf{i}}))\alpha_{w_0(k)}(\gamma^*(M_{\mathbf{i}})) = \prod_{m \in I(k)} q_{i_m}^2 \prod_{m' \in I(k') : a_{i_m, k} = -1} q_{i_{m'}}, \tag{14}$$

where, for \mathbf{i} , $I(k) = \{j : i_j = k\}$, $\mathbf{q} := CA^-(\mathcal{S}(\mathbf{i}))$.

Denote by $G^l(\hat{\mathbf{h}})$ a leaf of $G^{w_0, e}$ with fixed $\Delta_{\omega_i, \omega_i} =: \hat{h}_i, i \in I$.

Theorem 2 For each $\hat{\mathbf{h}}$, the above defined crystal actions $f_a^*, a \in I$, the decoration (9), and the functions (10)–(12) define a geometric crystal on the \mathcal{A} -cluster variety $G^l(\hat{\mathbf{h}})$.

Proof For the Verma relations, it suffices to check for SL_3 , and this is a rather straightforward. Then the relations among the actions, decorations and the functions in order to fulfill the axioms of the geometric crystal, follows from the above property of the potential Ψ_{*K} in each seed tail optimal for $a \in I$. \square

4.1 SL_3

For SL_3 , the cluster seed for 121 is tail optimal for 1 and hence for the second action f_2^* .

Thus, we have

$$f_2^*(c, M_{121}) = \begin{pmatrix} t_1 & 0 & 0 \\ t_2 & \frac{t_{12}}{t_1} & 0 \\ ct_3 & \frac{t_1 t_{23} + ct_3 t_{12}}{t_1 t_2} & \frac{1}{t_{12}} \end{pmatrix}$$

The cluster seed for 212 is tail-optimal for 2 and hence, is optimal for the crystal action f_1^* , we have

$$f_1^*(c, M_{212}) := \begin{pmatrix} t_1 & 0 & 0 \\ \frac{ct_1 t_{23} + t_3 t_{12}}{t_{13}} & \frac{t_{12}}{t_1} & 0 \\ t_3 & \frac{t_{13}}{t_1} & \frac{1}{ct_{12}} \end{pmatrix},$$

and, hence, $f_1^*(c, \cdot)$ acts in the chart for 121 as follows

$$f_1^*(c, M_{121}) = \begin{pmatrix} t_1 & 0 & 0 \\ t_2 + \frac{(c-1)t_1 t_{23} t_2}{t_1 t_{23} + t_3 t_{12}} & \frac{t_{12}}{t_1} & 0 \\ t_3 & \frac{t_1 t_{23} + t_3 t_{12}}{t_1 t_2} & \frac{1}{ct_{12}} \end{pmatrix}.$$

Note that in the cluster chart for 121, the specialization lead to the right action of H of the form

$$\begin{pmatrix} \frac{t_1}{t_3} & 0 & 0 \\ \frac{t_2}{t_3} & \frac{t_{12} t_3}{t_1 t_{23}} & 0 \\ 1 & \frac{(t_1 t_{23} + t_3 t_{12}) t_3}{t_1 t_2 t_3} & \frac{t_{23}}{t_{12}} \end{pmatrix} \cdot \begin{pmatrix} t_3 & 0 & 0 \\ 0 & \frac{t_{23}}{t_3} & 0 \\ 0 & 0 & \frac{1}{t_{23}} \end{pmatrix}.$$

For cluster chart for 121, the decoration Ψ_{*K} is of the form

$$\Psi_{*K}^{121} = \frac{t_4}{t_1 t_2} + \frac{t_2}{t_4 t_5} + \frac{t_5}{t_2 t_3} + \frac{t_2}{t_1} + \frac{t_5 t_1}{t_4 t_2} + \frac{t_3}{t_2}.$$

Restricted to the leaf for $t_1 =: \hat{h}_1, t_4 =: \hat{h}_2$, we get

$$\Psi^{121}(\hat{h}_1, \hat{h}_2) = \frac{\hat{h}_2}{\hat{h}_1 t_2} + \frac{t_2}{t_5 \hat{h}_2} + \frac{t_5}{t_2 t_3} + \frac{t_2}{\hat{h}_1} + \frac{t_5 \hat{h}_1}{\hat{h}_2 t_2} + \frac{t_3}{t_2}.$$

For the cluster seed, corresponding to a reduced word 212, we have the cluster variables $t'_1 = \Delta_{\omega_1, \omega_1}, t'_4 = \Delta_{s_2 \omega_2, \omega_2}, t'_2 = \Delta_{s_2 s_1 \omega_1, \omega_1}, t'_3 = \Delta_{\omega_2, \omega_2}, t'_5 = \Delta_{w_0 \omega_2, \omega_2}$. For this seed, Ψ_{*K}^{212} is of the form

$$\Psi_{*K}^{212} = \frac{t'_1}{t'_4 t'_3} + \frac{t'_4}{t'_1 t'_2} + \frac{t'_2}{t'_4 t'_5} + \frac{t'_4}{t'_3} + \frac{t'_2 t'_3}{t'_1 t'_4} + \frac{t'_5}{t'_4}.$$

Note that Ψ_{*K}^{212} coincides with Φ_{BK} computed in the chart for 121 under ‘reversing’ of the variables $t_k = t'_{w_0(k)}, k = 1, \dots, 5$.

As a generalization this remark we get the following

Proposition 3 *For a reduced decomposition $\mathbf{i} \in R(w_0)$ and the cluster chart $\mathcal{S}(\mathbf{i})$, we have*

$$\Phi_{BK}^{\mathbf{i}}(\mathcal{S}(\mathbf{i})) = \Psi_{*K}^{\mathbf{i}^*}((\mathcal{S}(\mathbf{i}^*))^{op}). \tag{15}$$

We establish an explicit crystal bijection between the geometric crystal and * dual geometric crystal below.

5 Piece-Wise Linear Combinatorics and RSK-Correspondences

5.1 Elementary Maps from Which We Make Geometric RSK-Correspondences

For $w \in W$ and a reduced decomposition $\mathbf{i} \in R(w)$, we define the geometric \mathbf{i} -RSK as the composition of $l(w)$ primitive maps, where $l(w)$ is the length of \mathbf{i} . (For simplicity we regard \mathbf{i} as a word of $I^{l(w)}$.)

For any $\mathbf{i} \in I^{l(w)}$ and $k = 1, \dots, l(w)$ we define a primitive map as the rational map $\kappa_k = \kappa_k^{A, \mathbf{i}} : \mathbb{T}^{l(w)} \rightarrow \mathbb{T}^{l(w)}, \mathbb{T}^{l(w)} := (\mathbb{C}^*)^{l(w)}$, by

$$\kappa_k(\mathbf{t})_{k'} = \begin{cases} t_{k'} & \text{if } k' > k \\ \sigma_{0,k}(\mathbf{t}) & \text{if } k' = k \\ t_{k'} \cdot \sigma_{k',k}(\mathbf{t})^{-a_{i_{k'}, i_k}} & \text{if } k' < k, i_{k'} \neq i_k \\ \frac{t_{k'}}{\sigma_{k',k}(\mathbf{t}) \cdot (t_{k'} + \sigma_{k',k}(\mathbf{t}))} & \text{if } k' < k, i_{k'} = i_k \end{cases},$$

where we abbreviated $\sigma_{k',k}(\mathbf{t}) := \sum_{k' < \ell \leq k: i_\ell = i_k} t_\ell$. Clearly, each κ_k is a positive birational isomorphism of $\mathbb{T}^{l(w)}$.

For a word $\mathbf{i} \in R(w)$, the coordinates of $\mathbb{T}^{l(w)}$ are labeled (colored) by simple roots follow to \mathbf{i} ,

$$\begin{matrix} t_1 & t_2 & t_3 & \cdots & t_k & \cdots & t_{l(w)-1} & t_{l(w)} \\ \alpha_{i_1} & \alpha_{i_2} & \alpha_{i_3} & \cdots & \alpha_{i_k} & \cdots & \alpha_{i_{l(w)-1}} & \alpha_{i_{l(w)}} \end{matrix}$$

Suppose, for example, that $s_{i_2} = s_{i_k} = s_{i_{l(w)}}$ and $s_{i_j} \neq s_{i_{l(w)}}$ for other j , than $\kappa_{l(w)}(\mathbf{t})$ is the following map

$$\begin{array}{ccccccc} & & t_1 & & & t_2 & & t_3 & & \cdots \\ & & \downarrow & & & \downarrow & & \downarrow & & \cdots \\ t_1(t_2 + t_k + t_{l(w)})^{-a_{i_1, i_{l(w)}}} & & & & & \frac{t_2}{(t_2 + t_k + t_{l(w)})(t_k + t_{l(w)})} & & t_3(t_k + t_{l(w)})^{-a_{i_3, i_{l(w)}}} & & \cdots \\ \cdot & & t_k & & t_{k+1} & \cdots & & t_{l(w)-1} & & t_{l(w)} \\ \cdot & & \downarrow & & \downarrow & \cdots & & \downarrow & & \downarrow \\ \cdot & & \frac{t_k}{t_{l(w)}(t_k + t_{l(w)})} & & t_{k+1}t_{l(w)}^{-a_{i_{k+1}, i_{l(w)}}} & \cdots & & t_{l(w)-1}t_{l(w)}^{-a_{i_{l(w)-1}, i_{l(w)}}} & & t_2 + t_k + t_{l(w)} \end{array}$$

Definition 1 For a Cartan matrix A , a reduced decomposition \mathbf{i} of $w \in W$, the composition of maps

$$\mathbf{K}_i^A := \kappa_1 \circ \cdots \circ \kappa_{l(w)} \tag{16}$$

is a *geometric i-RSK*.

(This definition is a slight generalization of that introduced in [6].)

Geometric \mathbf{i} -RSK is a positive birational isomorphism of $\mathbb{T}^{l(w)}$ which depends on \mathbf{i} .

Example For SL_3 , and the word 121, we get

$$K_{121}(t_1, t_2, t_3) = \left(\frac{t_1 t_2}{t_1 + t_3}, t_2 t_3, t_1 + t_3 \right).$$

In this example, we have $K_{121} = K_{212}$, but this is because $212 = w_0(1)w_0(2)w_0(1)$.

5.2 Inverse Geometric RSK

The composition of the following maps provide us with the inverse map for RSK.

Let $w \in W$ and $\mathbf{i} \in R(w)$. Then the map $(\kappa_k^{-1})^{A, \mathbf{i}} : (\mathbb{C}^*)^m \rightarrow (\mathbb{C}^*)^m, m := l(w)$, sends the vector (p_1, \dots, p_m) to the vector defined as follows.

Denote by $I(i_k) = \{j \in [k] \mid i_j = i_k\}$, and let $j_1 < j_2 < \dots < j_{|I(i_k)|} = k$ be elements of this set.

Then, for $s \in [j_1 - 1]$,

$$\kappa_k^{-1}(p_s) = p_s p_k^{a_{i_s, i_k}},$$

for $s = j_1$,

$$\kappa_k^{-1}(p_s) = \frac{p_s p_k^2}{p_k p_s + 1},$$

and we redefine $p_k := p_k(1)$ as $p_k(2) := \frac{p_k}{p_k p_s + 1}$;
for $s \in [j_l - 1] \setminus [j_{l-1}]$,

$$\kappa_k^{-1}(p_s) = p_s (p_k(l))^{a_{i_s, i_k}},$$

for $s = j_l, l < |I(i_k)|$,

$$\kappa_k^{-1}(p_s) = \frac{p_s p_k(l)^2}{p_s p_k(l) + 1},$$

and we define

$$p_k(l + 1) := \frac{p_k(l)}{p_k(l) p_{j_l} + 1};$$

and continue as above for the next interval $[j_{l+1}] \setminus [j_l]$;

then, for $s = j_{|I(i_k)|}$, we set

$$\kappa_k^{-1}(p_k) = p_k(|I(i_k)|),$$

and for $s > k, \kappa_k^{-1}(p_s) = p_s$.

We define $\mathbf{K}_\mathbf{i}^{-1} : (\mathbb{C}^*)^m \rightarrow (\mathbb{C}^*)^m$ by the rule

$$\mathbf{K}_\mathbf{i}^{-1}(p_1, \dots, p_m) := \kappa_m^{-1} \circ \dots \circ \kappa_2^{-1} \circ \kappa_1^{-1}(p_1, \dots, p_m).$$

Note that $\kappa_1^{-1}(p_1, \dots, p_m) = (p_1, \dots, p_m)$ is the identical map.

For example, for SL_3 and a reduced word $s_1s_2s_1$, we get

$$\kappa_2^{-1}(p_1, p_2, p_3) = \left(\frac{p_1}{p_2}, p_2, p_3\right),$$

$$\kappa_3^{-1}(q_1, q_2, q_3) = \left(\frac{q_1q_3^2}{q_1q_3 + 1}, q_2\left(\frac{q_3}{q_1q_3 + 1}\right)^{-1}, \frac{q_3}{q_1q_3 + 1}\right),$$

The composition of these maps is

$$\mathbf{K}_{121}^{-1} : (p_1, p_2, p_3) \rightarrow \left(\frac{p_1p_3^2}{p_1p_3 + p_2}, \frac{p_1p_3 + p_2}{p_3}, \frac{p_2p_3}{p_1p_3 + p_2}\right).$$

5.3 Geometric Lusztig Mutations

Piece-wise linear combinatorics of canonical bases was defined by Lusztig [3, 15–17] as tropicalization the following birational mappings between tori $(\mathbb{G}_m)^l$ coordinates of which are colored by corresponding transpositions of a reduced decomposition $\mathbf{i} \in R(w)$, $w \in W$, l is the length of w .

Here we give the rule for simply-laced groups: Positive birational mappings between tori for different reduced decompositions \mathbf{i} and \mathbf{i}' are either swapping coordinates for 2-move, if the decompositions are related by the corresponding 2-move, or

$$(\dots, p, q, r, \dots) \rightarrow \left(\dots, \frac{qr}{p+r}, p+r, \frac{pq}{p+r}, \dots\right)$$

for corresponding 3-move of the decompositions, and is the identical map on the torus T .

5.4 Commutativity Elementary Maps κ_l and Lusztig Moves

Proposition 4 *For any \mathbf{i} , the mapping κ_l and any geometric Lusztig move are commutative.*

Proof The statement is clear for 2-moves. It suffices to check the statement for a 3-move and a mapping κ_l with $i_l \in \{i_s, i_{s+1}, i_{s+2}\}$, where the latter set of indexes corresponds to the triple of the 3-move, and $s + 2 \leq l$.

For $i_l = i_s$, we have

$$\kappa_l(\dots, a, b, c, \dots) = \left(\frac{a}{(t+c)(t+a+c)}, b(t+c), \frac{c}{t(t+c)}\right),$$

where we denote by t the ‘running value’ of t_{i_l} at the $k - (s + 3) + 1$ -step.

Then the composition of the Lusztig 3-move and κ_l is

$$\left(\dots, \frac{bc(t + a + c)}{a + c}, \frac{a + c}{t(t + a + c)}, \frac{abt}{a + c}, \dots\right),$$

and ‘running value’ at $k - s + 1$ steps is $t + a + c$.

On the other hand side we have, the Lusztig map sends

$$\left(\dots, a, b, c, \dots\right) \rightarrow \left(\dots, \frac{bc}{a + c}, a + c, \frac{ab}{a + c}, \dots\right),$$

and

$$\kappa_l\left(\dots, \frac{bc}{a + c}, a + c, \frac{ab}{a + c}, \dots\right) = \left(\dots, \frac{bc(t + a + c)}{a + c}, \frac{a + c}{t(t + a + c)}, \frac{abt}{a + c}, \dots\right),$$

and ‘running value’ at $k - s + 1$ steps is $t + a + c$.

Checking of other possible cases we leave to the reader. □

Remark Let us note that in diagram (1), we can consider an expanded version by replacing the geometric RSK and its inverse by the elementary maps of which they are composed. On this way we will obtain new family of tori and corresponding tropicalizations of corresponding potentials. Explaining of meaning the corresponding potentials and crystal structures will be in done in another paper.

5.5 Lusztig Variety and the Map CA^+

Consider the part of cluster variety, corresponding to seeds labeled by reduced decompositions, $\mathcal{S}(\mathbf{i}), \mathbf{i} \in R(w_0)$.

For a reduced word \mathbf{i} , the mutations of the tuples $CA_{\mathbf{i}}^+(\Delta_{\mathbf{i}})$ of the cluster variables of seeds $\mathcal{S}(\mathbf{i})$ (specialized at the frozen $\Delta_{w_0\omega_i, \omega_i}, i \in I$), $\mathbf{i} \in R(w_0)$, at vertices corresponding to 3-moves follow the Lusztig rule [10]. Recall that, for a reduced decomposition $\mathbf{i} \in R(w_0)$, the Chamber variables are defined as

$$t_k(\mathbf{i}) = \frac{\prod_{l; i_k^- < i_l < i_k} \Delta_{\mathbf{i}|_{\leq l} \omega_{i_l}, \omega_{i_l}}^{-a(i_k, i_l)}}{\Delta_{\mathbf{i}|_{\leq i_k^-} \omega_{i_k}, \omega_{i_k}} \Delta_{\mathbf{i}|_{\leq i_k} \omega_{i_k}, \omega_{i_k}}}, \quad k \in [l(w_0)], \tag{17}$$

plus the frozen variables $\Delta_{w_0\omega_i, \omega_i}, i \in I$.

We denoted that map CA^+ in the diagram (1), specifically, for a cluster seed $\mathcal{S}(\mathbf{i})$, this transformation sends cluster variables $\Delta_{\mathbf{i}}$ to $(t_k(\mathbf{i}))_{k=1, \dots, l(w_0)} =: CA^+(\Delta(\mathbf{i}))$ and leaves unchanged the half of the frozen variables $\Delta_{w_0\omega_i, \omega_i}$, $i \in I$.

Proposition 5 *The tuples $\{t_k(\mathbf{i}), k \in [l(w_0)]\}$, $\mathbf{i} \in R(w_0)$, form the Lusztig variety in the sense of Definition 2.2.1 [4].*

Proof See, for example [10]. □

This Lusztig variety has the following implementation using elementary matrices. The following proposition a cluster version of the Chamber Ansatz of [4].

Proposition 6 *For each $\mathbf{i} \in R(w_0)$, the matrix*

$$x_{-\mathbf{i}}(K_{\mathbf{i}}^A(\{t_k(\mathbf{i}), k \in [l(w_0)]\})) \tag{18}$$

coincides with $M_{\mathbf{i}}$ under change of cluster variables (17).

5.6 Berenstein-Zelevinsky Variety

In [3] it was considered the following positive birational maps for tori $(\mathbb{C}^*)^{l(w_0)}$ labeled by reduced decompositions \mathbf{i} and \mathbf{i}' : swapping coordinates for \mathbf{i} and \mathbf{i}' related by a 2-move and the birational positive transformations of the form

$$(\dots, p, q, r, \dots) \rightarrow (\dots, \frac{q}{p + \frac{q}{r}}, pr, p + \frac{q}{r}, \dots)$$

for that related by the corresponding 3-move.

The *BZ-variety* is the collection of tori labeled by elements of $R(w_0)$ and glued together follows 3-moves by the above BZ-map. For an element $\{p_k^{\mathbf{i}}, k \in [l(w_0)]\}$ of the BZ-variety, the product

$$x_{-i_1}(p_1^{\mathbf{i}}) \cdots x_{-i_{l(w_0)}}(p_{l(w_0)}^{\mathbf{i}})$$

does not depend on the choice of a reduced word $\mathbf{i} \in R(w_0)$, see [3].

Moreover, the BZ-variety endow $G^{w_0, e}$ with a positive structure. This result essentially appears in [2].

5.7 Graded Nakashima-Zelevinsky Cone

Theorem 7

1. For $\mathbf{i} \in R(w_0)$, the tropicalization of Φ_{BK} corresponding to the BZ-torus, labeled by \mathbf{i} , defines $gr \mathcal{N} \mathcal{L}_{\mathbf{i}}$, the graded Nakashima-Zelevinsky cone for \mathbf{i}^2 ;
2. Tropicalization of the BK-potential Φ_{BK} defined by the torus obtained as the composition of geometric RSK $K_{\mathbf{i}}^A$ and the map (17) is $gr \mathcal{L}_{\mathbf{i}}$, the graded Lusztig cone for \mathbf{i} ;
3. The geometric RSK $K_{\mathbf{i}}^A$ sends the Lusztig (geometric) crystal actions $f_{\alpha_k}^c$ defined on the variables (17) to the Berenstein-Kazhdan geometric crystal actions [1] defined on the BZ-variety.

From Theorem 7 follows that the composition of the CA^+ and the geometric RSK-correspondence provides birational maps between the cluster positive structure and the BZ-positive structure for the same geometric crystal on $G^{w_0, e}$.

Before proving we give an example.

Example For SL_3 and a reduced word 121, we have

$$x_{-1} \left(\frac{t_1 t_2}{t_1 + t_3} \right) x_{-2} (t_2 t_3) x_{-1} (t_1 + t_3) = \begin{pmatrix} \frac{1}{t_1 t_2} & 0 & 0 \\ \frac{1}{t_3} & \frac{t_1}{t_3} & 0 \\ 1 & t_1 + t_3 & t_2 t_3 \end{pmatrix}$$

Here $t_1 := \frac{\Delta_{12}}{\Delta_1 \Delta_2}$, $t_2 := \frac{\Delta_2}{\Delta_{12} \Delta_{23}}$, $t_3 := \frac{\Delta_{23}}{\Delta_2 \Delta_3}$, and hence, we have the following form of the above matrix

$$\begin{pmatrix} \Delta_1 \Delta_{23} & 0 & 0 \\ \frac{\Delta_2 \Delta_3}{\Delta_{23}} & \frac{\Delta_3 \Delta_{12}}{\Delta_1 \Delta_{23}} & 0 \\ 1 & \frac{\Delta_{12} \Delta_3 + \Delta_{23} \Delta_1}{\Delta_2 \Delta_1 \Delta_3} & \frac{1}{\Delta_3 \Delta_{12}} \end{pmatrix}$$

Note that de-specialization is obtained by multiplication on the left by the diagonal matrix

$$\begin{pmatrix} \frac{1}{\Delta_{23}} & 0 & 0 \\ 0 & \frac{\Delta_{23}}{\Delta_3} & 0 \\ 0 & 0 & \Delta_3 \end{pmatrix}$$

²The graded Nakashima-Zelevinsky cone for \mathbf{i} is obtained of the realization of highest weight Kashiwara crystal with the highest weight $\sum_{a \in I} c_a \omega_a$ of the form the integer points of a polytope obtained as the intersection of the string cone $\mathcal{S}(\mathbf{i})$ and the polyhedron defined by inequalities: for each $a \in I$ $r\gamma \leq c_a$, while γ runs the set of the crossings $\mathbf{R}_a^l(\mathbf{i})$ wrt the left boundary (see [9]) and $r\gamma$ denotes the corresponding Reineke vector.

that is

$$\begin{pmatrix} \frac{1}{\Delta_{23}} & 0 & 0 \\ 0 & \frac{\Delta_{23}}{\Delta_3} & 0 \\ 0 & 0 & \Delta_3 \end{pmatrix} \cdot \begin{pmatrix} \Delta_1 \Delta_{23} & 0 & 0 \\ \frac{\Delta_2 \Delta_3}{\Delta_{23}} & \frac{\Delta_3 \Delta_{12}}{\Delta_1 \Delta_{23}} & 0 \\ 1 & \frac{\Delta_{12} \Delta_3 + \Delta_{23} \Delta_1}{\Delta_2 \Delta_1 \Delta_3} & \frac{1}{\Delta_3 \Delta_{12}} \end{pmatrix} = \begin{pmatrix} \Delta_1 & 0 & 0 \\ \Delta_2 & \frac{\Delta_{12}}{\Delta_1} & 0 \\ \Delta_3 & \frac{\Delta_{12} \Delta_3 + \Delta_{23} \Delta_1}{\Delta_2 \Delta_1} & \frac{1}{\Delta_{12}} \end{pmatrix}$$

the latter is nothing but the cluster torus M_{121} for the reduced word 121.

Thus, the Berenstein-Kazhdan potential computed in coordinates t_i 's and h_i 's this torus

$$\begin{pmatrix} h_1 \frac{1}{t_1 t_2} & 0 & 0 \\ \frac{h_2}{h_1} \frac{1}{t_3} & \frac{h_2}{h_1} \frac{t_1}{t_3} & 0 \\ \frac{1}{h_2} & \frac{1}{h_2} (t_1 + t_3) & \frac{1}{h_2} t_2 t_3 \end{pmatrix}$$

is

$$\Phi_{BK}(\mathbf{t}, \mathbf{h}) := t_1 + t_3 + t_2 + \frac{h_2^2}{h_1} \frac{1}{t_3} + \frac{h_1^2}{h_2} \frac{t_1 + t_3}{t_1 t_2}.$$

Recall that the potential Φ_{BK} computed at the torus in coordinates p_i 's and h_i 's

$$\alpha_1(h_1)\alpha_2(h_2)x_{-1}(p_1)x_{-2}(p_2)x_{-1}(p_3) = \begin{pmatrix} h_1 \frac{1}{p_1 p_3} & 0 & 0 \\ \frac{h_2}{h_1} (\frac{p_1}{p_2} + \frac{1}{p_3}) & \frac{h_2}{h_1} \frac{p_1 p_3}{p_2} & 0 \\ \frac{1}{h_2} & \frac{1}{h_2} p_3 & \frac{1}{h_2} p_2 \end{pmatrix}$$

is

$$\Phi_{BK}(\mathbf{p}, \mathbf{h}) = p_3 + p_1 + \frac{p_2}{p_3} + \frac{h_2^2}{h_1} (\frac{p_1}{p_2} + \frac{1}{p_3}) + \frac{h_1^2}{h_2} \frac{1}{p_1}.$$

Formal tropicalization of $\Phi_{BK}(\mathbf{t}, \mathbf{h})$ defines the graded Lusztig cone for 121 and that of $\Phi_{BK}(\mathbf{p}, \mathbf{h})$ defines the graded Nakashima-Zelevinsky cone for 121.

5.8 Proof of Theorem 7

We consider SL_n . Items 1 and 2 are slight generalization of the Chamber Ansatz of [3].

Item 3: because of Proposition 4, we can make proof for the lexmin reduced decomposition $\mathbf{i}_{\min} := 1(21)(321) \dots (n-1)n-2 \dots 1$. We have explicit form of geometric crystal actions. Let us check the statement for $f_{\alpha_{n-1}}$. Namely, we have to show that the composition $K_{\mathbf{i}}(f_{\alpha_{n-1}}(c, CA^+(\Delta_{\mathbf{i}})))$ turns into multiplication by c the

coordinate $(n(n-1)/2 - n - 2)$ th coordinate of $K_{\mathbf{i}}(CA^+(\Delta_{\mathbf{i}}))$. The latter coordinate corresponds to s_{n-1} of the lexmin reduced decomposition.

With help of $(n-1)$ 3-moves and $(n-3)(n-3)/2$ 2-moves, we can get from the lexmin decomposition, the following one $\mathbf{i}_{\min}(n-1) := 1(21)(321) \dots (n-3 \ n - 4 \dots 1)(n-1 \ n - 2 \ n - 1 \ n - 3 \ n - 2 \ n - 4 \ n - 3 \dots 12)$ of $I(n-1)$. The corresponding sequence of moves the following point of the Lusztig variety corresponding to \mathbf{t}^{\min} , a tuple of coordinates for the lexmin reduced decomposition. We denote $\mathbf{t} := \mathbf{t}^{\min}$ for simplicity. Denote by $b_s := t_{l(w_0) - (n-1) - (n-2-s)}$, $s = n-2, \dots, 1$, the coordinates of \mathbf{t} which correspond to the segment $(n-2 \ n - 3 \dots 1)$ of \mathbf{i}_{\min} and by $a_s := t_{l(w_0) - (n-1-s)}$, \dots , $s = n-1, \dots, 1$, the coordinates of \mathbf{t} which correspond to the segment $(n-1 \ n - 2 \dots 1)$ of \mathbf{i}_{\min} . Then we have

$$\mathbf{i}_{\min}^{(n-1)} = \begin{cases} t_k, & k < \frac{n(n-1)}{2} - (2n-3) \\ \frac{1}{\frac{1}{a_{n-1}} + \frac{b_{n-2}}{a_{n-2}a_{n-1}} + \frac{b_{n-3}b_{n-2}}{a_{n-3}a_{n-2}a_{n-1}} + \dots + \frac{b_1 \dots b_{n-2}b_{n-1}}{a_1 \dots a_{n-3}a_{n-2}a_{n-1}}}, & k = \frac{(n-2)(n-3)}{2} \\ b_{n-2} + \frac{1}{\frac{b_{n-3}b_{n-2}}{a_{n-3}a_{n-2}a_{n-1}} + \dots + \frac{b_1 \dots b_{n-2}b_{n-1}}{a_1 \dots a_{n-3}a_{n-2}a_{n-1}}}, & k = \frac{(n-2)(n-3)}{2} + 1 \\ b_{n-s} + \frac{1}{\frac{b_1 \dots b_{n-s-1}}{a_{n-3}a_{n-2}a_{n-1}} + \dots + \frac{b_1 \dots b_{n-2}b_{n-1}}{a_1 \dots a_{n-3}a_{n-2}a_{n-1}}}, & k = \frac{(n-2)(n-3)}{2} + s, s \leq n-2 \\ (b_{n-s+1}a_{n-s}(b_{n-s} + \frac{1}{\frac{b_1 \dots b_{n-s-1}}{a_{n-3}a_{n-2}a_{n-1}} + \dots + \frac{b_1 \dots b_{n-2}b_{n-1}}{a_1 \dots a_{n-3}a_{n-2}a_{n-1}}}))^{-1}, & k = \frac{(n-2)(n-1)}{2} + s, s \leq n-1 \end{cases}$$

By definition, $f_{\alpha_{n-1}}$ changes only one coordinate

$$\frac{1}{\frac{1}{a_{n-1}} + \frac{b_{n-2}}{a_{n-2}a_{n-1}} + \frac{b_{n-3}b_{n-2}}{a_{n-3}a_{n-2}a_{n-1}} + \dots + \frac{b_1 \dots b_{n-2}b_{n-1}}{a_1 \dots a_{n-3}a_{n-2}a_{n-1}}}$$

of $\mathbf{t}^{\min(n-1)}$ to

$$\frac{c}{\frac{1}{a_{n-1}} + \frac{b_{n-2}}{a_{n-2}a_{n-1}} + \frac{b_{n-3}b_{n-2}}{a_{n-3}a_{n-2}a_{n-1}} + \dots + \frac{b_1 \dots b_{n-2}b_{n-1}}{a_1 \dots a_{n-3}a_{n-2}a_{n-1}}}.$$

We have

$$\mathbf{i}_{\mathbf{k}}^{\min(n-1)} + \mathbf{i}_{\mathbf{k}+n-2}^{\min(n-1)} = b_s + a_s, \quad k = \frac{n(n-1)}{2} - (2n-3) + s, \quad s \leq n-1, \quad b_{n-1} := 0.$$

Because of this property and since each elementary κ_l has the same conservation law, we get the statement. □

6 Inverse Geometric RSK and * Dual Geometric Crystals

One can see that elementary maps of which the reverse geometric RSK is composed are also commute with transformations of the Lusztig variety.

6.1 Lusztig Variety and the Map CA^-

For each seed $\mathcal{S}(\mathbf{i})$, with $\mathbf{i} \in R(w_0)$, we make the following change of cluster variables wrt the specialization of the frozen variables $\Delta_{\omega_i, \omega_i}$, denoted by $CA^-(\Delta(\mathbf{i}))$, that is inverting the map $grNA_{\mathbf{i}}$ of [10], and defined by

$$q_{l(w_0)-l+1} := \frac{\Delta_{\mathbf{i}|_{\leq l} \omega_i, \omega_i}}{\Delta_{\mathbf{i}|_{< l} \omega_i, \omega_i}}, l \in [l(w_0)]. \tag{19}$$

Then for a reduced word $\mathbf{i}(k)$ which is tail optimal for k , we get that the *-dual $f_k * (c, \cdot)$ action changes only one variable q_1 , sending it to $c \cdot q_1$.

Note, that in coordinates q_l 's, the cluster transformations between seeds labeled by reduced decompositions are nothing else but the inverse BZ-moves for 3-braid moves between the corresponding words. Inverse means the following

$$(\dots, p, q, r, \dots) \rightarrow (\dots, p + \frac{q}{r}, pr, \frac{q}{p + \frac{q}{r}}, \dots).$$

For a reduced word $\mathbf{i} \in R(w_0)$, we have the corresponding variant of the Chamber Ansatz

Proposition 8 *The factorization*

$$y_{\mathbf{i}}(K_{\mathbf{i}}^{-1}(\mathbf{q})) \tag{20}$$

defines the cluster torus factorization $M_{\mathbf{i}}$ written in coordinates (19).

Here is an example.

Example Consider SL_3 and reduced word 121. Then the reverse RSK sends

$$(q_1, q_2, q_3) \rightarrow \left(\frac{q_1 q_3^2}{q_1 q_3 + q_2}, \frac{q_1 q_3 + q_2}{q_3}, \frac{q_2 q_3}{q_1 q_3 + q_2} \right),$$

and

$$y_1\left(\frac{q_1 q_3^2}{q_1 q_3 + q_2}\right) y_2\left(\frac{q_1 q_3 + q_2}{q_3}\right) y_1\left(\frac{q_2 q_3}{q_1 q_3 + q_2}\right) = \begin{pmatrix} 1 & 0 & 0 \\ q_3 & 1 & 0 \\ q_1 q_3 & \frac{q_1 q_3 + q_2}{q_3} & 1 \end{pmatrix}.$$

Recalling that $q_3 := \frac{\Delta_2}{\Delta_1}$, $q_2 := \frac{\Delta_{23}}{\Delta_{12}}$, $q_1 := \frac{\Delta_3}{\Delta_2}$, we get the latter matrix as

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{\Delta_2}{\Delta_1} & 1 & 0 \\ \frac{\Delta_3}{\Delta_1} & \frac{\Delta_3 \Delta_{12} + \Delta_1 \Delta_{23}}{\Delta_2 \Delta_{12}} & 1 \end{pmatrix}.$$

Multiplying the latter matrix on the right by diagonal matrix $\begin{pmatrix} \Delta_1 & 0 & 0 \\ 0 & \frac{\Delta_{12}}{\Delta_1} & 0 \\ 0 & 0 & \frac{1}{\Delta_{12}} \end{pmatrix}$ we get

$$\begin{pmatrix} \Delta_1 & 0 & 0 \\ \Delta_2 & \frac{\Delta_{12}}{\Delta_1} & 0 \\ \Delta_3 & \frac{\Delta_3 \Delta_{12} + \Delta_1 \Delta_{23}}{\Delta_1 \Delta_2} & \frac{1}{\Delta_{12}} \end{pmatrix},$$

that is M_{121} . □

6.2 Lusztig Variety and Decoration Ψ_{*K}

For SL_n , we have

$$\Psi_{*K}(y_i(\mathbf{q})\alpha_1(h_1) \cdots \alpha_{n-1}(h_{n-1})) = \sum_{i \in I(w_0)} q_i + \sum_{k \in I} \frac{h_k^2}{h_{k-1}h_{k+1}} \sum_{\gamma \in \mathbf{R}_k^l(\mathbf{i})} \mathbf{q}^{-s\gamma}, \tag{21}$$

where $\mathbf{R}_k^l(\mathbf{i})$ denotes the set of crossings wrt the left boundary [9], and $s\gamma$ is the Reineke statistics.

Thus the tropicalization of this potential defines the $*$ -Kashiwara dual Lusztig cone.

In such a case, we have

Theorem 9 For any $\mathbf{i} \in R(w_0)$,

$$\begin{aligned} & \Psi_{*K}(y_i(\mathbf{K}_i^{-1}(\mathbf{q}))\alpha_1(h_1) \cdots \alpha_{n-1}(h_{n-1})) = \\ & \sum_{k \in I} \sum_{\gamma \in \mathbf{R}_k^l(\mathbf{i})} \mathbf{q}^{r\gamma} + \sum_{k \in I} \frac{h_k^2}{h_{k-1}h_{k+1}} \sum_{s \in I(k)} \mathbf{q}^{x_{i_k} + 2 \sum_{m \geq k} x_{i_m} - \sum_{(i \pm 1)_l \geq i_k} x_{(i \pm 1)_l}} \end{aligned}$$

Note that the tropicalization of the latter potential defines the Littelmann graded cone $gr\mathcal{S}_i$.

We also have the following theorem

Theorem 10

1. For $\mathbf{i} \in R(w_0)$, tropicalization of the Ψ_{*K} corresponding to the torus of the Lusztig variety labeled by \mathbf{i} , defines $*$ -Kashiwara dual graded Lusztig cone, $gr\mathcal{L}_{\mathbf{i}}^*$.
2. The inverse geometric RSK $K_{\mathbf{i}}^{-1}$ sends the dual Kashiwara (geometric) crystal action $f_{\alpha_k}^c$, defined on the variables (19), to the geometric $*$ -dual Lusztig crystal action on the Lusztig variety.

6.3 Transposed BZ-Twist

Berenstein and Zelevinsky [3, Definition 4.1]) defined twist map between reduced double Bruhat cells. We use this map for $G^{w_0, e}$, and in such a case it is

$$\eta_{w_0, e} : N \cap B_- w_0 B \rightarrow B \cap N_- w_0 N_-, \quad \eta_{w_0, e}(x) = [(x^t)^{-1}]_+ ([\bar{w}^{-1}x]_+)^t,$$

where $x \rightarrow x^t$ is the involutive antiautomorphism of G given by

$$\mathbf{h}^t = \mathbf{h}^{-1}, \mathbf{h} \in H, x_i(t)^t = x_i(t), y_i(t)^t = y_i(t).$$

By transposing the BZ-twist we get a map

$$N_- \cap B w_0 B \rightarrow B \cap N_- w_0 N_-$$

which is a crystal isomorphism between the $*$ -dual geometric crystals for the Lusztig-variety and geometric crystal for the BZ-variety. This result is a reformulation of Theorem 5.10 of [3].

Thus all the maps are in place in the diagrams (1) and (2) in order to define the Donaldson-Thomas transformation through commutativity.

Note that all these maps are crystal isomorphism between corresponding cluster geometric crystals.

Thus we have as a corollary

Theorem 11 *The Donaldson-Thomas transformation defined above is an isomorphism of geometric crystals.*

In particular, the tropical DT-transformation is a crystal isomorphism between the Littelmann graded cone $gr\mathcal{S}_{\mathbf{i}}$ and the Lusztig graded cone $gr\mathcal{L}_{\mathbf{i}}$.

Acknowledgements I thank Arkady Berenstein, Volker Genz and Bea Schumann for inspired and fruitful discussions, organizers of the MATRIX workshop, and especially Paul Zinn-Justin, and the RSF grant 16-11-10075 for financial support.

References

1. Berenstein, A., Kazhdan, D.: Geometric and unipotent crystals II: from unipotent bicrystals to crystal bases. *Contemporary Mathematics*, vol. 433, pp. 13–88. American Mathematical Society, Providence (2007)
2. Berenstein, A., Zelevinsky, A.: Total positivity in Schubert varieties. *Comment. Math. Helv.* **72**(1), 128–166 (1997)
3. Berenstein, A., Zelevinsky, A.: Tensor product multiplicities, canonical bases and totally positive varieties. *Invent. Math.* **143**, 77–128 (2001)
4. Berenstein, A., Fomin, S., Zelevinsky, A.: Parametrizations of canonical bases and totally positive matrices (with S. Fomin and A. Zelevinsky). *Adv. Math.* **122**, 49–149 (1996)
5. Berenstein, A., Fomin, S., Zelevinsky, A.: Cluster algebras III: upper bounds and double Bruhat cells (2004). [arXiv:math/0305434v3](https://arxiv.org/abs/math/0305434v3)
6. Berenstein, A., Kirillov, A., Koshevoy, G.: Generalized RSK correspondences, *Obervolfach reports*, 23/2015, 1303–1305
7. Danilov, V., Karzanov, A., Koshevoy, G.: Tropical Plücker functions and Kashiwara crystals. *Contemporary Mathematics*, vol. 616, pp. 77–100. American Mathematical Society, Providence (2014)
8. Fomin, S., Zelevinsky, A.: Cluster algebras I: foundations. *J. Am. Math. Soc.* **15**, 497–529 (2002)
9. Genz, V., Koshevoy, G., Schumann, B.: Combinatorics of canonical bases revisited: type A (2017). [arXiv:1611.03465](https://arxiv.org/abs/1611.03465)
10. Genz, V., Koshevoy, G., Schumann, B.: Polyhedral parametrizations of canonical bases & cluster duality (2017). [arXiv:1711.07176](https://arxiv.org/abs/1711.07176)
11. Goncharov, A., Shen, L.: Donaldson-Thomas transformations of moduli spaces of G-local systems (2016). [arXiv:1602.06479](https://arxiv.org/abs/1602.06479)
12. Gross, M., Hacking, P., Keel, S., Kontsevich, M.: Canonical bases for cluster algebras, preprint (2014). [arXiv:1411.1394v2](https://arxiv.org/abs/1411.1394v2) [math.AG]
13. Kashiwara, M.: On crystal bases. In: *Representations of Groups (Banff, AB, 1994)*. CMS Conference Proceedings, vol. 16, pp. 155–197. American Mathematical Society, Providence (1995)
14. Lee, K., Schiffler, R.: Positivity for cluster algebras. *Ann. Math.* **182**, 73–125 (2015)
15. Lusztig, G.: *Introduction to Quantum Groups*. Progress in Mathematics, vol. 110. Birkhäuser, Boston (1993)
16. Lusztig, G.: Total positivity in reductive groups. In: *Lie Theory and Geometry*. Progress in Mathematics, vol. 123, pp. 531–568. Birkhäuser, Boston (1994)
17. Lusztig, G.: Piecewise linear parametrization of canonical bases (2008). [arXiv:0807.2824](https://arxiv.org/abs/0807.2824)
18. Nakashima, T.: Decorations on geometric crystals and monomial realizations of crystal bases for classical groups. *J. Algebra* **399**, 712–769 (2014)
19. Nakashima, T., Zelevinsky, A.V.: Polyhedral realizations of crystal bases for quantized Kac-Moody algebras. *Adv. Math.* **131**, 253–278 (1997)
20. Weng, D.: Donaldson-Thomas transformation of double Bruhat cells in semisimple lie groups (2016). [arXiv:1611.04186](https://arxiv.org/abs/1611.04186)

Part II
Other Contributed Articles

Fields of Definition of Finite Hypergeometric Functions



Frits Beukers

Abstract Finite hypergeometric functions are functions of a finite field \mathbb{F}_q to \mathbb{C} . They arise as Fourier expansions of certain twisted exponential sums and were introduced independently by John Greene and Nick Katz in the 1980s. They have many properties in common with their analytic counterparts, the hypergeometric functions. One restriction in the definition of finite hypergeometric functions is that the hypergeometric parameters must be rational numbers whose denominators divide $q - 1$. In this note we use the symmetry in the hypergeometric parameters and an extension of the exponential sums to circumvent this problem as much as possible.

1 Introduction

In the 1980s Greene [4] and Katz [5] independently introduced functions from finite fields to the complex numbers which can be interpreted as finite sum analogues of the classical one variable hypergeometric functions. These functions, also known as Clausen–Thomae functions, are determined by two multisets of d entries in \mathbb{Q} each. We denote them by $\alpha = (\alpha_1, \dots, \alpha_d)$ and $\beta = (\beta_1, \dots, \beta_d)$. Throughout we assume that these sets have empty intersection when considered modulo \mathbb{Z} . The Clausen–Thomae functions satisfy a linear differential equation of order d with rational function coefficients. See [1].

Let \mathbb{F}_q be the finite field with q elements. Let ζ_p be a primitive p -th root of unity and define the additive character $\psi_q(x) = \zeta_p^{\text{Tr}(x)}$ where Tr is the trace from \mathbb{F}_q to \mathbb{F}_p . For any multiplicative character $\chi : \mathbb{F}_q^\times \rightarrow \mathbb{C}^\times$ we define the Gauss sum

$$g(\chi) = \sum_{x \in \mathbb{F}_q^\times} \chi(x) \psi_q(x).$$

F. Beukers (✉)
University of Utrecht, Utrecht, The Netherlands
e-mail: f.beukers@uu.nl

Let ω be a generator of the character group on \mathbb{F}_q^\times . We use the notation $g(m) = g(\omega^m)$ for any $m \in \mathbb{Z}$. Note that $g(m)$ is periodic in m with period $q - 1$. Note that the dependence of $g(m)$ on ζ_p and ω is not made explicit. Very often we shall need characters on \mathbb{F}_q^\times of a given order. For that we use the notation $\mathfrak{q} = q - 1$ so that a character of order d can be given by $\omega^{\mathfrak{q}/d}$ for example, provided that d divides \mathfrak{q} of course.

Now we define finite hypergeometric sums. Let again α and β be multisets of d rational numbers each, and disjoint modulo \mathbb{Z} . We need the following crucial assumption.

Assumption 1.1 *Suppose that*

$$(q - 1)\alpha_i, (q - 1)\beta_j \in \mathbb{Z}$$

for all i and j .

Definition 1.2 (Finite Hypergeometric Sum) Keep the above notation and Assumption 1.1. We define for any $t \in \mathbb{F}_q$,

$$H_q(\alpha, \beta|t) = \frac{1}{1 - q} \sum_{m=0}^{q-2} \prod_{i=1}^d \left(\frac{g(m + \alpha_i \mathfrak{q})g(-m - \beta_i \mathfrak{q})}{g(\alpha_i \mathfrak{q})g(-\beta_i \mathfrak{q})} \right) \omega((-1)^d t)^m .$$

It is an exercise to show that the values of $H_q(\alpha, \beta|t)$ are independent of the choice of ζ_p .

The hypergeometric sums above were considered without the normalizing factor

$$\left(\prod_{i=1}^d g(\alpha_i \mathfrak{q})g(-\beta_i \mathfrak{q}) \right)^{-1}$$

by Katz in [5, p. 258]. Greene, in [4], has a definition involving Jacobi sums which, after some elaboration, amounts to

$$\omega(-1)^{|\beta| \mathfrak{q}} q^{-d} \prod_{i=1}^d \frac{g(\alpha_i \mathfrak{q})g(-\beta_i \mathfrak{q})}{g(\alpha_i \mathfrak{q} - \beta_i \mathfrak{q})} H_q(\alpha, \beta|t) ,$$

where $|\beta| = \beta_1 + \dots + \beta_d$. The normalization we adopt in this paper coincides with that of McCarthy, [6, Def 3.2].

Let

$$A(x) = \prod_{j=1}^d (x - e^{2\pi i \alpha_j}), \quad B(x) = \prod_{j=1}^d (x - e^{2\pi i \beta_j}).$$

An important special case is when $A(x), B(x) \in \mathbb{Z}[x]$. In that case we say that the hypergeometric sum is defined over \mathbb{Q} . Another way of describing this case is that $k\alpha \equiv \alpha \pmod{\mathbb{Z}}$ and $k\beta \equiv \beta \pmod{\mathbb{Z}}$ for all integers k relatively prime to the common denominator of the α_i, β_j . In other words, multiplication by k of the $\alpha_i \pmod{\mathbb{Z}}$ simply permutes these elements. Similarly for the β_j . From work of Levelt [1, Thm 3.5] it follows that in such a case the monodromy group of the classical hypergeometric equation can be defined over \mathbb{Z} . It also turns out that hypergeometric sums defined over \mathbb{Q} occur in point counts in \mathbb{F}_q of certain algebraic varieties, see [2, Thm 1.5] and the references therein. It is an easy exercise to show that $H_q(\alpha, \beta|t)$ is independent of the choice of ω (it is already independent of the choice of ψ_q).

One of the obstacles in the definition of finite hypergeometric sums over \mathbb{Q} is Assumption 1.1 which has to be made on q , whereas one has the impression that such sums can be defined for any q relatively prime with the common denominator of the α_i, β_j . This is resolved in [2, Thm 1.3] by an extension of the definition of hypergeometric sum. The idea is to apply the theorem of Hasse–Davenport to the products of Gauss sums which occur in the coefficients of the hypergeometric sum. Another way of dealing with this problem is given by McCarthy, who uses the Gross–Koblitz theorem which expresses Gauss sums as values of the p -adic Γ -function.

Theorem 1.3 (Gross–Koblitz) *Let ω be the inverse of the Teichmüller character. Let $\pi^{p-1} = -p$ and ζ_p such that $\zeta_p \equiv 1 + \pi \pmod{\pi^2}$. Let Γ_p be the p -adic Morita Γ -function. Let $q = p^f$ and $g_q(m)$ denote the Gauss-sum over \mathbb{F}_q with multiplicative character ω^m . Then, for any integer m we have*

$$g_q(m) = - \prod_{i=0}^{f-1} \pi^{(p-1)\left\{\frac{p^i m}{q-1}\right\}} \Gamma_p \left(\left\{ \frac{p^i m}{q-1} \right\} \right).$$

Here $\{x\} = x - [x]$ is the fractional part of x . In particular, when $q = p$ we get

$$g_p(m) = -\pi^{(p-1)\left\{\frac{m}{p-1}\right\}} \Gamma_p \left(\left\{ \frac{m}{p-1} \right\} \right).$$

See Henri Cohen’s book [3] for a proof. When p does not divide the common denominator of the α_i, β_j one easily writes down a p -adic version of our hypergeometric sum for the case $q = p$.

Definition 1.4 We define $G_p(\alpha, \beta|t)$ by the sum

$$\frac{1}{1-p} \sum_{m=0}^{q-2} \omega((-1)^d t)^m (-p)^{\Lambda(m)} \prod_{i=1}^d \frac{\Gamma_p \left(\left\{ \alpha_i + \frac{m}{p-1} \right\} \right)}{\Gamma_p(\{\alpha_i\})} \frac{\Gamma_p \left(\left\{ -\beta_i - \frac{m}{p-1} \right\} \right)}{\Gamma_p(\{-\beta_i\})},$$

where

$$\Lambda(m) = \sum_{i=1}^d \left\{ \alpha_i + \frac{m}{p-1} \right\} - \{\alpha_i\} + \left\{ -\beta_i - \frac{m}{p-1} \right\} - \{-\beta_i\}.$$

Note that

$$\Lambda(m) = \sum_{i=1}^d - \left\lfloor \alpha_i + \frac{m}{p-1} \right\rfloor + \lfloor \alpha_i \rfloor - \left\lfloor -\beta_i - \frac{m}{p-1} \right\rfloor + \lfloor -\beta_i \rfloor.$$

In particular $\Lambda(m) \in \mathbb{Z}$. Definition 1.4 almost coincides with McCarthy’s function ${}_dG_d$ from [6, Def 1.1] in the sense that our function coincides with ${}_dG_d(1/t)$. We prefer to adhere to the definition given above. The advantage of Definition 1.4 is that Assumption 1.1 is not required, it is well-defined for all parameters α_i, β_j as long as they are p -adic integers. Define

$$\delta = \delta(\alpha, \beta) = \max_{x \in [0,1]} \sum_{i=1}^d \lfloor x + \alpha_i \rfloor - \lfloor \alpha_i \rfloor + \lfloor -x - \beta_i \rfloor - \lfloor -\beta_i \rfloor.$$

Then, using Definition 1.4 and the fact that $-\Lambda(m) \leq \delta$ one easily deduces that $p^\delta G_p(\alpha, \beta|t)$ is a p -adic integer. In [6, Prop 3.1] we find this in a slightly different formulation. However, it is not clear from the definition whether this value is algebraic or not over \mathbb{Q} . It is the purpose of the present note to be a bit more specific by proving the following theorem.

Theorem 1.5 *Let notations be as above and let K be the field extension of \mathbb{Q} generated by the coefficients of $A(x)$ and $B(x)$. Suppose p splits in K , i.e. p factors into $[K : \mathbb{Q}]$ distinct prime ideals in the integers of K . Let $\Delta = \max_k \delta(k\alpha, k\beta)$ over all integers k relatively prime with the common denominator of the α_i, β_j . Then $p^\Delta G_p(\alpha, \beta|t)$ is an algebraic integer in K .*

For the proof we construct in Sect. 2 a generalization of the hypergeometric function $H_q(A, B|t)$ involving two semisimple finite algebras A and B over \mathbb{F}_q . We show that it belongs to K and then, in Sect. 3 identify its p -adic evaluation with $G_p(\alpha, \beta|t)$.

2 Gauss Sums on Finite Algebras

The main idea of the proof of Theorem 1.5 is to use Gauss sums on finite commutative algebras over \mathbb{F}_q with 1. Let A be such an algebra. For any $x \in A$ we define the trace $\text{Tr}(x)$ and norm $N(x)$ as the trace and norm of the \mathbb{F}_p -linear map given by multiplication with x on A .

Choose an additive character ψ on A which is *primitive*. That is, to any ideal $I \subset A, I \neq (0)$ there exists $x \in I$ such that $\psi(x) \neq 1$. Any other non-degenerate additive character is of the form $\psi(ax)$ with $a \in A^\times$. A multiplicative character χ is called *primitive* if its kernel does not contain any subgroup of the form $\{1+a|a \in I\}$ for some non-zero ideal I in A .

For any multiplicative character χ on A^\times we can define a Gauss sum

$$g_A(\psi, \chi) = \sum_{x \in A^\times} \psi(x)\chi(x).$$

When A is not semisimple, the Gauss sum can be 0, as illustrated by the following example.

Example 2.1 Let $A = \mathbb{F}_p[x]/(x^2)$. Choose the additive character $\psi(a + bx) = \zeta_p^b$. It is easy to see that this is a primitive character. Note that $a + bx \in A^\times \iff a \in \mathbb{F}_p^\times$. Let χ be a nontrivial multiplicative character on \mathbb{F}_p^\times and extend it to A^\times by $\chi(a + bx) = \chi(a)$. Then

$$g_A(\psi, \chi) = \sum_{a \in \mathbb{F}_p^\times, b \in \mathbb{F}_p} \zeta_p^b \chi(a) = 0.$$

◇

So we restrict ourselves to semisimple algebras. These are precisely the finite sums of finite field extensions of \mathbb{F}_q . In this case there is an obvious choice for the additive character.

Lemma 2.2 *Suppose A is a direct sum of finite field extensions of \mathbb{F}_q . Then $\psi(x) = \zeta_p^{\text{Tr}(x)}$ is a primitive additive character.*

Proof Let $A \cong \bigoplus_{i=1}^r F_i$ with F_i a finite field extension of \mathbb{F}_q for all i . Then $\psi(x) = \zeta_p^{\text{Tr}_1(x_1) + \dots + \text{Tr}_r(x_r)}$, where Tr_i stands for the trace function on F_i . If ψ were not primitive then there exists $a \in A, a \neq 0$ such that $\psi(ax) = 1$ for all $x \in A$. Suppose $a = (a_1, \dots, a_r)$ and assume, without loss of generality, $a_1 \neq 0$. Then $\psi(x, 0, \dots, 0) = \zeta_p^{\text{Tr}(a_1 x)} = 1$ for all $x \in F_1$. By the properties of the trace of a field this is not possible. □

From now on we use the trace character on a semisimple algebra A as additive character and write $g_A(\chi)$ for the Gauss sum. So we dropped the dependence of the Gauss sum on the additive character. The only amount of freedom in the additive character rests on the choice of ζ_p .

Proposition 2.3 *Let A be a direct sum of finite fields over \mathbb{F}_q and $\psi(x) = \zeta_p^{\text{Tr}(x)}$ the additive character. Let χ a multiplicative character. Then there exists a non-negative integer f such that*

$$|g_A(\chi)|^2 = q^f.$$

Proof Again, write $A = \bigoplus_{i=1}^r F_i$. Then χ can be written as $\chi(x_1, \dots, x_r) = \chi_1(x_1) \cdots \chi_r(x_r)$, where χ_i is a multiplicative character on F_i^\times . This implies that

$$g_A(\psi, \chi) = \prod_{i=1}^r g(\chi_i),$$

where $g(\chi_i)$ is the usual Gauss sum on the field F_i . The additive character on F_i is $\zeta_p^{\text{Tr}_i(x)}$ with the same choice of ζ_p for each i . Our assertion follows directly. \square

Choose two finite semisimple algebras A, B over \mathbb{F}_q . Choose the trace characters on each of them with the same choice of ζ_p and call them ψ_A, ψ_B . Let χ_A, χ_B be multiplicative characters on A^\times, B^\times . Denote the norms on A, B by N_A, N_B .

Definition 2.4 We define

$$H_q(A, B|t) = \frac{-1}{g_A(\chi_A)g_B(\chi_B)} \sum_{x \in A^\times, y \in B^\times, tN_A(x) = N_B(y)} \psi_A(x)\psi_B(-y)\chi_A(x)\overline{\chi_B(y)},$$

for any $t \in \mathbb{F}_q^\times$.

The following theorem gives its Fourier expansion in t .

Theorem 2.5 *Let ω be a generator of the multiplicative characters on \mathbb{F}_q^\times . When the context is clear we denote both functions $\omega(N_A(x))$ and $\omega(N_B(y))$ by ω_N . We then have,*

$$H_q(A, B|t) = \frac{1}{1-q} \sum_{m=0}^{q-2} \frac{g_A(\chi_A \omega_N^m)g_B(\overline{\chi_B} \omega_N^{-m})}{g_A(\chi_A)g_B(\chi_B)} \omega(N_B(-1)t)^m.$$

Proof We compute the Fourier expansion $\sum_{m=0}^{q-2} c_m \omega(t)^m$ of $H_q(A, B|t)$. The coefficient c_m can be computed using

$$c_m = \frac{1}{q-1} \sum_{t \in \mathbb{F}_q^\times} H_q(A, B|t) \omega(t)^{-m}.$$

When we substitute the definition for $H_q(A, B|t)$ in the summation over t , we get a summation over $t \in \mathbb{F}_q^\times, x \in A^\times, y \in B^\times$ with the restriction $tN_A(x) = N_B(y)$. So we might as well substitute $t = N_B(y)/N_A(x)$ and sum over x, y . We get,

$$c_m = \frac{1}{1-q} \sum_{x \in A^\times, y \in B^\times} \frac{1}{g_A g_B} \psi_A(x)\psi_B(-y)\chi_A(x)\chi_B(y)^{-1} \omega(N_A(x))^m \omega(N_B(y))^{-m}.$$

The summation over x yields $g_A(\chi_A \omega_N^m)$. To sum over y we first replace y by $-y$ and then perform the summation. We get $\omega(N_B(-1))^m g_B(\overline{\chi_B} \omega_N^{-m})$. This proves our theorem. \square

Example 2.6 As in the previous section take two multisets of hypergeometric parameters α, β . Suppose that $(q - 1)\alpha_i, (q - 1)\beta_j$ are in \mathbb{Z} for all i, j . Take $A = B = \mathbb{F}_q^d$, the direct sum of d copies of \mathbb{F}_q with componentwise addition and multiplication. The norm on A, B is given by $N(x_1, \dots, x_d) = x_1 \cdots x_d$. In particular $N_B(-1) = (-1)^d$. For both A, B we take the additive character $\psi(x_1, \dots, x_d) = \zeta_p^{\text{Tr}(x_1 + \cdots + x_d)}$, where Tr the trace function on \mathbb{F}_q . As multiplicative characters we take

$$\chi_A(x_1, \dots, x_d) = \prod_{i=1}^d \omega(x_i)^{(q-1)\alpha_i}, \quad \chi_B(x_1, \dots, x_d) = \prod_{j=1}^d \omega(y_j)^{(q-1)\beta_j}.$$

An easy calculation shows that $g_A(\chi_A \omega_N^m) = \prod_{i=1}^d g(m + (q - 1)\alpha_i)$ and similarly for g_B . So we see that we recover the finite hypergeometric sum of the previous section. \diamond

Lemma 2.7 *Suppose $\dim_{\mathbb{F}_q}(A) = \dim_{\mathbb{F}_q}(B)$. Then $H_q(A, B|t)$ does not depend on the choice of ζ_p in the additive characters.*

As a corollary, in this equi-dimensional case the values of $H_q(A, B|t)$ are contained in the field generated by the character values of χ_A, χ_B .

Proof When we choose $\zeta_p^a, a \in \mathbb{F}_p^\times$ instead of ζ_p in the definition of the additive character it is easy to check that $g_A(\chi_A)$ gets replaced by $\chi_A(a)^{-1} g_A(\chi_A)$. And similarly for B . As a corollary any term in the sum in the hypergeometric sum in Theorem 2.4 is multiplied by $\omega(N_B(a)/N_A(a))^m$. Since $a \in \mathbb{F}_p$ is a scalar, $N_A(a) = N_B(a) = a^d$, where $d = \dim_{\mathbb{F}_q}(A) = \dim_{\mathbb{F}_q}(B)$. Hence, in the case of equal dimensions of A, B the multiplication factor is 1.

Let $\sigma \in \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ be such that it fixes the values of χ_A, χ_B but sends ζ_p to ζ_p^a . According to the above calculation $H_q(A, B|t)$ is fixed under this substitution and hence under σ . \square

Let us return momentarily to Example 2.6 and suppose that the parameters α have the property that $k\alpha \equiv \alpha \pmod{\mathbb{Z}}, k\beta \equiv \beta \pmod{\mathbb{Z}}$ for all k relative prime with the common denominator of the α_i, β_j . Then, for any $\sigma \in \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ there exists a permutation ρ of the summands of $A = \bigoplus_{i=1}^d \mathbb{F}_p$ such that $\chi_A(\rho(x)) = \chi_A(x)^\sigma$ for all $x \in A^\times$. A similar permutation exists for B . Notice also that $\text{Tr}(\rho(x)) = \text{Tr}(x)$ and $N(\rho(x)) = N(x)$.

A similar situation arises in the case $A = \mathbb{F}_{p^r}$ as \mathbb{F}_p -algebra. Let χ_A be a character of order d dividing $p^r - 1$. Let ρ be the p -th power Frobenius on A , then $\chi_A(\rho(x)) = \chi_A(x)^p$, a conjugate of $\chi_A(x)$ for all $x \in A^\times$. Notice also that $\text{Tr}(\rho(x)) = x$ and $N(\rho(x)) = N(x)$.

Definition 2.8 Let A be a finite dimensional \mathbb{F}_q -algebra. A ring automorphism $\rho : A \rightarrow A$ is called an \mathbb{F}_q -automorphism if it is \mathbb{F}_q -linear and it fixes both norm and trace of A .

Proposition 2.9 Let A, B be finite commutative semisimple \mathbb{F}_q -algebras. Let χ_A, χ_B be multiplicative characters. Consider the subgroup G of $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ of elements σ for which there exists an \mathbb{F}_q -automorphisms ρ_A of A and ρ_B of B with the property that $\chi_A(\rho_A(x)) = \chi_A(x)^\sigma$ and $\chi_B(\rho_B(x)) = \chi_B(x)^\sigma$ for every $\sigma \in G$. Then $H_q(A, B|t)$ lies in the fixed field of G for every $t \in \mathbb{F}_q^\times$.

Proof Let $\sigma \in G$. We first compute the action of σ on $g_A(\chi_A)$. Suppose that $\sigma(\zeta_p) = \zeta_p^a$.

$$\begin{aligned} g_A(\chi_A)^\sigma &= \sum_{x \in A^\times} \zeta_p^{a\text{Tr}(x)} \chi_A(x)^\sigma \\ &= \sum_{x \in A^\times} \zeta_p^{a\text{Tr}(x)} \chi_A(\rho(x)) \\ &= \sum_{x \in A^\times} \zeta_p^{\text{Tr}(\rho^{-1}(x))} \chi_A(a^{-1}x) \\ &= \chi_A(a)^{-1} g_A(\chi_A) \end{aligned}$$

A similar calculation holds for B . Now apply σ to the terms in the sum in Definition 2.4. A similar calculation as above shows that the sum gets multiplied with $\chi_A(a)^{-1} \chi_B(a)^{-1}$. This cancels the factor coming from $g_A(\chi_A)g_B(\chi_B)$. Hence $H_q(A, B|t)$ is fixed under all $\sigma \in G$. □

3 Proof of Theorem 1.5

We use the notations from the introduction. In particular

$$A(x) = \prod_{j=1}^d (x - e^{2\pi i \alpha_j}), \quad B(x) = \prod_{j=1}^d (x - e^{2\pi i \beta_j})$$

and K is the field generated by the coefficients of $A(x)$ and $B(x)$. Let p be a prime which splits completely in K . Then we can consider $A(x)$ as element of $\mathbb{F}_p[x]$. Let $A(x) = A_1(x) \cdots A_r(x)$ be the irreducible factorization of $A(x)$ in $\mathbb{F}_p[x]$. For the \mathbb{F}_p -algebra we take $\bigoplus_{i=1}^r \mathbb{F}_p[x]/(A_i(x))$. The construction of a multiplicative character on A is as follows. First we choose a multiplicative character ω on $\overline{\mathbb{F}_p}$ such that its restriction to \mathbb{F}_{p^r} has order $p^r - 1$ for all $r \geq 1$ and fix in the remainder of the proof.

Since p splits in K multiplication by p gives a permutation of the multiset α modulo \mathbb{Z} . Under this action $\alpha \pmod{\mathbb{Z}}$ decomposes into a union of orbits, which we call p -orbits. Let O be such a p -orbit. Then $\prod_{\alpha \in O} (x - e^{2\pi i \alpha})$ is a polynomial and p splits in the field generated by its coefficients. So we can consider it modulo a prime ideal dividing p and hence as an element of $\mathbb{F}_p[x]$. It is one of the factors $A_i(x)$ of the mod p factorization of $A(x)$. The orbit O will now be denoted by O_i . There are r orbits and we renumber the indices of the α_i such that $\alpha_i \in O_i$ for $i = 1, \dots, r$. On $\mathbb{F}_p[x]/(A_i)$ we define the multiplicative character $\chi_i = \omega^{\alpha_i(q_i-1)}$, where $q_i = p^{\deg(A_i)}$. If we would have chosen $p\alpha_i$ instead of α_i , the new character would simply consist of the Frobenius transform followed by χ_i . For the character χ_A on $A = \sum_{i=1}^r \mathbb{F}_p[x]/(A_i)$ we choose

$$\chi_A(x_1, \dots, x_r) = \prod_{i=1}^r \omega(x_i)^{\alpha_i(q_i-1)}.$$

Let $\sigma \in \text{Gal}(\overline{\mathbb{Q}}/K)$. It acts as $\omega(x) \mapsto \omega(x)^k$ for some integer k . Hence

$$\chi_A(x_1, \dots, x_r)^\sigma = \prod_{i=1}^r \omega(x_i)^{k\alpha_i(q_i-1)}.$$

This permutes the factors by a permutation $s \in S_r$ and we get

$$\chi_A(x_1, \dots, x_r)^\sigma = \prod_{i=1}^r \omega(x_{s^{-1}(i)})^{p^{l_i} \alpha_i(q_i-1)},$$

where $0 \leq l_i < \deg(A_i)$ for each i . We used $q_{s(i)} = q_i$. We finally get

$$\chi_A(x_1, \dots, x_r)^\sigma = \prod_{i=1}^r \omega(x_{s^{-1}(i)}^{p^{l_i}})^{\alpha_i(q_i-1)} = \chi_A(x_{s^{-1}(1)}^{p^{l_1}}, \dots, x_{s^{-1}(r)}^{p^{l_r}}).$$

In other words, $\chi_A(x)^\sigma = \chi_A(\rho(x))$ for a suitable \mathbb{F}_p -automorphism ρ of A . Notice that norm and trace of A are preserved by ρ . A similar construction can be performed for $B(x)$. According to Proposition 2.9 we get $H_p(A, B|t) \in K$ for all $t \in \mathbb{F}_q^\times$.

In order to connect to the p -adic function G_p we take the inverse of the Teichmüller character for ω and compute the terms given in Definition 2.4 p -adically. The Gauss sum $g_A(\chi_A \omega_N^m)$ is the product of ordinary Gauss sums of the form $g(\omega^{(q-1)\alpha+m(1+p+\dots+p^{l-1})})$ over the field \mathbb{F}_q with $q = p^l$. The occurrence of $m(1 + p + \dots + p^{l-1})$ is due to $\omega(N_{\mathbb{F}_q/\mathbb{F}_p}(x)^m) = \omega(x)^{m(1+\dots+p^{l-1})}$. The Gross-Koblitz theorem for Gauss sums over \mathbb{F}_q with $q = p^l$ gives us

$$g_q(\omega^a) = - \prod_{i=0}^{l-1} \pi \left\{ \frac{p^i a}{q-1} \right\} \Gamma_p \left(\left\{ \frac{p^i a}{q-1} \right\} \right)$$

for every integer a . When applied to $a = (q-1)\alpha + m(q-1)/(p-1)$ this amounts to

$$-\prod_{i=0}^{l-1} \pi^{\left\{p^i \alpha + \frac{m}{p-1}\right\}} \Gamma_p \left(\left\{p^i \alpha + \frac{m}{p-1}\right\} \right).$$

Note that this is a product over the p -orbit containing α and each factor is precisely of the type that occur in the definition of the p -adic hypergeometric sum. A similar story goes for $B(x)$. As a result we get

$$\frac{g_A(\chi_A \omega_N^m) g_B(\overline{\chi_B} \omega_N^{-m})}{g_A(\chi_A) g_B(\overline{\chi_B})} = (-p)^{\Lambda(m)} \prod_{i=1}^d \frac{\Gamma_p \left(\left\{ \alpha_i + \frac{m}{p-1} \right\} \right) \Gamma_p \left(\left\{ -\beta_i - \frac{m}{p-1} \right\} \right)}{\Gamma_p(\{\alpha_i\}) \Gamma_p(\{-\beta_i\})},$$

where $\Lambda(m)$ is as defined in the introduction. So we find that p -adically

$$H_p(A, B|t) = G_p(\alpha, \beta|t).$$

Hence we conclude that the values of G_p are in K . It remains to give an estimate for the denominator. The conjugates of $H_p(A, B|t)$ are obtained by taking χ_A^k, χ_B^k as multiplicative characters. The corresponding hypergeometric parameters are $k\alpha, k\beta$. From McCarthy’s work it follows that $p^\Delta G_p(k\alpha, k\beta|t)$ is a p adic integer for all k relatively prime to the common denominator of α_i, β_j . This implies that $p^\Delta H_p(A, B|t)$ is an algebraic integer in K .

References

1. Beukers, F., Heckman, G.: Monodromy for the hypergeometric function ${}_nF_{n-1}$. *Invent. Math.* **95**, 325–354 (1989)
2. Beukers, F., Cohen, H., Mellit, A.: Finite hypergeometric functions. *Pure Appl. Math. Q.* **11**, 559–589 (2015), arXiv:1505.02900
3. Cohen, H.: *Number Theory, Volume II: Analytic and Modern Tools*. Graduate Texts in Mathematics, vol. 240. Springer, New York (2007)
4. Greene, J.: Hypergeometric functions over finite fields. *Trans. Am. Math. Soc.* **301**, 77–101 (1987)
5. Katz, N.M.: *Exponential Sums and Differential Equations*. *Annals of Math Studies*, vol. 124. Princeton University Press, Princeton (1990)
6. McCarthy, D.: The trace of Frobenius of elliptic curves and the p -adic gamma function. *Pac. J. Math.* **261**(1), 219–236 (2013)

L-Series and Feynman Integrals



David Broadhurst and David P. Roberts

Abstract Integrals from Feynman diagrams with massive particles soon outgrow polylogarithms. We consider the simplest situation in which this occurs, namely for diagrams with two vertices in two space-time dimensions, with scalar particles of unit mass. These comprise vacuum diagrams, on-shell sunrise diagrams and diagrams obtained from the latter by cutting internal lines. In all these cases, the Feynman integral is a moment of $n = a + b$ Bessel functions, of the form $M(a, b, c) := \int_0^\infty I_0^a(t) K_0^b(t) t^c dt$. The corresponding L-series are built from Kloosterman sums over finite fields. Prior to the Creswick conference, the first author obtained empirical relations between special values of L-series and Feynman integrals with up to $n = 8$ Bessel functions. At the conference, the second author indicated how to extend these. Working together we obtained empirical relations involving Feynman integrals with up to 24 Bessel functions, from sunrise diagrams with up to 22 loops. We have related results for moments that lie beyond quantum field theory.

1 Physical and Mathematical Context

The context for our work is given in [1]. At the conference, the first author reported on the magnificent progress made by Stefano Laporta, whose solitary decade-long effort on the magnetic moment of the electron has come to fruition [4]. This

D. Broadhurst (✉)
Open University, Milton Keynes, UK
e-mail: David.Broadhurst@open.ac.uk

D. P. Roberts
University of Minnesota Morris, Morris, MN, USA
e-mail: roberts@morris.umn.edu

involves, inter alia, moments of 6 Bessel functions, one of which

$$M(1, 5, 1) := \int_0^\infty I_0(t)K_0^5(t)tdt = \frac{\pi^2 L_6(2)}{2} \tag{1}$$

is empirically related in [1] to the central value of the L-series of modular weight 4 and conductor 6, with Fourier series $q \prod_{k>0} (1 - q^k)^2 (1 - q^{2k})^2 (1 - q^{3k})^2 (1 - q^{6k})^2 = \sum_{k>0} A_6(k)q^k$. This modular form also delivers the Bessel moments

$$M(2, 4, 1) := \int_0^\infty I_0^2(t)K_0^4(t)tdt = \frac{3L_6(3)}{2} \tag{2}$$

$$M(3, 3, 1) := \int_0^\infty I_0^3(t)K_0^3(t)tdt = \frac{3L_6(2)}{2}. \tag{3}$$

For $n = 8$ Bessel functions, the L-series of a modular form of weight 6 and conductor 6 likewise gives evaluations of $M(1, 7, 1)$, $M(2, 6, 1)$, $M(3, 5, 1)$ and $M(4, 4, 1)$. The challenge presented at the conference was to find comparable relations between L-series and moments of $n > 8$ Bessel functions.

2 Progress at Creswick

This challenge was met by considering determinants of Feynman integrals, by allowing for adjustment of the local Kloosterman data, at primes that divide the conductor, and by empirical determination of the conductor, sign and gamma factors that enter the functional equation for the L-series. The good factors are classical and much useful information towards conductors was available in [5]. The formalism of [2] pointed us to the correct determinants. Using the methods in [3] for numerical computation of L-series, we were able to progress beyond the modular forms studied in [1].

We were successful for odd Bessel numbers up to $n = 17$ and even Bessel numbers up to $n = 20$. Let $\Omega_{a,b}$ be the determinant of the $r \times r$ matrix with $M(a, b, 1)$ at top left, size $r = \lceil (a + b)/4 - 1 \rceil$, powers of t^2 increasing to the right and powers of $I_0^2(t)$ increasing downwards. Then

$$L_{17}(8) = \frac{2^{15} \times 29 \Omega_{2,15}}{3^5 \times 5^2 \times 7\pi^{12}} \tag{4}$$

$$L_{20}(10) = \frac{2^{12} \times 11 \times 131 \Omega_{1,19}}{3^{11} \times 5^6 \times 7^3 \pi^{20}} \tag{5}$$

$$L_{20}(11) = \frac{2^{19} \times 17 \times 19 \times 23 \Omega_{2,18}}{3^{13} \times 5^7 \times 7^3 \pi^{12}} \tag{6}$$

are among findings made and presented at the conference.

3 Subsequent Progress

From data on Kloosterman sums in finite fields \mathbf{F}_q with $q < 200,000$, we found

$$L_{19}(8) = \frac{2^{14} \times 1093 \times 13171 \Omega_{2,17}}{3^4 \times 5^4 \times 7 \times 11\pi^{20}} \quad (7)$$

$$L_{24}(12) = \frac{2^{29} \times 12558877 \Omega_{1,23}}{3^{19} \times 5^9 \times 7^3 \times 11\pi^{30}} \quad (8)$$

$$L_{24}(13) = \frac{2^{27} \times 17 \times 19^2 \times 23^2 \times 46681 \Omega_{2,22}}{3^{23} \times 5^{12} \times 7^4 \times 11^2\pi^{20}}. \quad (9)$$

In parallel with the above results for odd moments, we obtained relations between even moments of Bessel functions and L-series determined by a quadratic twist of Kloosterman data. We have conjecturally complete sets of quadratic relations for both types of moment, encoded by Betti and de Rham matrices, for which we provide explicit constructions. In some cases where the sign of the functional equation is odd, we are able to define regulated moments that deliver central derivatives.

Acknowledgements We thank Ling Long, Masha Vlasenko and Wadim Zudilin for their splendid organization of a three-week conference on *Hypergeometric motives and Calabi–Yau differential equations*, held at Creswick, Australia, in January 2017, and the MATRIX Institute for financial support. DB thanks the University of Newcastle, NSW, and the Mainz Institute for Theoretical Physics, for hospitality and support that enabled further progress. DPR’s research is supported by grant DMS-1601350 from the National Science Foundation.

References

1. Broadhurst, D.: Feynman integrals, L-series and Kloosterman moments. *Commun. Number Theory Phys.* **10**, 527–569 (2016). <http://arxiv.org/abs/1604.03057>
2. Deligne, P.: Valeurs de fonctions L et périodes d’intégrales. *Proc. Sympos. Pure Math.* **33**, 313–346 (1979). <http://publications.ias.edu/deligne/paper/379>
3. Dokchitser, T.: Computing special values of motivic L-functions. *Exp. Math.* **13**, 137–149 (2004). <http://arxiv.org/abs/math/0207280>
4. Laporta, S.: High-precision calculation of the 4-loop contribution to the electron $g - 2$ in QED. <http://arxiv.org/abs/1704.06996>
5. Yun, Z.: Galois representations attached to moments of Kloosterman sums and conjectures of Evans. *Compos. Math.* **151**, 68–120 (2015). <http://arxiv.org/abs/1308.3920>

Arithmetic Properties of Hypergeometric Mirror Maps and Dwork's Congruences



Éric Delaygue

Abstract Mirror maps are power series which occur in Mirror Symmetry as the inverse for composition of $q(z) = \exp(f(z)/g(z))$, called local q -coordinates, where f and g are particular solutions of the Picard–Fuchs differential equations associated with certain one-parameter families of Calabi–Yau varieties. In several cases, it has been observed that such power series have integral Taylor coefficients at the origin. In the case of hypergeometric equations, we discuss p -adic tools and techniques that enable one to prove a criterion for the integrality of the coefficients of mirror maps. This is a joint work with T. Rivoal and J. Roques. This note is an extended abstract of the talk given by the author in January 2017 at the conference “Hypergeometric motives and Calabi–Yau differential equations” in Creswick, Australia.

1 Arithmetic Conditions for Operators of Calabi–Yau Type

An irreducible fourth order differential operator \mathcal{L} in $\mathbb{Q}(z)[d/dz]$ is of Calabi–Yau type if it is of Fuchsian type, self-dual, has 0 as MUM-point and it satisfies certain arithmetic conditions including that

- (i) \mathcal{L} has a solution $\omega_1(z) \in 1 + z\mathbb{C}[[z]]$ at $z = 0$ which is N -integral¹;
- (ii) \mathcal{L} has a linearly independent solution $\omega_2(z) = G(z) + \log(z)\omega_1(z)$ at $z = 0$ with $G(z) \in z\mathbb{C}[[z]]$ and $\exp(\omega_2(z)/\omega_1(z))$ is N -integral.

An additional condition is usually considered: the instanton numbers n_d associated with \mathcal{L} belong to $\frac{1}{N}\mathbb{Z}$ for some non-zero integer N . As far as we know, a systematic approach to prove the integrality of the n_d 's has not yet been developed, even in

¹A power series $f(z) \in 1 + z\mathbb{Q}[[z]]$ is N -integral if there is $c \in \mathbb{Q}^*$ such that $f(cz) \in \mathbb{Z}[[z]]$.

É. Delaygue (✉)

Camille Jordan Institute, Université Claude Bernard Lyon 1, Villeurbanne Cedex, France
e-mail: delaygue@math.univ-lyon1.fr

the case of hypergeometric equations. In this note, we discuss useful p -adic tools to prove or disprove Conditions (i) and (ii). A classical example of a differential operator satisfying both (i) and (ii) is

$$\mathcal{L} = \theta^4 - 5z(5\theta + 1)(5\theta + 2)(5\theta + 3)(5\theta + 4),$$

where $\theta = z \frac{d}{dz}$. Consider the two solutions

$$\omega_1(z) = \sum_{n=0}^{\infty} \frac{(5n)!}{n!^5} z^n \quad \text{and} \quad \omega_2(z) = G(z) + \log(z)\omega_1(z),$$

with

$$G(z) = \sum_{n=1}^{\infty} \frac{(5n)!}{n!^5} (5H_{5n} - 5H_n)z^n \quad \text{and} \quad H_n := \sum_{k=1}^n \frac{1}{k}.$$

Then $\omega_1(z)$ has integers coefficients and Lian and Yau proved in [10] that

$$\exp\left(\frac{\omega_2(z)}{\omega_1(z)}\right) \in \mathbb{Z}[[z]].$$

We shall see that hypergeometric techniques presented in this note allow to prove the integrality of the coefficients of q -coordinates associated with non-hypergeometric operators. For example, consider the differential operator

$$\mathcal{L} = \theta^3 - z(34\theta^3 + 51\theta^2 + 27\theta + 5) + z^2(\theta + 1)^3,$$

whose holomorphic solution is the generating series of the Apéry numbers used by Apéry in its proof of the irrationality of $\zeta(3)$ (see [1]):

$$\omega_1(z) = \sum_{n=0}^{\infty} \sum_{k=0}^n \binom{n}{k}^2 \binom{n+k}{k}^2 z^n.$$

A second solution is given by the method of Frobenius and reads $\omega_2(z) = G(z) + \log(z)\omega_1(z)$, with

$$G(z) = \sum_{n=1}^{\infty} \sum_{k=0}^n \binom{n}{k}^2 \binom{n+k}{k}^2 (2H_{n+k} - 2H_{n-k})z^n.$$

As we will see, a consequence of the results of the author [5] is that

$$\exp\left(\frac{\omega_2(z)}{\omega_1(z)}\right) \in \mathbb{Z}[[z]].$$

First, we present criteria on the integrality of hypergeometric terms.

2 Integrality of Hypergeometric Terms

2.1 Factorial Ratios

Let $\mathbf{e} = (e_1, \dots, e_u)$ and $\mathbf{f} = (f_1, \dots, f_v)$ be vectors of positive integers. For every non-negative integer n , we set

$$\mathcal{Q}(n) = \frac{(e_1 n)! \cdots (e_u n)!}{(f_1 n)! \cdots (f_v n)!}$$

and we consider the generating series of \mathcal{Q} :

$$F(z) = \sum_{n=0}^{\infty} \mathcal{Q}(n)z^n,$$

which is a rescaling of a hypergeometric function. We consider the function Δ of Landau defined for every x in \mathbb{R} by

$$\Delta(x) := \sum_{i=1}^u \lfloor e_i x \rfloor - \sum_{j=1}^v \lfloor f_j x \rfloor.$$

Let p be a prime number. By Legendre’s formula, we have

$$v_p(n!) = \sum_{\ell=1}^{\infty} \left\lfloor \frac{n}{p^\ell} \right\rfloor,$$

which yields

$$v_p(\mathcal{Q}(n)) = \sum_{\ell=1}^{\infty} \Delta\left(\frac{n}{p^\ell}\right).$$

Furthermore, we have $\Delta(x) = \Delta(\{x\}) + (|\mathbf{e}| - |\mathbf{f}|)\lfloor x \rfloor$, where $\{x\}$ is the fractional part of x and $|\mathbf{e}| = e_1 + \dots + e_u$. Hence the graph of Δ is essentially determined by its values on $[0, 1]$. Landau’s function provides a useful criterion for the N -integrality of $F(z)$.

Theorem 2.1 (Landau [9], Bober [2]) *The following assertions are equivalent.*

- (i) $F(z)$ is N -integral;
- (ii) $F(z) \in \mathbb{Z}[[z]]$;
- (iii) For all x in $[0, 1]$, we have $\Delta(x) \geq 0$.

Landau proved the equivalence of (ii) and (iii) in 1900 while Bober proved in 2009 a result which implies the equivalence with (i). One can easily compute the jumps of Δ on $[0, 1]$ to check Assertion (iii).

The generating series of factorial ratios are rescaling of hypergeometric functions whose parameters have a certain symmetry. Namely, if $\alpha = (\alpha_1, \dots, \alpha_r)$ and $\beta = (\beta_1, \dots, \beta_s)$ are tuples of parameters in $\mathbb{Q} \cap (0, 1]$, then there is $C \in \mathbb{Q}^*$ such that, for every $n \in \mathbb{N}$, we have

$$C^n \frac{(\alpha_1)_n \cdots (\alpha_r)_n}{(\beta_1)_n \cdots (\beta_s)_n} = \frac{(e_1 n)! \cdots (e_u n)!}{(f_1 n)! \cdots (f_v n)!},$$

if, and only if

$$\frac{(X - e^{2i\pi\alpha_1}) \cdots (X - e^{2i\pi\alpha_r})}{(X - e^{2i\pi\beta_1}) \cdots (X - e^{2i\pi\beta_s})}$$

is a ratio of cyclotomic polynomials. We will see that, when this is not the case, we still have a criterion for the N -integrality of hypergeometric functions but it involves several Landau’s functions: the functions of Christol.

2.2 Generalized Hypergeometric Functions

Let $\alpha = (\alpha_1, \dots, \alpha_r)$ and $\beta = (\beta_1, \dots, \beta_s)$ be tuples of elements in $\mathbb{Q} \setminus \mathbb{Z}_{\leq 0}$. We set

$$F(z) = \sum_{n=0}^{\infty} \frac{(\alpha_1)_n \cdots (\alpha_r)_n}{(\beta_1)_n \cdots (\beta_s)_n} z^n.$$

If $\beta_i = 1$ for some i , then $F(z)$ is annihilated by the hypergeometric differential operator

$$\mathcal{L} = \prod_{i=1}^s (\theta + \beta_i - 1) - z \prod_{i=1}^r (\theta + \alpha_i),$$

which is irreducible if, and only if $\alpha_i \not\equiv \beta_j \pmod{\mathbb{Z}}$. Elementary calculations show that $F(z)$ is N -integral if and only if, for almost all primes p , we have $F(z) \in \mathbb{Z}_{(p)}[[z]]$, where $\mathbb{Z}_{(p)}$ is the set of the rational numbers whose denominator is not divisible by p .

We introduce some definitions to construct useful functions defined by Christol in [3]. If x is a rational number, then we set

$$\langle x \rangle = \begin{cases} \{x\} & \text{if } x \notin \mathbb{Z}, \\ 1 & \text{otherwise.} \end{cases}$$

We write \preceq for the total order on \mathbb{R} defined by

$$x \preceq y \iff (\langle x \rangle < \langle y \rangle \text{ or } (\langle x \rangle = \langle y \rangle \text{ and } x \geq y)).$$

Let d be the common multiple of the exact denominators of the α_i ’s and β_j ’s. For all a coprime to d , $1 \leq a \leq d$, we set

$$\xi_a(x) := \#\{1 \leq i \leq r : a\alpha_i \preceq x\} - \#\{1 \leq j \leq s : a\beta_j \preceq x\}.$$

Then we have the following criterion for the N -integrality of $F(z)$.

Theorem 2.2 (Christol [3]) *The following assertions are equivalent.*

- (i) $F(z)$ is N -integral;
- (ii) For all a coprime to d , $1 \leq a \leq d$, and all x in \mathbb{R} , we have $\xi_a(x) \geq 0$.

If $F(z)$ is N -integral, then the set of constants $c \in \mathbb{Q}$ such that $F(cz) \in \mathbb{Z}[[z]]$ is $C\mathbb{Z}$ for some $C \in \mathbb{Q}^*$. When $F(z)$ is algebraic over $\overline{\mathbb{Q}}(z)$, then $F(z)$ is N -integral and C is called the Eisenstein constant of F . Hence we shall also call C the *Eisenstein constant* of $F(z)$. Rivoal, Roques and the author gave in [6] a formula for C when the parameters of the hypergeometric function belong to $(0, 1]$.

For every prime p , we set

$$\lambda_p = \#\{1 \leq i \leq r : \alpha_i \in \mathbb{Z}_{(p)}\} - \#\{1 \leq j \leq s : \beta_j \in \mathbb{Z}_{(p)}\}.$$

If α is a rational number, then we write $\text{den}(\alpha)$ for its exact denominator. As a particular case of Theorem 1 in [6], we have the following formula.

Theorem 2.3 *If α and β are tuples of elements in $(0, 1]$, $r = s$ and $F(z)$ is N -integral, then the Eisenstein constant of F is*

$$C = \frac{\prod_{i=1}^r \text{den}(\alpha_i)}{\prod_{j=1}^s \text{den}(\beta_j)} \prod_{p|d} p^{-\lfloor \frac{\lambda_p}{p-1} \rfloor}.$$

In the case of factorial ratios, if $\mathbf{e} = (e_1, \dots, e_u)$ and $\mathbf{f} = (f_1, \dots, f_v)$ are tuples of positive integers, then we have

$$\frac{(e_1 n)! \cdots (e_u n)!}{(f_1 n)! \cdots (f_v n)!} = \left(\frac{e_1^{e_1} \cdots e_u^{e_u}}{f_1^{f_1} \cdots f_v^{f_v}} \right)^n \frac{\prod_{i=1}^u \prod_{r=1}^{e_i} (r/e_i)_n}{\prod_{j=1}^v \prod_{r=1}^{f_j} (r/f_j)_n}.$$

If the associated generating series is (N) -integral then the Eisenstein constant is indeed

$$C = \frac{e_1^{e_1} \cdots e_u^{e_u}}{f_1^{f_1} \cdots f_v^{f_v}}.$$

2.2.1 Landau-Like Functions

To prove Theorem 2.3, we use Landau-like functions to calculate the p -adic valuation of Pochhammer’s symbols. To define those functions, we first consider a map \mathfrak{D}_p introduced by Dwork as follows.

Let p be a prime and α in $\mathbb{Z}_{(p)}$. We write $\mathfrak{D}_p(\alpha)$ for the unique element in $\mathbb{Z}_{(p)}$ satisfying $p\mathfrak{D}_p(\alpha) - \alpha \in \{0, \dots, p - 1\}$. We have $\mathfrak{D}_p(1) = 1$ and if $\alpha = r/N$ with r coprime to $N \geq 2$, $1 \leq r \leq N$, then

$$\mathfrak{D}_p(\alpha) = \frac{s_N(\pi_N(p)^{-1}\pi_N(r))}{N},$$

where s_N is the section of the canonical morphism $\pi_N : \mathbb{Z} \rightarrow \mathbb{Z}/N\mathbb{Z}$ with values in $\{0, \dots, N - 1\}$.

If p does not divide d , then, for all positive integers ℓ , we define the Landau-like function

$$\Delta_{p,\ell}(x) = \sum_{i=1}^r \left[x - \mathfrak{D}_p^\ell(\alpha_i) - \frac{\lfloor 1 - \alpha_i \rfloor}{p^\ell} \right] - \sum_{j=1}^s \left[x - \mathfrak{D}_p^\ell(\beta_j) - \frac{\lfloor 1 - \beta_j \rfloor}{p^\ell} \right] + r - s.$$

Christol proved in [3] a Legendre-like formula involving Landau-like function.

Theorem 2.4 *If p does not divide d , then we have*

$$v_p \left(\frac{(\alpha_1)_n \cdots (\alpha_r)_n}{(\beta_1)_n \cdots (\beta_s)_n} \right) = \sum_{\ell=1}^{\infty} \Delta_{p,\ell} \left(\frac{n}{p^\ell} \right).$$

Our first task to prove Theorem 2.3 was to find a convenient analog of Legendre’s formula when p is a divisor of d . In this case, Dwork’s maps are not defined for every parameters α_i and β_j . To that end, we proved in [6] an average formula for primes dividing d .

Theorem 2.5 *Assume that α and β are tuples of r elements in $(0, 1]$ such that $F(z)$ is N -integral. Let p be a prime divisor of d and write $d = p^f D$ where D is coprime to p .*

For every a coprime to p , $1 \leq a \leq p^f$, and all positive integers ℓ , we choose a prime $p_{a,\ell}$ satisfying $p_{a,\ell} \equiv p^\ell \pmod{D}$ and $p_{a,\ell} \equiv a \pmod{p^f}$. Then

$$v_p \left(C^n \frac{(\alpha_1)_n \cdots (\alpha_r)_n}{(\beta_1)_n \cdots (\beta_s)_n} \right) = \frac{1}{\varphi(p^f)} \sum_{\substack{a=1 \\ \gcd(a,p)=1}}^{p^f} \sum_{\ell=1}^{\infty} \Delta_{p_{a,\ell},1} \left(\frac{n}{p^\ell} \right) + n \left\{ \frac{\lambda_p}{p-1} \right\}.$$

In the case of factorial ratios, we have again

$$\frac{(e_1n)! \cdots (e_un)!}{(f_1n)! \cdots (f_vn)!} = C^n \frac{(\alpha_1)_n \cdots (\alpha_r)_n}{(\beta_1)_n \cdots (\beta_s)_n}.$$

α and β are tuples of elements in $(0, 1]$. For every p not dividing d and every ℓ , the map \mathfrak{D}_p^ℓ induces a permutation on α and β . Hence we have

$$\begin{aligned} \Delta_{p,\ell}(x) &= \sum_{i=1}^r \left[x - \mathfrak{D}_p^\ell(\alpha_i) - \frac{\lfloor 1 - \alpha_i \rfloor}{p^\ell} \right] - \sum_{j=1}^s \left[x - \mathfrak{D}_p^\ell(\beta_j) - \frac{\lfloor 1 - \beta_j \rfloor}{p^\ell} \right] + r - s \\ &= \sum_{i=1}^r \left[x - \mathfrak{D}_p^\ell(\alpha_i) \right] - \sum_{j=1}^s \left[x - \mathfrak{D}_p^\ell(\beta_j) \right] + r - s \\ &= \sum_{i=1}^r [x - \alpha_i] - \sum_{j=1}^s [x - \beta_j] + r - s \\ &= \sum_{i=1}^u [e_i x] - \sum_{j=1}^v [f_j x] \\ &= \Delta(x). \end{aligned}$$

Hence, in both cases, the formulas of Theorems 2.4 and 2.5 reduce to Legendre's one.

3 Integrality of the Coefficients of q -Coordinates

3.1 A Glimpse of Dwork's Result

Consider the power series

$$\begin{aligned} F(z) &= \sum_{n=0}^{\infty} \frac{(\alpha_1)_n \cdots (\alpha_r)_n}{(\beta_1)_n \cdots (\beta_s)_n} z^n, \\ G(z) &= \sum_{n=0}^{\infty} \frac{(\alpha_1)_n \cdots (\alpha_r)_n}{(\beta_1)_n \cdots (\beta_s)_n} \left(\sum_{i=1}^r H_{\alpha_i}(n) - \sum_{j=1}^s H_{\beta_j}(n) \right) z^n, \end{aligned}$$

where, for $n \in \mathbb{N}$ and $x \in \mathbb{Q} \setminus \mathbb{Z}_{\leq 0}$, we set $H_x(n) = \sum_{k=0}^{n-1} \frac{1}{x+k}$.

Then $G(z) + \log(z)F(z)$ is annihilated by the hypergeometric operator \mathcal{L} if there are at least two 1's in β . The q -coordinate is

$$q(z) = \exp\left(\frac{G(z) + \log(z)F(z)}{F(z)}\right) = z \exp\left(\frac{G(z)}{F(z)}\right).$$

A consequence of a lemma of Dieudonné and Dwork is that, for every prime p , we have

$$q(z) \in \mathbb{Z}_p[[z]] \iff \frac{G}{F}(z^p) - p \frac{G}{F}(z) \in p\mathbb{Z}_p[[z]].$$

Let p be a prime not dividing d and write $F_1(z)$ (resp. $G_1(z)$) for $F(z)$ (resp. $G(z)$) with the substitutions

$$\alpha \leftrightarrow (\mathfrak{D}_p(\alpha_1), \dots, \mathfrak{D}_p(\alpha_r)) \quad \text{and} \quad \beta \leftrightarrow (\mathfrak{D}_p(\beta_1), \dots, \mathfrak{D}_p(\beta_s)).$$

Then Dwork proved in [7] the following. Assume that $r = s$, for all $\ell \in \mathbb{N}$, $\mathfrak{D}_p^\ell(\beta_i) \in \mathbb{Z}_p^\times$, plus some fundamental but hard to read interlacing conditions (depending on p) on elements of α and β . Then we have

$$\frac{G_1}{F_1}(z^p) - p \frac{G}{F}(z) \in p\mathbb{Z}_p[[z]].$$

In particular, if \mathfrak{D}_p induces a permutation on α and β , which is the case for factorial ratios, then $F_1 = F$, $G_1 = G$ and Dwork's result yields

$$\frac{G}{F}(z^p) - p \frac{G}{F}(z) \in p\mathbb{Z}_p[[z]],$$

so that $q(z) \in \mathbb{Z}_p[[z]]$.

3.2 Factorial Ratios

If the interlacing conditions hold for every (explicitly) large enough primes p , then $q(z)$ is N -integral. Methods for the remaining primes were developed by Lian-Yau [10], Zudilin [11], Krattenthaler–Rivoal [8] for infinite families of factorial ratios, yielding proofs of $q(Cz) \in \mathbb{Z}[[z]]$ where C is the Eisenstein constant of $F(z)$.

In the case of factorial ratios, we have

$$G(z) = \sum_{n=0}^{\infty} \frac{(e_1 n)! \cdots (e_u n)!}{(f_1 n)! \cdots (f_v n)!} \left(\sum_{i=1}^u e_i H_{e_i n} - \sum_{j=1}^v f_j H_{f_j n} \right) z^n$$

and

$$\Delta(x) = \sum_{i=1}^u \lfloor e_i x \rfloor - \sum_{j=1}^v \lfloor f_j x \rfloor.$$

We gave a criterion for the integrality of the Taylor coefficients of $q(z)$ in 2012 (see [4]).

Theorem 3.1 *If $F(z)$ is N -integral with Eisenstein constant C , then the following assertions are equivalent.*

- (i) $q(z)$ is N -integral;
- (ii) $q(Cz) \in \mathbb{Z}[[z]]$;
- (iii) we have $|\mathbf{e}| = |\mathbf{f}|$ and, for all $x \in [1/M, 1)$, we have $\Delta(x) \geq 1$, where M is the largest element in \mathbf{e} and \mathbf{f} .

The proof of (iii) \Rightarrow (i) is essentially a consequence of Dwork’s results. Legendre’s formula and Landau’s functions play an important role in the proof of Theorem 3.1. When $F(z)$ is the generating series of multisums of binomial coefficients (such as Apéry numbers), it seems impossible to apply an analog of the proof of Theorem 3.1. To prove the integrality of the coefficients of the associated q -coordinate, we prove a generalization of Theorem 3.1 to several variables and then we specialize the multivariate q -coordinates.

3.3 Factorial Ratios of Linear Forms

Let $e = (\mathbf{e}_1, \dots, \mathbf{e}_u)$ and $f = (\mathbf{f}_1, \dots, \mathbf{f}_v)$ be tuples of nonzero vectors in \mathbb{N}^d . Consider

$$F(\mathbf{z}) = \sum_{\mathbf{n} \in \mathbb{N}^d} \frac{(\mathbf{e}_1 \cdot \mathbf{n})! \cdots (\mathbf{e}_u \cdot \mathbf{n})!}{(\mathbf{f}_1 \cdot \mathbf{n})! \cdots (\mathbf{f}_v \cdot \mathbf{n})!} \mathbf{z}^{\mathbf{n}}.$$

For every $k \in \{1, \dots, d\}$, write

$$G_k(\mathbf{z}) = \sum_{\mathbf{n} \in \mathbb{N}^d} \frac{(\mathbf{e}_1 \cdot \mathbf{n})! \cdots (\mathbf{e}_u \cdot \mathbf{n})!}{(\mathbf{f}_1 \cdot \mathbf{n})! \cdots (\mathbf{f}_v \cdot \mathbf{n})!} \left(\sum_{i=1}^u e_i^{(k)} H_{\mathbf{e}_i \cdot \mathbf{n}} - \sum_{j=1}^v f_j^{(k)} H_{\mathbf{f}_j \cdot \mathbf{n}} \right) \mathbf{z}^{\mathbf{n}},$$

where $e_i^{(k)}$ is the k -th component of \mathbf{e}_i . The q -coordinates are

$$q_k(\mathbf{z}) = z_k \exp \left(\frac{G_k(\mathbf{z})}{F(\mathbf{z})} \right), \quad 1 \leq k \leq n.$$

The associated Landau function is

$$\Delta(\mathbf{x}) = \sum_{i=1}^u \lfloor \mathbf{e}_i \cdot \mathbf{x} \rfloor - \sum_{j=1}^v \lfloor \mathbf{f}_j \cdot \mathbf{x} \rfloor, \quad (\mathbf{x} \in \mathbb{R}^d).$$

The non-trivial zone for Δ is defined by

$$\mathcal{D} := \{ \mathbf{x} \in [0, 1)^d : \text{there is } \mathbf{d} \text{ in } \mathbf{e} \text{ or } \mathbf{f} \text{ such that } \mathbf{d} \cdot \mathbf{x} \geq 1 \}.$$

Observe that if \mathbf{x} belongs to $[0, 1)^d \setminus \mathcal{D}$, then we have $\Delta(\mathbf{x}) = 0$. We proved in [5] the following criterion.

Theorem 3.2 *Assume that $F(\mathbf{z}) \in \mathbb{Z}[[\mathbf{z}]]$. Then the following assertions are equivalent:*

- (i) *For every k , we have $q_k(\mathbf{z}) \in \mathbb{Z}[[\mathbf{z}]]$;*
- (ii) *we have $|e| = |f|$ and, for every $\mathbf{x} \in \mathcal{D}$, $\Delta(\mathbf{x}) \geq 1$.*

To apply Theorem 3.2 to the case of Apéry numbers (associated with $\zeta(3)$), we consider the bivariate power series

$$F(x, y) = \sum_{n_1, n_2 \geq 0} \frac{(2n_1 + n_2)!^2}{n_1!^4 n_2!^2} x^{n_1} y^{n_2}$$

and

$$G_2(x, y) = \sum_{n_1, n_2 \geq 0} \frac{(2n_1 + n_2)!^2}{n_1!^4 n_2!^2} (2H_{2n_1+n_2} - 2H_{n_2}) x^{n_1} y^{n_2}.$$

In this case, we have

$$\Delta(x, y) = 2\lfloor 2x + y \rfloor - 4\lfloor x \rfloor - 2\lfloor y \rfloor.$$

We have

$$\mathcal{D} = \{ (x, y) \in [0, 1)^2 : 2x + y \geq 1 \}$$

and if $\mathbf{x} \in \mathcal{D}$, then $\Delta(\mathbf{x}) \geq 2$. Hence we have $q_2(x, y) \in \mathbb{Z}[[x, y]]$ by Theorem 3.2. Taking $x = y$ yields

$$q_2(x, x) = \exp\left(\frac{G_2(x, x)}{F(x, x)}\right) \in \mathbb{Z}[[x]],$$

where

$$F(x, x) = \sum_{n=0}^{\infty} \sum_{k=0}^n \binom{n}{k}^2 \binom{n+k}{k}^2 x^n$$

and

$$G_2(x, x) = \sum_{n=1}^{\infty} \sum_{k=0}^n \binom{n}{k}^2 \binom{n+k}{k}^2 (2H_{n+k} - 2H_{n-k})x^n,$$

as expected.

3.4 Generalized Hypergeometric q -Coordinates

In this section, we briefly comment analog results in the (univariate) general case.

Write $m(a)$ for the smallest element in $(\{a\alpha_1, \dots, a\alpha_r, a\beta_1, \dots, a\beta_s\}, \leq)$. We consider the following assertion, denoted H : For all a coprime to d , $1 \leq a \leq d$, for all $x \in \mathbb{R}$ satisfying $m(a) \leq x < a$, we have $\xi_a(x) \geq 1$. We consider a product of q -coordinates whose N -integrality is strongly related to the one of $q(z)$:

$$\tilde{q}(z) = \prod_{a=1, \gcd(a,d)=1}^d q_{(a\alpha), (a\beta)}(z).$$

Then we proved in [6] the following criterion.

Theorem 3.3 *Assume that \mathcal{L} is irreducible and that $F(z)$ is N -integral. Then*

- (i) *if $r = s$ and Assertion H holds, then $\tilde{q}(z)$ is N -integral.*

Furthermore, the following assertions are equivalent:

- (ii) *$q(z)$ is N -integral;*
- (iii) *$\tilde{q}(z)$ is N -integral and $\tilde{q}(z) = q(z)^{\varphi(d)}$.*

3.5 A Brief Overview of the p -Adic Strategy

The first step is to reduce the problem for each prime by the following classical result: if $x \in \mathbb{Q}$, then $x \in \mathbb{Z}$ if and only if $x \in \mathbb{Z}_p$ for all primes p .

Then we get ride of the exponential by applying the lemma of Dieudonné and Dwork.

Lemma 3.4

$$z \exp\left(\frac{G(z)}{F(z)}\right) \in \mathbb{Z}_p[[z]] \iff \frac{G}{F}(z^p) - p \frac{G}{F}(z) \in pz\mathbb{Z}_p[[z]].$$

Then, in all proofs, one has to generalize a theorem on formal congruences of Dwork to prove that

$$F_{s-1}(z^p)F(z) \equiv F(z^p)F_s(z) \pmod{p^s \mathbb{Z}_p[[z]]}, \quad (\forall s \geq 1),$$

where $F_s(z) := \sum_{n=0}^{p^s-1} a_n z^n$ and $F(z) = \sum_{n=0}^{\infty} a_n z^n$.

The last main step is to prove congruences for harmonic numbers $H_\alpha(n)$ and the p -adic Gamma function.

Acknowledgements This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No 648132.

References

1. Apéry, R.: Irrationalité de $\zeta(2)$ et $\zeta(3)$. *Astérisque* **61**, 11–13 (1979)
2. Bober, J.W.: Factorial ratios, hypergeometric series, and a family of step functions. *J. Lond. Math. Soc. (2)* **79**, 422–444 (2009)
3. Christol, G.: Fonctions hypergéométriques bornées. *Groupe de travail d'analyse ultramétrique*, tome **14**, exp. 8, 1–16 (1986–1987)
4. Delaygue, É.: Critère pour l'intégralité des coefficients de Taylor des applications miroir. *J. Reine Angew. Math.* **662**, 205–252 (2012)
5. Delaygue, É.: A criterion for the integrality of the Taylor coefficients of mirror maps in several variables. *Adv. Math.* **234**, 414–452 (2013)
6. Delaygue, É., Rivoal, T., Roques, J.: On Dwork's p -adic formal congruences theorem and hypergeometric mirror maps. *Mem. Am. Math. Soc.* **246**(1163), 100 pp (2017)
7. Dwork, B.: On p -adic differential equations IV generalized hypergeometric functions as p -adic analytic functions in one variable. *Annales Scientifiques de l'É.N.S. 4^o série*, tome **6**(3), 295–316 (1973)
8. Krattenthaler, C., Rivoal, T.: On the integrality of the Taylor coefficients of mirror maps. *Duke Math. J.* **151**, 175–218 (2010)
9. Landau, E.: Sur les conditions de divisibilité d'un produit de factorielles par un autre. *Collected works*, I, p. 116. Thales-Verlag, New Cairo (1985)
10. Lian, B.H., Yau, S.T.: Integrality of certain exponential series. *Algebra and Geometry (Taipei, 1995)*, pp. 215–227. *Lectures on Algebraic Geometry*, vol. 2. International Press, Cambridge (1998)
11. Zudilin, V.V.: Integrality of power expansions related to hypergeometric series. *Math. Notes* **71**(5), 604–616 (2002)

Appell–Lauricella Hypergeometric Functions over Finite Fields, and a New Cubic Transformation Formula



Sharon Frechette, Holly Swisher, and Fang-Ting Tu

Abstract We define a finite-field version of Appell–Lauricella hypergeometric functions built from period functions in several variables, paralleling the development by Fuselier et al. (Hypergeometric functions over finite fields, arXiv:1510.02575v2) in the single variable case. We develop geometric connections between these functions and the family of generalized Picard curves. In our main result, we use finite-field Appell–Lauricella functions to establish a finite-field analogue of Koike and Shiga’s cubic transformation (Koike and Shiga, *J. Number Theory* 124:123–141, 2007) for the Appell hypergeometric function F_1 , proving a conjecture of Ling Long. We also prove a finite field analogue of Gauss’ quadratic arithmetic geometric mean. We use our multivariable period functions to construct formulas for the number of \mathbb{F}_p -points on the generalized Picard curves. Lastly, we give some transformation and reduction formulas for the period functions, and consequently for the finite-field Appell–Lauricella functions.

1 Cubic Transformation Formulas

Classical hypergeometric functions are among the most versatile of all special functions. These functions and their finite-field analogues have numerous applications in number theory and geometry. For instance, finite-field hypergeometric functions play a role in proving congruences and supercongruences, they count points modulo p over algebraic varieties and affine hypersurfaces, and in certain instances they

S. Frechette (✉)
College of the Holy Cross, Worcester, MA, USA
e-mail: sfrechet@holycross.edu

H. Swisher (✉)
Oregon State University, Corvallis, OR, USA
e-mail: swisherh@math.oregonstate.edu

F.-T. Tu
Louisiana State University, Baton Rouge, LA, USA
e-mail: tu@math.lsu.edu

provide formulas for the Fourier coefficients of modular forms. We define finite-field hypergeometric functions $\mathbb{F}_D^{(n)}$ in several variables, as an analogue of the classical Appell–Lauricella hypergeometric functions of type D . Lauricella’s series of type D [15] give a natural generalization of Appell’s F_1 functions [1, 2] to n variables and are closely related to generalized Picard curves [18]. Following the literature, we refer to these generalizations as Appell–Lauricella functions. For a comprehensive survey of Appell–Lauricella functions, we refer the reader to the article by Schlosser [19], and to the monograph by Slater [20]. Furthermore, we note that classical hypergeometric functions as well as Appell–Lauricella functions are examples of a more general class called A -hypergeometric functions introduced and studied by Gelfand et al. [10], and further studied by Beukers [4].

We develop the theory of these $\mathbb{F}_D^{(n)}$ finite field hypergeometric functions in several variables, with a focus on their geometric connections to the generalized Picard curves. This parallels the construction (by the second and third authors et al.) in [9], categorizing the interplay between classical and finite-field hypergeometric functions in the single-variable setting.

Our results are motivated by a conjecture of Ling Long, related to identities proved by Koike and Shiga [13, 14]. In [13], Koike and Shiga applied Appell’s F_1 hypergeometric function in two variables to establish a new three-term arithmetic geometric mean result (AGM), related to Picard modular forms. As a consequence of this cubic AGM, Koike and Shiga proved the following cubic transformation for Appell’s F_1 -function. Let $x, y \in \mathbb{C}$, and let ω be a primitive cubic root of unity. Then

$$\begin{aligned}
 &F_1 \left[\frac{1}{3}; \frac{1}{3}, \frac{1}{3}; 1 \mid 1 - x^3, 1 - y^3 \right] \\
 &= \frac{3}{1 + x + y} F_1 \left[\frac{1}{3}; \frac{1}{3}, \frac{1}{3}; 1 \mid \left(\frac{1 + \omega x + \omega^2 y}{1 + x + y} \right)^3, \left(\frac{1 + \omega^2 x + \omega y}{1 + x + y} \right)^3 \right].
 \end{aligned}
 \tag{1}$$

As an application of Appell–Lauricella functions over finite fields, we prove the following finite-field analogue of Koike and Shiga’s transformation, as conjectured by Ling Long.

Theorem 1 *Let q be a prime power with $q \equiv 1 \pmod{3}$, let ω be a primitive cubic root of unity, and let η_3 be a primitive cubic character in $\widehat{\mathbb{F}_q^\times}$. If $\lambda, \mu \in \mathbb{F}_q$ satisfy $1 + \lambda + \mu \neq 0$, then*

$$\begin{aligned}
 &\mathbb{F}_D^{(2)} \left[\begin{matrix} \eta_3; \eta_3 \eta_3 \\ \varepsilon \end{matrix}; 1 - \lambda^3, 1 - \mu^3 \right] \\
 &= \mathbb{F}_D^{(2)} \left[\begin{matrix} \eta_3; \eta_3 \eta_3 \\ \varepsilon \end{matrix}; \left(\frac{1 + \omega\lambda + \omega^2\mu}{1 + \lambda + \mu} \right)^3, \left(\frac{1 + \omega^2\lambda + \omega\mu}{1 + \lambda + \mu} \right)^3 \right].
 \end{aligned}$$

When $\lambda = \mu$, we have the following corollary.

Corollary 1 *For q a prime power with $q \equiv 1 \pmod{3}$, and ω as above, if $\lambda \in \mathbb{F}_q$ satisfies $1 + 2\lambda \neq 0$, then*

$${}_2F_1 \left[\begin{matrix} \eta_3 & \eta_3^2 \\ \varepsilon \end{matrix} ; 1 - \lambda^3 \right] = {}_2F_1 \left[\begin{matrix} \eta_3 & \eta_3^2 \\ \varepsilon \end{matrix} ; \left(\frac{1 - \lambda}{1 + 2\lambda} \right)^3 \right].$$

The result of Corollary 1 was first established in [9], using a different method of proof. It is a finite-field version of the cubic transformation

$${}_2F_1 \left[\begin{matrix} \frac{1}{3} & \frac{2}{3} \\ 1 \end{matrix} ; 1 - x^3 \right] = \frac{3}{1 + 2x} {}_2F_1 \left[\begin{matrix} \frac{1}{3} & \frac{2}{3} \\ 1 \end{matrix} ; \left(\frac{1 - x}{1 + 2x} \right)^3 \right], \tag{2}$$

proved by Borwein and Borwein [5, 6] for $x \in \mathbb{R}$ with $0 < x < 1$, as a cubic analogue of Gauss’ quadratic AGM.

2 Quadratic Transformations: Revisiting Gauss’ Quadratic AGM

In [9], the authors give a dictionary for the correspondence between results on classical hypergeometric functions and finite-field hypergeometric functions. Given a transformation for classical hypergeometric functions, this dictionary can be used to predict the form of the analogous transformation for finite-field hypergeometric functions. They also use a calculus-style method of converting the proofs of classical identities to the finite-field setting, provided the classical identity satisfies the following condition: It can be proved using only the binomial theorem, the reflection and multiplication formulas [for the gamma function], or their corollaries (such as the Pfaff-Saalschütz formula) [9].

We illustrate this calculus-style method of translating classical results by proving the following theorem.

Theorem 2 *The quadratic arithmetic-geometric mean of Gauss, given for $x \in \mathbb{C}$ by*

$${}_2F_1 \left[\begin{matrix} \frac{1}{2}, \frac{1}{2} \\ 1 \end{matrix} \middle| 1 - x^2 \right] = \frac{2}{1 + x} {}_2F_1 \left[\begin{matrix} \frac{1}{2}, \frac{1}{2} \\ 1 \end{matrix} \middle| \left(\frac{1 - x}{1 + x} \right)^2 \right], \tag{3}$$

can be proved using only the binomial theorem, the reflection and duplication formulas for the gamma function, and special evaluations of the ${}_3F_2$ and ${}_2F_1$ functions, using the Pfaff-Saalschütz formula and Gauss’ formula, respectively.

As a consequence, translating this alternate proof of Gauss’ quadratic AGM and analyzing the associated error terms on the finite-field side, we also obtain the following corollary.

Corollary 2 *Let $p \equiv 1 \pmod{4}$ be prime, let ϕ be the quadratic character in $\widehat{\mathbb{F}_p^\times}$, and let ε be the trivial character in $\widehat{\mathbb{F}_p^\times}$. If $\lambda \in \mathbb{F}_p$ satisfies $1 + \lambda \neq 0$, then*

$${}_2\mathbb{F}_1 \left[\begin{matrix} \phi & \phi \\ \varepsilon \end{matrix} \middle| 1 - \lambda^2 \right] = {}_2\mathbb{F}_1 \left[\begin{matrix} \phi & \phi \\ \varepsilon \end{matrix} \middle| \left(\frac{1 - \lambda}{1 + \lambda} \right)^2 \right], \tag{4}$$

3 Connections to Picard Curves

Taking the approach used in [9], our finite-field Appell–Lauricella hypergeometric functions are defined as normalizations of finite-field period functions $\mathbb{P}_D^{(n)}$, which we also define. These period functions are naturally related to periods of the generalized Picard curves

$$C_\lambda^{[N;i,j,\mathbf{k}]} : y^N = x^i(1-x)^j(1-\lambda_1x)^{k_1} \cdots (1-\lambda_nx)^{k_n}, \tag{5}$$

defined for distinct complex numbers $\lambda_1, \dots, \lambda_n \neq 0, 1$ and positive integers N, i, j, k_1, \dots, k_n that satisfy the conditions $\gcd(N, i, j, k_1, \dots, k_n) = 1$ and $N \nmid i + j + k_1 + \cdots + k_n$. As a consequence, the $\mathbb{P}_D^{(n)}$ functions are ideally suited for counting \mathbb{F}_p -points on Picard curves. We prove a theorem which gives the number of \mathbb{F}_p -points on the generalized Picard curves in a simple, elegant formula. This is analogous to the point-counting result for the generalized Legendre curves that was established by the second and third authors, et al. in [7, 8]. We also compute the genus of the generalized Picard curves $C_\lambda^{[N;i,j,\mathbf{k}]}$, following methods of Archinard [3].

4 Transformation and Reduction Formulas

Transformation and reduction formulas for classical hypergeometric functions have been successfully translated to the finite-field setting, first by Greene and also by authors such as McCarthy, and Fuselier et al. (See [9, 11, 17] for details.) Transformation formulas for classical Appell–Lauricella hypergeometric functions, many of which can be found in the monograph by Slater [20] or the survey paper of Schlosser [19], may be translated into the finite-field setting using the same methods. We carry out this process, proving several identities for the period functions $\mathbb{P}_D^{(n)}$

and hypergeometric functions $\mathbb{F}_D^{(n)}$. Among other things, these include a finite-field analogue of the Pfaff-Kummer transformation,

$$F_1 \left[a; b_1, b_2; c \mid x, y \right] = (1-x)^{-b_1} (1-y)^{-b_2} F_1 \left[c-a; b_1, b_2; c \mid \frac{x}{x-1}, \frac{y}{y-1} \right],$$

and Euler's transformation,

$$F_1 \left[a; b_1, b_2; c \mid x, y \right] = (1-x)^{c-a-b_1} (1-y)^{-b_2} F_1 \left[c-a; c-b_1-b_2, b_2; c \mid x, \frac{x-y}{1-y} \right],$$

which hold for all $a, b_1, b_2, c \in \mathbb{C}$ and all x, y for which the series are defined.

We note that another version of finite-field Appell–Lauricella functions is independently defined by He [12] and Li et al. [16], which closely follows Greene's definition. For their version, they establish several degree 1 transformation and reduction formulas, including some that are analogous to the identities we prove.

Acknowledgements We thank Ling Long for providing the inspiration for this work. We also thank the Institute for Computational and Experimental Research in Mathematics (ICERM) for the special semester program on Computational Aspects of the Langlands Program where this work was initiated, and the International Mathematical Research Institute MATRIX in Australia for the workshop on Hypergeometric Motives and Calabi–Yau Differential Equations, where we further collaborated on this project.

References

1. Appell, P.: Sur les fonctions hypergéométriques de deux variables. *Journal de Mathématiques Pures et Appliquées 3e série* **8**, 173–216 (1882)
2. Appell, P.: Sur les Fonctions hypérgéométriques de plusieurs variables les polynomes d'Hermite et autres fonctions sphériques dans l'hyperspace. Gauthier–Villars, Paris (1925)
3. Archinard, N.: Hypergeometric abelian varieties. *Can. J. Math.* **55**(5), 897–932 (2003)
4. Beukers, F.: Algebraic A-hypergeometric functions. *Invent. Math.* **180**(3), 589–610 (2010)
5. Borwein, J.M., Borwein, P.M.: A Remarkable cubic mean iteration. In: Ruseheweyh, St., Saff, E.B., Salinas, L.C., Varga, R.S. (eds.) *Proceedings of the Valparaiso Conference. Lecture Notes in Mathematics*, vol. 1435, pp. 27–31. Springer, Berlin (1989)
6. Borwein, J.M., Borwein, P.B.: A cubic counterpart of Jacobi's identity and the AGM. *Trans. Am. Math. Soc.* **323**(2), 691–701 (1991)
7. Deines, A., Fuselier, J., Long, L., Swisher, H., Tu, F.: Generalized legendre curves and quaternionic multiplication. *J. Number Theory* (2014, in press). arXiv:1412.6906
8. Deines, A., Fuselier, J.G., Long, L., Swisher, H., Tu, F.-T.: Hypergeometric series, truncated hypergeometric series, and Gaussian hypergeometric functions. In: *Directions in Number Theory. Association for Women in Mathematics Series*, vol. 3, pp. 125–159. Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-319-30976-7_5 (2016)
9. Fuselier, J., Long, L., Ramakirshma, R., Swisher, H., Tu, F.: Hypergeometric functions over finite fields. arXiv:1510.02575v2
10. Gelfand, I.M., Kapranov, M.M., Zelevinsky, A.V.: Generalized Euler integrals and A-hypergeometric functions. *Adv. Math.* **84**(2), 255–271 (1990)

11. Greene, J.: Hypergeometric functions over finite fields. *Trans. Am. Math. Soc.* **301**, 77–101 (1987)
12. He, B.: A Lauricella hypergeometric function over finite fields. arXiv:1610.04473
13. Koike, K., Shiga, H.: Isogeny formulas for the Picard modular form and a three terms arithmetic geometric mean. *J. Number Theory* **124**, 123–141 (2007)
14. Koike, K., Shiga, H.: An extended Gauss AGM and corresponding Picard modular forms. *J. Number Theory* **128**, 2097–2126 (2008)
15. Lauricella, G.: Sulle funzioni ipergeometriche a più variabili. *Rendiconti del Circolo Matematico di Palermo (in Italian)* **7(S1)**, 111–158 (1893)
16. Li, L., Li, X., Mao, R.: Some new formulas for Appell series over finite fields. arXiv:1701.02674
17. McCarthy, D.: Extending Gaussian hypergeometric series to the p -adic setting. *Int. J. Number Theory* **8(7)**, 1581–1612 (2012)
18. Picard, E.: Sur une extension aux fonctions de deux variables du problème de Riemann relatif aux fonctions hypergéométriques. *Annales scientifiques de l'École Normale Supérieure. (2ème série)* **10**, 305–322 (1881, in French); *Acta Math.* **2**, 114–135 (1883)
19. Schlosser, M.: Multiple hypergeometric series – Appell series and beyond. In: Schneider, C., Blümlein, J. (eds.) *Computer Algebra in Quantum Field Theory. Springer Texts & Monographs in Symbolic Computation*, pp. 305–345. Springer, Vienna (2013)
20. Slater, L.J.: *Generalized Hypergeometric Functions*. Cambridge University Press, Cambridge (1966)

Sequences, Modular Forms and Cellular Integrals



Dermot McCarthy, Robert Osburn, and Armin Straub

Abstract It is well-known that the Apéry sequences which arise in the irrationality proofs for $\zeta(2)$ and $\zeta(3)$ satisfy many intriguing arithmetic properties and are related to the p th Fourier coefficients of modular forms. Here, we briefly indicate that the connection to modular forms persists for sequences associated to Brown's cellular integrals and state a general conjecture concerning supercongruences.

1 Introduction and Statement of Results

Recently, Brown [4] introduced a program where period integrals on the moduli space $\mathcal{M}_{0,N}$ of curves of genus 0 with N marked points play a central role in understanding irrationality proofs of values of the Riemann zeta function. The main idea of [4] is to associate a rational function f_σ and a differential $(N-3)$ -form ω_σ to a given permutation $\sigma = \sigma_N$ on $\{1, 2, \dots, N\}$. Consider the *cellular integral*

$$I_\sigma(n) := \int_{S_N} f_\sigma^n \omega_\sigma,$$

where

$$S_N = \{(t_1, \dots, t_{N-3}) \in \mathbb{R}^{N-3} : 0 < t_1 < \dots < t_{N-3} < 1\}.$$

D. McCarthy

Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX, USA
e-mail: dermot.mccarthy@ttu.edu

R. Osburn (✉)

School of Mathematics and Statistics, University College Dublin, Dublin, Ireland
e-mail: robert.osburn@ucd.ie

A. Straub

Department of Mathematics and Statistics, University of South Alabama, Mobile, AL, USA
e-mail: straub@southalabama.edu

For “convergent” σ , the integral $I_\sigma(n)$ converges and, for $n = 0$, we obtain the cell-zeta values $\zeta_\sigma(N - 3) = I_\sigma(0)$ studied in [5], which are multiple zeta values of weight $N - 3$. More generally, Brown [4] showed that $I_\sigma(n)$ is a \mathbb{Q} -linear combination of multiple zeta values of weight less than or equal to $N - 3$. If this linear combination is of the form $A_\sigma(n)\zeta_\sigma(N - 3)$, for some rational $A_\sigma(n)$, plus a combination of multiple zeta values of weight less than $N - 3$, then we say that $A_\sigma(n)$ is the *leading coefficient* of the cellular integral $I_\sigma(n)$.

This construction recovers Beukers’ integrals [3] which appear in the irrationality proofs of $\zeta(2)$ and $\zeta(3)$ and the Apéry numbers

$$a(n) = \sum_{k=0}^n \binom{n}{k}^2 \binom{n+k}{k}, \quad b(n) = \sum_{k=0}^n \binom{n}{k}^2 \binom{n+k}{k}^2$$

as leading coefficients. This framework also raises some natural questions. Is there an analogue of the results in [1] and [2] to higher weight modular forms? Do the leading coefficients $A_\sigma(n)$ satisfy supercongruences akin to those in [6]?

In [7], we prove that there is a higher weight version of Theorem 5 in [2] and Theorem 3 in [1] can be extended to all odd weights greater than or equal to 3. Based on numerical evidence, we also conjecture that for each $N \geq 5$ and convergent σ_N , the leading coefficients $A_{\sigma_N}(n)$ satisfy

$$A_{\sigma_N}(mp^r) \equiv A_{\sigma_N}(mp^{r-1}) \pmod{p^{3r}}$$

for all primes $p \geq 5$ and integers $m, r \geq 1$.

References

1. Ahlgren, S.: Gaussian hypergeometric series and combinatorial congruences. In: Symbolic Computation, Number Theory, Special Functions, Physics and Combinatorics (Gainesville, FL, 1999). Developments in Mathematics, vol. 4, pp. 1–12. Kluwer Academic Publishers, Dordrecht (2001)
2. Ahlgren, S., Ono, K.: A Gaussian hypergeometric series evaluation and Apéry number congruences, *J. Reine Angew. Math.* **518**, 187–212 (2000)
3. Beukers, F.: A note on the irrationality of $\zeta(2)$ and $\zeta(3)$. *Bull. Lond. Math. Soc.* **11**(3), 268–272 (1979)
4. Brown, F.: Irrationality proofs for zeta values, moduli spaces and dinner parties. *Mosc. J. Comb. Number Theory* **6**(2–3), 102–165 (2016)
5. Brown, F., Carr, S., Schneps, L.: The algebra of cell-zeta values. *Compos. Math.* **146**(3), 731–771 (2010)
6. Coster, M.: Supercongruences. Ph.D. thesis, Universiteit Leiden (1988)
7. McCarthy, D., Osburn, R., Straub, A.: Sequences, modular forms and cellular integrals. *Math. Proc. Camb. Philos. Soc.* (2018). arXiv: 1705.05586, <https://doi.org/10.1017/S0305004118000774>

Some Supercongruences for Truncated Hypergeometric Series



Ling Long and Ravi Ramakrishna

Abstract We prove various supercongruences involving truncated hypergeometric sums. These include a strengthened version of a conjecture of van Hamme. Our method is to employ various hypergeometric transformation and evaluation formulae to convert the truncated sums to quotients of Γ -values. We then convert these to quotients of Γ_p -values and use Taylor's Theorem to make p -adic approximations. In the cases under consideration higher order coefficients often vanish leading to the supercongruences.

In [3], van Hamme conjectured that

$$\begin{aligned} {}_7F_6 \left[\begin{matrix} \frac{1}{3} & \frac{7}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{6} & 1 & 1 & 1 & 1 & 1 & 1 \end{matrix} \right]_{p-1} &:= \sum_{k=0}^{p-1} (6k+1) \frac{((1/3)_k)^6}{(k!)^6} \\ &\equiv -p\Gamma_p(1/3)^9 \pmod{p^4} \text{ for } p \equiv 1 \pmod{6}. \end{aligned}$$

We have, in [2], proved

Theorem 1 *Let $p > 11$. Then*

$${}_7F_6 \left[\begin{matrix} \frac{1}{3} & \frac{7}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{6} & 1 & 1 & 1 & 1 & 1 & 1 \end{matrix} \right]_{p-1} \equiv -p\Gamma_p(1/3)^9 \pmod{p^6} \text{ for } p \equiv 1 \pmod{6}$$

L. Long
LSU, Baton Rouge, LA, USA
e-mail: llong@lsu.edu

R. Ramakrishna (✉)
Cornell University, Ithaca, NY, USA
e-mail: ravi@math.cornell.edu

and

$${}_7F_6 \left[\begin{matrix} \frac{1}{3} & \frac{7}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{6} & 1 & 1 & 1 & 1 & 1 \end{matrix} \right]_{p-1} \equiv -\frac{10}{27} p^4 \Gamma_p(1/3)^9 \pmod{p^6} \text{ for } p \equiv 5 \pmod{6}.$$

Note these are congruences cover (almost) all primes and are stronger than the van Hamme Conjecture. We proved a number of other supercongruences including the ${}_3F_2$ ones below:

Theorem 2

$${}_3F_2 \left[\begin{matrix} \frac{1}{3} & \frac{1}{3} \\ 1 & 1 \end{matrix} \right]_{p-1} \equiv \Gamma_p(1/3)^6 \pmod{p^3} \text{ for } p \equiv 1 \pmod{6}$$

and

$${}_3F_2 \left[\begin{matrix} \frac{1}{3} & \frac{1}{3} \\ 1 & 1 \end{matrix} \right]_{p-1} \equiv -\frac{p^2}{3} \Gamma_p(1/3)^6 \pmod{p^3} \text{ for } p \equiv 5 \pmod{6}.$$

For $p \equiv 1 \pmod{6}$ the right side of the above congruence corresponds to Dwork’s unit root for ordinary primes of a certain modular form that is part of the corresponding hypergeometric motive.

We outline the strategies for $p \equiv 1 \pmod{6}$. Various minor differences (and one on medium-sized technical issue in Theorem 1) arise for $p \equiv 5 \pmod{6}$. The idea in both theorems is to perturb the entries so that the series naturally truncate at $\frac{p-1}{3}$ (for $p \equiv 1 \pmod{6}$).

Let ζ_3 be a primitive cube root of unity. For instance in Theorem 2 we study ${}_3F_2 \left[\begin{matrix} \frac{1-p}{3} & \frac{1-\zeta_3 p}{3} & \frac{1-\zeta_3^2 p}{3} \\ 1 & 1 \end{matrix} \right]$ for $p \equiv 1 \pmod{6}$. The corresponding infinite series truncates at $\frac{p-1}{3}$. The Galois symmetry and a simple congruence argument imply

$${}_3F_2 \left[\begin{matrix} \frac{1-p}{3} & \frac{1-\zeta p}{3} & \frac{1-\zeta^2 p}{3} \\ 1 & 1 \end{matrix} \right] \equiv {}_3F_2 \left[\begin{matrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & 1 \end{matrix} \right]_{p-1} \pmod{p^3}.$$

At this point we can use the Pfaff-Saalschütz formula (see, for instance, Theorem 2.2.6 of [1])

$${}_3F_2 \left[\begin{matrix} -n & a & b \\ c & 1+a+b-c-n \end{matrix} \right] = \frac{(c-a)_n (c-b)_n}{(c)_n (c-a-b)_n}$$

with $n = \frac{p-1}{3}$ to write ${}_3F_2 \left[\begin{matrix} \frac{1-p}{3} & \frac{1-\zeta p}{3} & \frac{1-\zeta^2 p}{3} \\ 1 & 1 \end{matrix} \right]$ as a quotient of Γ -values. One can rewrite this as a quotient of Γ_p -values and then use a Taylor approximation to get the desired result.

For Theorem 1 a similar argument with primitive 5th roots of unity and Dougall’s formula below, which holds when $1 + 2a = b + c + d + e + f$,

$$\begin{aligned}
 & {}_7F_6 \left[\begin{matrix} a & 1 + \frac{a}{2} & b & c & d & e & f \\ & \frac{a}{2} & 1 + a - b & 1 + a - c & 1 + a - d & 1 + a - e & 1 + a - f \end{matrix} \right] \\
 &= \frac{(a + 1)_{-f} (a - b - c + 1)_{-f} (a - b - d + 1)_{-f} (a - c - d + 1)_{-f}}{(a - b + 1)_{-f} (a - c + 1)_{-f} (a - d + 1)_{-f} (a - b - c - d + 1)_{-f}}
 \end{aligned}$$

gives the van Hamme congruences mod p^5 .

To obtain the congruence mod p^6 involves an extra argument. It is not difficult to show the terminating

$${}_7F_6 \left[\begin{matrix} \frac{1}{3} & \frac{7}{6} & \frac{1}{3} - \zeta_5 x & \frac{1}{3} - \zeta_5^2 x & \frac{1}{3} - \zeta_5^3 x & \frac{1}{3} - \zeta_5^4 x & \frac{1}{3} - x \\ & \frac{1}{6} & 1 + \zeta_5 x & 1 + \zeta_5^2 x & 1 + \zeta_5^3 x & 1 + \zeta_5^4 x & 1 + x \end{matrix} \right]_{\frac{p-1}{3}} \in \mathbb{Z}_p[[x^5]].$$

Call this power series $G(x)$. Using a result of Bailey relating ${}_9F_8$ expressions one can in fact prove the above series is in $p\mathbb{Z}_p[[x^5]]$. A somewhat subtle argument is required when $p \equiv 5 \pmod 6$ to obtain the divisibility of $G(x)$ by p .

Since $p \mid G(x)$, $G(0) \equiv G(p/3) \pmod{p^6}$. It is easy to show that $G(0)$ is congruent to the left side of Theorem 1 mod p^6 . The argument using Dougall’s formula gives $G(p/3)$ is congruent to the right side of Theorem 1 mod p^6 .

References

1. Andrews, G., Askey, R., Roy, R.: Special Functions. Cambridge University Press, Cambridge (1999)
2. Long, L., Ramakrishna, R.: Some supercongruences occurring in truncated hypergeometric series. *Adv. Math.* **290**, 773–808 (2016)
3. van Hamme, L.: Some congruences involving the p-adic gamma function and some arithmetical consequences. In: *p-Adic Functional Analysis* (Ioannina, 2000). Lecture Notes in Pure and Applied Mathematics, vol. 222, pp. 133–138. Dekker, New York (2001)

The Explicit Formula and a Motivic Splitting



David P. Roberts

Abstract We apply the Guinand-Weil-Mestre explicit formula to resolve two questions about how a certain hypergeometric motive splits into two irreducible motives.

1 Introduction

The classical explicit formula of Guinand and Weil was generalized to a broader context by Mestre in [2]. This formula applies to any L -function satisfying standard analytic properties, and gives a family of formulas for its conductor N . Mestre used it to get lower bounds on conductors of abelian varieties. This extended abstract gives an example of how it can be used in more exotic motivic contexts.

The example we pursue here has the form $M = M_8 \oplus M_6$, the factor motives being indexed by their degree. We assume that the associated L -functions really do have the required analytic properties, and work numerically to a precision that is adequate for being very confident in the assertions. Presently, we can compute directly with M , but not with the individual factors. We know that its conductor is $\text{cond}(M) = 2^{15}$ and its local L -factor at 2 is just 1. These numerics imply that one of M_6 and M_8 is tame at 2, and the other is minimally wild. Also we know the order of central vanishing is $\text{rank}(M) = 2$. This raises two questions:

Q1: $(\text{rank}(M_6), \text{rank}(M_8))$ can only be $(2, 0)$, $(1, 1)$, or $(0, 2)$. Which is correct?

Q2: $(\text{cond}(M_6), \text{cond}(M_8))$ can only be $(2^6, 2^9)$ or $(2^7, 2^8)$. Which is correct?

The answers are given in the table at the end of this extended abstract. We provide enough computational details so that the reader can both reproduce our answers and attempt analogous calculations for other split motives.

D. P. Roberts (✉)
University of Minnesota Morris, Morris, MN, USA
e-mail: roberts@morris.umn.edu

2 The Motive $M = M_6 \oplus M_8$

One of the points of the talk was to illustrate how the *Magma* hypergeometric motives package by Mark Watkins lets one compute with hypergeometric motives of large degree. We use *Magma* language here as well [1], and the reader can repeat most computations using the free online *Magma* calculator.

To obtain the motive M and its L -function L , type

```
M:=HypergeometricData(
    [1/2: i in [1..16]], [0: i in [1..16]]);
L:=LSeries(M,1:Precision:=10,BadPrimes:=[<2,15,1>]);
```

Here *Magma* correctly understands that M has good reduction outside of 2. The optional argument ensures that it has the correct data at 2 as well, that being conductor 2^{15} and local L -factor 1. Other possibilities failing badly, correctness of the choice $\langle 2, 15, 1 \rangle$ is confirmed by `CheckFunctionalEquation(L)` returning 0.0000000000. The command `HodgeStructure(L:PHV)` says that M has weight $w = 15$ with Hodge vector

$$(h^{0,15}, h^{1,14}, \dots, h^{14,1}, h^{15,0}) = (1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1).$$

In particular M can only appear in the cohomology of varieties of dimension ≥ 15 .

In general, if d is even and the α_i 's and the β_j 's are obtained from one another by adding $1/2$ modulo \mathbb{Z} , then $H(\alpha, \beta|1)$ decomposes as a sum of two motives of specified degrees. In our case, we know a priori that $M = M_8 \oplus M_6$. `Factorization(EulerFactor(L, 3))` then yields $f_3(x)$ in 2s:

$$\begin{aligned} &(1 - 268 \cdot 3x + 204193 \cdot 3^4x^2 - 1001800 \cdot 3^9x^3 + 204193 \cdot 3^{19}x^4 \\ &\quad - 268 \cdot 3^{31}x^5 + 3^{45}x^6) \\ &(1 + 2992 \cdot x + 39116 \cdot 3^4x^2 - 7596496 \cdot 3^6x^3 - 203836426 \cdot 3^{12}x^4 \\ &\quad - 7596496 \cdot 3^{21}x^5 + 39116 \cdot 3^{34}x^6 + 2992 \cdot 3^{45}x^7 + 3^{60}x^8). \end{aligned}$$

Thus, M_6 and M_8 are both irreducible. Moreover Newton-over-Hodge forces the Hodge vector of M to decompose nicely into $h_6 + h_8$ with

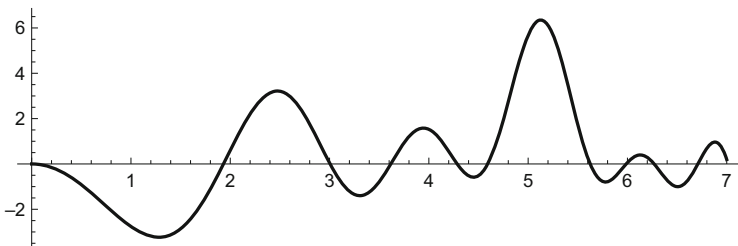
$$\begin{aligned} h_6 &:= (0, 1, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0), \\ h_8 &:= (1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 1). \end{aligned}$$

Likewise, but in 30 s, 8 min, and 2.5 h now,

$$\begin{aligned} f_5(x) &= (1 + 1614 \cdot 5^3x + \dots + 5^{45}x^6)(1 - 41208x + \dots + 5^{60}x^8), \\ f_7(x) &= (1 + 248232 \cdot 7x + \dots + 7^{45}x^6)(1 + 667104x + \dots + 7^{60}x^8), \\ f_{11}(x) &= (1 - 883812 \cdot 11x + \dots + 11^{45}x^6)(1 + 34438544x + \dots + 11^{60}x^8). \end{aligned}$$

Any two of the $f_p(x)$ are completely different Galois-theoretically, implying that the two factor \widehat{M}_k each have motivic Galois group as large as possible, namely $GSpk$.

L has a functional equation with respect to $s \leftrightarrow 16 - s$. `Sign(L)` immediately returns 1, so the analytic rank r of L is even. `Evaluate(L, 8)` takes 4s and returns 0.000000000, so $r \geq 2$. `Evaluate(L, 8:Derivative:=2)` takes 14s and returns 7.851654518, so $r = 2$. The Hardy Z-function $Z(t)$ is a vertically rescaled version of $L(M, 8 + ti)$. On $[0, 7]$ it graphs out to



The double zero at $t = 0$ is visible. The next three zeros are $\gamma_1 \approx 1.93195000805$, $\gamma_2 \approx 3.00559765$, and $\gamma_3 \approx 3.61679$. Note that this calculation does not give any hints as to the desired factorization $Z(t) = Z_6(t)Z_8(t)$. In other words, we do not know which motive a given γ_i belongs to.

3 The Explicit Formula

Let M be a motive of odd weight w with L -function assumed to satisfy the Riemann hypothesis. Then its Hodge vector h , conductor N , analytic rank r , Frobenius traces $c_{p^e} = \text{Tr}(\text{Fr}_p^e | M)$, and zeros $1/2 + \gamma_k i$ in the upper half plane are related by $\log N =$

$$2\pi r \widehat{F}(0) + 4\pi \sum_k \widehat{F}(\gamma_k) + 4 \sum_{j>0} h^j \int_0^\infty \widehat{F}(t) E_j(t) dt + 2 \sum_{p^e} c_{p^e} \frac{\log p}{p^{(ew+e)/2}} F(e \log p).$$

Here F is an allowed test function, $E_j(t) = \log 2\pi - \Psi((1 + j)/2 + it)$ with $\Psi(s) = \text{Re}(\Gamma'(s)/\Gamma(s))$, and $h^{p^{-q}} = h^{p \cdot q}$.

The standard Odlyzko test function and its Fourier transform are

$$F_{\text{Od}}(x) = \chi_{[-1,1]} \left((1 - |x|) \cos(\pi x) + \frac{\sin |\pi x|}{\pi} \right), \quad \widehat{F}_{\text{Od}}(t) = \frac{4\pi \cos^2(t/2)}{(\pi^2 - t^2)^2}.$$

Also allowed are the scaled functions $F_z(x) = F_{\text{Od}}(x/\log z)$ and their Fourier transforms $\widehat{F}_z(t) = (\log z) \widehat{F}_{\text{Od}}(t \log z)$.

4 Applying the Explicit Formula to M_6 and M_8

Computing c_{p^e} for our motive M is easily done by *Magma*. However, to get the decomposition $c_{p^e} = c_{p^e}^6 + c_{p^e}^8$, even for just $e = 1$, we need to factor $f_p(x)$, which we can do only for $p \leq 11$. From the factorizations above, one has $c_3^6 = 268 \cdot 3$, $c_9^6 = (268 \cdot 3)^2 - 2(204193 \cdot 3^4)$, etc. The explicit formula using (F_{13}, \hat{F}_{13}) , with all terms divided by $\log 2$ for greater clarity, answers Questions 1 and 2:

	(Tends to 6 or 7)		(Tends to 8 or 9)		Comments
	Term ₆	Total ₆	Term ₈	Total ₈	
h	3.11142	3.11142	4.85928	4.85928	Hodge contribution
3	0.17011	3.28154	-0.63306	4.22622	Contributions from the successively harder factorizations of Frobenius polynomials $f_p(x)$
5	-0.35472	2.92682	0.07245	4.29897	
7	-0.07386	2.85296	-0.02836	4.27031	
9	-0.02269	2.83027	0.00183	4.27214	
11	0.00028	2.83055	-0.00101	4.27114	
r	2.99946	5.83002	2.99946	7.27060	Forced! A1 : (1, 1)
γ_1		5.83002	1.68061	8.95121	Forced! A2 : $(2^6, 2^9)$
γ_2	0.13610	5.96612		8.95121	Forced!
\vdots	\vdots	\vdots	\vdots	\vdots	
Total		6.00000		9.00000	

Terms are positive starting with the line beginning r , and so these terms must be associated with either M_6 or M_8 so as to keep $(total_6, total_8)$ coordinatewise less than either $(6, 9)$ or $(7, 8)$. This forces the indicated answers. Thus, both motives have analytic rank 1. The prime 2 is tamely ramified in M_6 and minimally wildly ramified in M_8 .

Remarkably, the talk just described relates directly to two collaborative projects begun at the MATRIX Institute. The decomposition studied here is the $d = 16$ case of the sequence of decompositions mentioned in numbered sentence 2 in §4 in the abstract with Rodriguez Villegas. The Hodge vectors h_6 and h_8 also arise for the L -functions denoted L_{16} and L_{18} in the abstract with Broadhurst; conductors there are $1260 = 2^2 \cdot 3^2 \cdot 5 \cdot 7$ and $7560 = 2^3 \cdot 3^3 \cdot 5 \cdot 7$ respectively.

Acknowledgements I thank the organizers and local staff for the excellent conference. My attendance at the MATRIX Institute was supported by the conference and by grant DMS-1601350 from the National Science Foundation.

References

1. Bosma, W., Cannon, J.J., Fieker, C., Steel, A. (eds.): Handbook of Magma Functions, 2.19 edn., 5291 pp. (2013)
2. Mestre, J.-F.: Formules explicites et minoration de conducteurs de de variétés algébriques. *Compos. Math.* **58**, 209–232 (1986)

Hypergeometric Supercongruences



David P. Roberts and Fernando Rodriguez Villegas

Abstract We discuss two related principles for hypergeometric supercongruences, one related to accelerated convergence and the other to the vanishing of Hodge numbers.

1 Introduction

At the conference, we added two related principles to the study of supercongruences involving the polynomials obtained by truncating hypergeometric series. By a *supercongruence* we mean a congruence which somewhat unexpectedly remains valid when the prime modulus p is replaced by p^r for some integer $r > 1$. We call r the *depth* of the supercongruence.

The first principle is that a supercongruence is the first instance of a sequence of similar supercongruences, reflecting accelerated convergence of certain Dwork quotients. The second is that splittings of underlying motives can be viewed as the conceptual source of supercongruences, with the depth of the congruence being governed by the vanishing of Hodge numbers.

We present these principles here in a limited context, so that they can be seen as clearly as possible. Let $\alpha = (\alpha_1, \dots, \alpha_d)$ be a length d vector of rational numbers in $(0, 1)$ and let $\beta = 1^d = (1, \dots, 1)$. We assume that multiplication by any integer coprime to the least common multiple m of the denominators of the α_i 's preserves the multiset $\{\alpha_1, \dots, \alpha_d\}$ modulo \mathbb{Z} .

D. P. Roberts (✉)
University of Minnesota Morris, Morris, MN, USA
e-mail: roberts@morris.umn.edu

F. Rodriguez Villegas
The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
e-mail: villegas@ictp.it

The associated classical hypergeometric series and its p -power truncations, for p prime, are as follows.

$$F(\alpha, 1^d | t) := \sum_{k=0}^{\infty} \frac{(\alpha_1)_k \cdots (\alpha_d)_k}{k!^d} t^k, \quad F_s(\alpha, 1^d | t) := \sum_{k=0}^{p^s-1} \frac{(\alpha_1)_k \cdots (\alpha_d)_k}{k!^d} t^k.$$

Our starting point was the list *CY3* of fourteen $\alpha = (\alpha_1, \dots, \alpha_4)$ associated to certain families of Calabi–Yau threefolds discussed in [8]. Each has a corresponding normalized Hecke eigenform $f = \sum a_n q^n$ of weight four and trivial character. For each, it was conjectured in [8] that

$$F_1(\alpha, 1^d | 1) \equiv a_p \pmod{p^3}, \quad p \nmid ma_p. \tag{1}$$

Some of these cases have been settled. For example, the case $\alpha = (1/5, 2/5, 3/5, 4/5)$ was proved by McCarthy [6], the corresponding modular form having level 25 [9]. Just before submitting this note, Long et al. [5] announced two different proofs of (1) for all fourteen cases in *CY3*.

2 Convergence to the Unit Root and Hodge Gaps

The two principles stem from observations about common behavior of the examples in *CY3*. The first observation is that each supercongruence (1) seems to be part of a sequence. Dwork proved [2] that for $p \nmid m$

$$\frac{F_{s+1}(\alpha, 1^d | t)}{F_s(\alpha, 1^d | t^{p^s})} \equiv \frac{F_s(\alpha, 1^d | t)}{F_{s-1}(\alpha, 1^d | t^{p^{s-1}})} \pmod{p^s}, \quad s \geq 0. \tag{2}$$

Moreover, the rational functions $F_{s+1}(\alpha, 1^d | t)/F_s(\alpha, 1^d | t^{p^s})$ converge as $s \rightarrow \infty$ to a Krasner analytic function which can be evaluated at a Teichmüller representative $\text{Teich}(\tau)$ which is not a zero of F_1 giving the *unit root* γ_p of the corresponding local L -series at p .

For $\alpha \in \text{CY3}$, computations suggest

$$\frac{F_s(\alpha, 1^d | 1)}{F_{s-1}(\alpha, 1^d | 1)} \equiv \gamma_p \pmod{p^{3s}}, \quad p \nmid ma_p, \quad s > 0, \tag{3}$$

where $\gamma_p \in \mathbb{Z}_p$ is the root of $T^2 - a_p T + p^3$ not divisible by p . Note that the case $s = 1$ reduces to (1) since $\gamma_p \equiv a_p \pmod{p^3}$.

Our second observation is that the appearance of a congruence to a power p^{3s} as opposed to the expected p^s is related to Hodge theory. Consider the hypergeometric family of motives $H(\alpha, 1^d | t)$ (see [1] for a computer implementation). For any $\tau \in \mathbb{P}^1(\mathbb{Q}) \setminus \{0, 1, \infty\}$ the motive $H(\alpha, 1^d | \tau)$ is defined over \mathbb{Q} , has rank d , weight

$d - 1$ and its only non-zero Hodge numbers are $(h^{d-1,0}, \dots, h^{0,d-1}) = (1, \dots, 1)$. When $\tau = 1$ there is a mild degeneration and the rank drops to $d - 1$.

For $\alpha \in CY3$, the motive for $\tau = 1$ is the direct sum, up to semi-simplification, of a Tate motive $\mathbb{Q}(-1)$ and the motive $A = M(f)$ of the corresponding Hecke eigenform f of weight four. The Hodge numbers of A are $(1, 0, 0, 1)$. We view the gap of three between the initial 1 and the next 1 as explaining the supercongruences (3).

3 A Congruence of Depth Five

To illustrate our two observations further, we use the decomposition established in [3, Cor. 2.1] for the case $\alpha = (1/2, 1/2, 1/2, 1/2, 1/2, 1/2)$. We learned at the conference that this example was recently studied further by Osburn et al. [7], who proved (4) below for $s = 1$ modulo p^3 and report that Mortenson conjectured it modulo p^5 .

Again after semisimplifying, the motive $H(\alpha, 1^6 | 1)$ has a distinguished summand isomorphic to the Tate motive $\mathbb{Q}(-2)$ of rank 1 and weight 4. The complement of this $\mathbb{Q}(-2)$ breaks up into two pieces A and B . They are both rank 2 motives of weight 5. Namely, $A = M(f_6)$ is the motive associated to the unique normalized eigenform $f_6 = \sum_{n \geq 1} a_n q^n$ of level 8 and weight 6 and $B = M(f_4)(-1)$ is a Tate twist of the motive associated to the unique normalized eigenform $f_4 = \sum_{n \geq 1} b_n q^n$ of level 8 and weight 4. The LMFDB [4] conveniently gives data on modular forms, including the a_n and b_n here.

The trace of Frob_p on the full rank 5 motive $H(\alpha, 1^6 | 1)$ is given by

$$a_p + b_p p + p^2.$$

Numerically, we observe the following supercongruences

$$\frac{F_s(\alpha, 1^d | 1)}{F_{s-1}(\alpha, 1^d | 1)} \equiv \gamma_p \pmod{p^{5s}}, \quad p \nmid 2a_p, \quad s \geq 1, \tag{4}$$

where $\gamma_p \in \mathbb{Z}_p$ is the root of $T^2 - a_p T + p^5$ not divisible by p .

The Hodge numbers for A and B are $(1, 0, 0, 0, 0, 1)$ and $(1, 0, 0, 1)$ respectively, with the gap of five in the Hodge numbers for A nicely matching the exponent of the supercongruences.

4 A Summarizing Conjecture

We now state a conjecture that generalizes the situations discussed so far.

Conjecture 1 For fixed $\tau = \pm 1$, let A be the unique submotive of $H(\alpha, 1^d | \tau)$ with $h^{0,d-1}(A) = 1$ and let r the smallest positive integer such that $h^{r,d-1-r}(A) = 1$. For $p \nmid m$ such that $F_1(\alpha, 1^d | \tau) \in \mathbb{Z}_p^\times$, let γ_p be the unit root of A . Then

$$\frac{F_s(\alpha, 1^d | \tau)}{F_{s-1}(\alpha, 1^d | \tau)} \equiv \gamma_p \pmod{p^{rs}}, \quad s \geq 1. \quad (5)$$

In particular, for $s = 1$ we have

$$F_1(\alpha, 1^d | \tau) \equiv a_p \pmod{p^r}, \quad (6)$$

where a_p is the trace of Frob_p acting on A .

1. For generic α, τ we expect $r = 1$ and (5) follows (see (2) and the subsequent paragraph). For the conjecture to predict $r > 1$, the motive has to split appropriately.
2. For $\alpha = (1/2, \dots, 1/2)$ and $\tau = (-1)^d$ the motive $H(\alpha, 1^d | \tau)$ acquires an involution and we expect $r = 2$ for any $d \geq 7$; all numerical evidence is consistent with this assertion.
3. For large d the unit roots involved are not in general related to classical modular forms since the motives A will typically have degrees greater than two.

Acknowledgements We thank the organizers for the wonderful conference *Hypergeometric motives and Calabi–Yau differential equations*, held at the MATRIX Institute in Creswick, Australia, in January 2017. DPR’s research is supported by grant DMS-1601350 from the National Science Foundation. FRV would like to thank the AMSI and the Australian Mathematical Society for their financial support for his participation in this conference. Special thanks go to our close collaborator M. Watkins for his continuous and significant contributions to the hypergeometric motives project in general and for the implementation of an associated package in *Magma* [1] in particular.

References

1. Bosma, W., Cannon, J.J., Fieker, C., Steel, A. (eds.): Handbook of Magma Functions, 2.19 edn., 5291 pp. (2013)
2. Dwork, B.: p-adic cycles. *Publ. Math. de l’IHÉS* **37**, 27–115 (1969)
3. Frechette, S., Ono, K., Papanikolas, M.: Gaussian hypergeometric functions and traces of Hecke operators. *Int. Math. Res. Not.* **2004**(60), 3233–3262 (2004)
4. LMFDB Collaboration: The L-functions and modular forms database (2017). <http://www.lmfdb.org>
5. Long, L., Tu, F.-T., Yui, N., Zudilin, W.: Supercongruences for rigid hypergeometric Calabi–Yau threefolds. arXiv:1705.01663v1

6. McCarthy, D.: On a supercongruence conjecture of Rodriguez-Villegas. *Proc. Am. Math. Soc.* **140**, 2241–2254 (2012)
7. Osburn, R., Straub, A., Zudilin, W.: A modular supercongruence for ${}_6F_5$: an Apéry-like story. arXiv:1701.04098v1
8. Rodriguez Villegas, F.: Hypergeometric families of Calabi–Yau manifolds. In: *Calabi–Yau Varieties and Mirror Symmetry* (Toronto, ON, 2001), pp. 223–231. Fields Institute Communications, vol. 38. American Mathematical Society, Providence (2003)
9. Schoen, C.: On the geometry of a special determinantal hypersurface associated to the Mumford–Horrocks vector bundle. *J. Reine Angew. Math.* **364**, 85–111 (1986)

Alternate Mirror Families and Hypergeometric Motives



Charles F. Doran, Tyler L. Kelly, Adriana Salerno, Steven Sperber,
John Voight, and Ursula Whitcher

Abstract Mirror symmetry predicts surprising geometric correspondences between distinct families of algebraic varieties. In some cases, these correspondences have arithmetic consequences. Among the arithmetic correspondences predicted by mirror symmetry are correspondences between point counts over finite fields, and more generally between factors of their Zeta functions. In particular, we will discuss our results on a common factor for Zeta functions of alternate families of invertible polynomials. We will also explore closed formulas for the point counts for our alternate mirror families of K3 surfaces and their relation to their Picard–Fuchs equations. Finally, we will discuss how all of this relates to hypergeometric motives. This report summarizes work from two papers.

C. F. Doran

University of Alberta, Department of Mathematics, Edmonton, AB, Canada
e-mail: doran@math.ualberta.edu

T. L. Kelly

School of Mathematics, University of Birmingham, Birmingham, UK
e-mail: t.kelly.1@bham.ac.uk

A. Salerno (✉)

Department of Mathematics, Bates College, Lewiston, ME, USA
e-mail: asalerno@bates.edu

S. Sperber

School of Mathematics, University of Minnesota, Minneapolis, MN, USA
e-mail: sperber@umn.edu

J. Voight

Department of Mathematics, Dartmouth College, Hanover, NH, USA

U. Whitcher

Mathematical Reviews, Ann Arbor, MI, USA
e-mail: uaw@umich.edu

© Springer Nature Switzerland AG 2019

D. R. Wood et al. (eds.), 2017 *MATRIX Annals*, MATRIX Book Series 2,
https://doi.org/10.1007/978-3-030-04161-8_34

1 Motivation

Calabi–Yau varieties—those smooth projective varieties with trivial canonical bundle—provide a rich and interesting source of arithmetic and geometry. Calabi–Yau varieties of dimension 1 are elliptic curves, ubiquitous in mathematics and theoretical physics. In dimensions two and above, we take our Calabi–Yau varieties to be simply connected. The two-dimensional Calabi–Yau varieties are better known as *K3 surfaces*, after the mathematicians Kummer, Kähler, and Kodaira and the mountain K2. Like elliptic curves, K3 surfaces are all diffeomorphic to each other, but the study of their complex and arithmetic structure remains deep. The study of higher dimensional Calabi–Yau varieties promise the same rewards in many areas of mathematics.

It is particularly important to study Calabi–Yau varieties in families, and interesting families of Calabi–Yau varieties arise in several ways. Perhaps the simplest method of obtaining Calabi–Yau varieties is to take smooth $(n + 1)$ -folds in projective space \mathbb{P}^n . A natural generalization of this construction is to take anticanonical hypersurfaces or complete intersections in certain toric varieties. Often, however, one wishes to consider subfamilies with further special properties. For example, a general smooth quartic in $\mathbb{P}_{\mathbb{C}}^3$ has Picard rank 1, but a general member of the pencil of K3 surfaces given by

$$x_0^4 + x_1^4 + x_2^4 + x_3^4 - 4\psi x_0 x_1 x_2 x_3 = 0$$

has Picard rank 19 and the Fermat quartic

$$x_0^4 + x_1^4 + x_2^4 + x_3^4 = 0$$

where $\psi = 0$ has Picard rank 20. As often happens, these special geometric properties are correlated with enhanced symmetry: each member of the pencil admits an action by the group $(\mathbb{Z}/4\mathbb{Z})^2$, and the Fermat quartic admits an action by a group of 384 elements [18, 19].

Calabi–Yau manifolds are also interesting from a physical perspective. Indeed, string theory posits that our universe consists of four space-time dimensions together with six extra, compact real dimensions which take the shape of a Calabi–Yau variety. Physicists have produced several consistent candidate theories, using properties of the underlying varieties. These theories are linked by *dualities* which transform physical observables described by one collection of geometric data into equivalent observables described by different geometric data. Attempts to build a mathematically consistent description of the duality between Type IIA and Type IIB string theories led to the thriving field of *mirror symmetry*, which is based on the philosophy that the complex moduli of a given family of Calabi–Yau varieties should correspond to the complexified Kähler moduli of a mirror family.

There are several methods of constructing the mirror to a family of Calabi–Yau varieties. The first mirror symmetry construction, due to Greene–Plesser [15], used a $(\mathbb{Z}/5\mathbb{Z})^3$ action on the one-parameter family of quintic threefold X_ψ given by

$$x_0^5 + x_1^5 + x_2^5 + x_3^5 + x_4^5 - 5\psi x_0 x_1 x_2 x_3 x_4 = 0$$

to construct the mirror family Y_ψ to all smooth quintic hypersurfaces in \mathbb{P}^4 . A directly analogous construction can be used to find the mirrors to families of Calabi–Yau hypersurfaces in weighted projective spaces. Batyrev gave combinatorial methods for constructing mirror families to Calabi–Yau varieties realized as hypersurfaces or complete intersection in toric varieties [1]. Though powerful, Batyrev’s construction relates families rather than individual varieties. In the current work, we use an alternative generalization of the Greene–Plesser construction due to Berglund–Hübsch–Krawitz [2, 17], allowing for a direct comparison of varieties on either side of the mirror correspondence.

When individual pairs of mirror varieties can be identified, mirror symmetry constructions have implications for their arithmetic and geometric structure. These implications were first explored by Candelas–de la Ossa–Rodriguez-Villegas [4] for their zeta functions, the generating function for the number of \mathbb{F}_{p^r} -valued points

$$Z(X, T) = \exp \left(\sum_{r=1}^{\infty} \frac{\#X(\mathbb{F}_{p^r})T^r}{r} \right)$$

for a variety X over \mathbb{F}_p ; we have $Z(X, T) \in \mathbb{Q}(T)$ by a theorem of Dwork [10]. These authors used the Greene–Plesser mirror construction and techniques from toric varieties to compare the zeta function of fibers of the diagonal Fermat pencil of threefold X_ψ and the mirror pencil of threefold Y_ψ [3–5]. They found that for general ψ , the zeta functions of X_ψ and Y_ψ share a common factor $R(T, \psi)$. This common factor is related to the period of the holomorphic form on X_ψ , and the number of points on X_ψ over a finite field is given by a truncation of a generalized hypergeometric function which solves the Picard–Fuchs equation associated to the holomorphic form. Furthermore, the other nontrivial factors of $Z(X_\psi, T)$ were closely related to the action of $(\mathbb{Z}/5\mathbb{Z})^3$ on homogeneous monomials.

The Greene–Plesser construction generalizes easily to smooth hypersurfaces of degree $n + 1$ in \mathbb{P}^n . Wan [20] has characterized the relationship between a member X_ψ of the diagonal Fermat pencil in \mathbb{P}^n and its mirror Y_ψ in terms of point counts via the congruence

$$\#X_\psi(\mathbb{F}_q) \equiv \#Y_\psi(\mathbb{F}_q) \pmod{q}$$

for all $q = p^r$ such that $\mathbb{F}_q \supseteq \mathbb{F}_p(\psi)$. Fu–Wan [12] generalized this result to other pairs of mirror pencils. More recently, Kloosterman [16] showed that one can use a group action to describe the distinct factors of the zeta function for

any one-parameter monomial deformation of a diagonal hypersurface in weighted projective space.

In our work, we take a slightly different approach. Rather than relating a pencil of Calabi–Yau varieties to its mirror, we instead consider those pencils whose mirrors are related in some geometric way. In other words, we seek to understand when common properties of mirrors translate into arithmetic, geometric, or physical implications for the original pencils themselves.

There is an intricate relationship between Picard–Fuchs equations and the zeta function, mediated by the action of the Frobenius map. Given a set of symmetric pencils in \mathbb{P}^n which yield alternate mirrors to smooth $n + 1$ -folds in \mathbb{P}^n , we hypothesize that the zeta functions of the members of each pencil and their mirror should share a common factor, corresponding to the Picard–Fuchs equation satisfied by the holomorphic form. In the current work, we apply the formalism of Berglund–Hübsch–Krawitz mirror symmetry to characterize appropriate symmetric pencils, and we study the resulting zeta functions.

We have followed four approaches, exploiting algebraic, geometric, and arithmetic properties of highly symmetric pencils.

2 Common Factor Theorem

Our first result, described in more detail in [8], is that invertible pencils whose mirrors have common properties share arithmetic similarities as well. Revisiting work of Gähns [14], we find that invertible pencils whose BHK mirrors are hypersurfaces in quotients of the same weighted-projective space have the same Picard–Fuchs equation associated to their holomorphic form. In turn, we show that the Picard–Fuchs equations for the pencil dictate a factor of the zeta functions of the pencil.

An *invertible polynomial* is a polynomial

$$F_A = \sum_{i=0}^n \prod_{j=0}^n x_j^{a_{ij}} \in \mathbb{Z}[x_0, \dots, x_n],$$

where $A = (a_{ij})_{i,j}$ is an $(n + 1) \times (n + 1)$ is a matrix with nonnegative integer entries, such that:

- $\det(A) \neq 0$,
- the polynomial F_A is homogeneous of degree $n + 1$, and
- the function $F_A : \mathbb{C}^{n+1} \rightarrow \mathbb{C}$ has exactly one singular point at the origin.

We further impose that these hypersurfaces are Calabi–Yau varieties, so the degree of the polynomial F_A is $n + 1$.

Inspired by Berglund–Hübsch–Krawitz (BHK) mirror symmetry, we look at the weights of the transposed polynomial

$$F_{A^T} := \sum_{i=0}^n \prod_{j=0}^n x_j^{a_{ji}},$$

which will be a quasihomogeneous polynomial, i.e., there exist nonnegative integral weights q_0, \dots, q_n so that $\gcd(q_0, \dots, q_n) = 1$ and F_{A^T} defines a hypersurface X_{A^T} in the weighted-projective space $W\mathbb{P}^n(q_0, \dots, q_n)$. We call q_0, \dots, q_n the *dual weights* of F_A . Let $d^T = \sum_i q_i$ be the sum of the weights.

Using the dual weights, we define a one-parameter deformation of our invertible pencil. Consider the polynomials

$$F_{A,\psi} = \sum_{i=0}^n \prod_{j=0}^n x_j^{a_{ij}} - d^T \psi x_0 \cdots x_n \in \mathbb{Z}[\psi][x_0, \dots, x_n].$$

We then have a family of hypersurfaces $X_{A,\psi} := Z(F_{A,\psi}) \subset \mathbb{P}^n$ in the parameter ψ , which we call an *invertible pencil*.

The Picard–Fuchs equation for the family $X_{A,\psi}$ is determined completely by the dual weights by work of Gährs [14, Theorem 3.6]. Indeed, Gährs computes the order of the Picard–Fuchs equation in terms of the q_i . There is an explicit formula for the order $D(\mathbf{q})$ of the Picard–Fuchs equation that depends solely on the $(n + 1)$ -tuple of dual weights $\mathbf{q} = (q_0, \dots, q_n)$. The Picard–Fuchs equation itself depends solely on \mathbf{q} as well. To be precise, we observe that the Picard–Fuchs equation is a hypergeometric differential equation whose motive descends to \mathbb{Q} .

For a smooth projective hypersurface X in \mathbb{P}^n , the zeta function is of the form

$$Z(X, T) = \frac{P_X(T)^{(-1)^n}}{(1 - T)(1 - qT) \cdots (1 - q^{n-1}T)},$$

with $P_X(T) \in \mathbb{Q}[T]$. Our main result exhibits a (fiber-wise) common factor of the zeta function in the general setting suggested above.

Theorem 1 *Let $X_{A,\psi}$ and $X_{B,\psi}$ be invertible pencils of Calabi–Yau $n - 1$ -folds in \mathbb{P}^n , determined by integer matrices A and B , respectively. Suppose A and B have the same dual weights q_i . Then for each $\psi \in \mathbb{F}_q$ such that the fibers $X_{A,\psi}$ and $X_{B,\psi}$ are smooth and $\gcd(q, (n + 1)d^T) = 1$, the polynomials $P_{X_{A,\psi}}(T)$ and $P_{X_{B,\psi}}(T)$ have a common factor $R_\psi(T) \in \mathbb{Q}[T]$ with $\deg R_\psi(T) \geq D(\mathbf{q})$.*

3 Explicit Computations and Hypergeometric Motives

We next focus our attention on invertible families of K3 surfaces, with dual weights $(1, 1, 1, 1)$, which are as follows:

In [9] we analyze the zeta functions of the families given in Table 1. Using a classical viewpoint, we find that the hypergeometricity of the Picard–Fuchs equations associated to the five families predicts a motivic decomposition of the point counts over finite fields for our families. We see that the hypergeometric Picard–Fuchs equations for the primitive middle cohomology of the five families correspond to nontrivial hypergeometric summands in the point counts over finite fields. The core of this paper is the following theorem:

Theorem 2 *Let $\diamond \in \mathcal{F} = \{F_4, F_2L_2, F_1L_3, L_2L_2, L_4\}$ signify one of the five K3 families in Table 1. There is a canonical decomposition of the finite field point count for $N_{\mathbb{F}_q}(X_{\diamond, \psi})$ whose summands are either trivial or hypergeometric. Moreover, there exists an element in $H^2_{\text{prim}}(X_{\diamond, \psi})$ that satisfies a hypergeometric Picard–Fuchs differential equation with parameters $\alpha_1, \dots, \alpha_n; \beta_1, \dots, \beta_{n-1}$ if and only if there exists a nontrivial summand in the canonical finite field point count $N_{\mathbb{F}_q}(X_{\diamond, \psi})$ corresponding to the hypergeometric function defined over \mathbb{F}_q with parameters $\alpha_1, \dots, \alpha_n; \beta_1, \dots, \beta_{n-1}$.*

This proof is done explicitly. First, we find the Picard–Fuchs equations via the diagrammatic method introduced in [4, 5] and fully developed in [7]. After establishing the hypergeometric forms of the Picard–Fuchs equations, we confirm that they do indeed correspond to those in the finite point counts using Gauss sums, using a classical method due to Delsarte [6] and Furtado Gomida [13].

Additionally, we obtain finer information by factoring the polynomial $Q_{\diamond, \psi}(T)$ in Theorem 1 further, giving a complete hypergeometric decomposition. Our result is as follows.

Table 1 The symmetric quartic K3 pencils with dual weights $(1,1,1,1)$

	Quartic Family	Symmetries
F_4	$x_0^4 + x_1^4 + x_2^4 + x_3^4 - 4\psi x_0 x_1 x_2 x_3 = 0$	$(\mathbb{Z}/4\mathbb{Z})^2$
F_1L_3	$x_0^4 + x_1^3 x_2 + x_2^3 x_3 + x_3^3 x_1 - 4\psi x_0 x_1 x_2 x_3 = 0$	$\mathbb{Z}/7\mathbb{Z}$
F_2L_2	$x_0^4 + x_1^4 + x_2^3 x_3 + x_3^3 x_2 - 4\psi x_0 x_1 x_2 x_3 = 0$	$\mathbb{Z}/8\mathbb{Z}$
L_2L_2	$x_0^3 x_1 + x_1^3 x_0 + x_2^3 x_3 + x_3^3 x_2 - 4\psi x_0 x_1 x_2 x_3 = 0$	$\mathbb{Z}/4\mathbb{Z}$
L_4	$x_0^3 x_1 + x_1^3 x_2 + x_2^3 x_3 + x_3^3 x_0 - 4\psi x_0 x_1 x_2 x_3 = 0$	$\mathbb{Z}/5\mathbb{Z}$

Corollary 1 *The polynomials $Q_{\diamond, \psi, q}(T)$ factor over $\mathbb{Z}[T]$ according to the following table.*

Family	Factorization	Hypothesis	r
F_4	$(\deg 1)^{12}(\deg 2)^3$	$q \equiv 1 \pmod 4$	2
F_2L_2	$(1 - qT)^6(\deg 1)^2(\deg 2)^5$	$q \equiv 1 \pmod 8$	4
F_1L_3	$(\deg 6)^3$	$q \equiv 1 \pmod{28}$	12
L_2L_2	$(1 - qT)^8(\deg 2)^1(\deg 4)^2$	$q \equiv 1 \pmod 4$	2
L_4	$(1 - qT)^2(\deg 4)^4$	$q \equiv 1 \pmod{20}$	10

(1)

In the above table, there may be further factorization depending on ψ and q , and some of these factors may agree. The integer r in above table is such that for $q = p^r$, we have $Q_{\diamond, \psi, q}(T) = (1 - qT)^{18}$ under the hypotheses of Theorem A: in other words, if we factor $Q_{\diamond, \psi, q}(T)$ as a product of cyclotomic polynomials ϕ_{m_i} , then $\text{lcm}(m_i) \mid r$.

The case of the Dwork pencil F_4 is due to Dwork [11, §6j, p. 73], and in this case we know that the degree 2 factor occurs with multiplicity 3 and the linear factor occurs with multiplicity 12, as the notation indicates. The factorization in Corollary 1 is motivated by similar work due to Candelas–de la Ossa–Rodriguez-Villegas [4, 5]. (Kloosterman [16] has shown that one can use a group action to describe the distinct factors of the zeta function for any one-parameter monomial deformation of a diagonal hypersurface in weighted projective space; only one of our families, the Dwork pencil, fits within the scope of this work.)

Acknowledgements The authors heartily thank Xenia de la Ossa for her input and many discussions about this project. They thank Simon Judes for sharing his expertise, Frits Beukers for numerous helpful discussions, and Edgar Costa for sharing his code for computing zeta functions. The authors would like to thank the American Institute of Mathematics and its SQuaRE program, the Banff International Research Station, SageMath, and the MATRIX Institute for facilitating their work together. Kelly acknowledges that this material is based upon work supported by the NSF under Award No. DMS-1401446 and the EPSRC under EP/N004922/1. Voight was supported by an NSF CAREER Award (DMS-1151047).

References

1. Batyrev, V.V.: Dual polyhedra and mirror symmetry for Calabi-Yau hypersurfaces in toric varieties. *J. Algebraic Geom.* **3**(3), 493–535 (1994)
2. Berglund, P., Hübsch, T.: A generalized construction of mirror manifolds. *Nuclear Phys. B* **393**(1–2), 377–391 (1993)
3. Candelas, P., de la Ossa, X.: The Zeta-Function of a p-adic Manifold, Dwork Theory for Physicists. arxiv:0705.2056v1 (2008)
4. Candelas, P., de la Ossa, X., Rodriguez Villegas, F.: Calabi–Yau manifolds over finite fields, I. arXiv:hep-th/0012233v1 (2000)
5. Candelas, P., de la Ossa, X., Rodriguez-Villegas, F.: Calabi–Yau manifolds over finite fields II. In: *Calabi–Yau Varieties and Mirror Symmetry*, Toronto, pp. 121–157, (2001). hep-th/0402133

6. Delsarte, J.: Nombre de solutions des équations polynomiales sur un corps fini. *Sém. Bourbaki* **39**(1), 321–329 (1951)
7. Doran, C.F., Greene, B., Judes, S.: Families of quintic Calabi–Yau 3-folds with discrete symmetries. *Commun. Math. Phys.* **280**, 675–725 (2008)
8. Doran, C.F., Kelly, T., Salerno, A., Sperber, S., Voight, J., Whitcher, U.: Zeta functions of alternate mirror Calabi–Yau families. *Israel J. Math.* **228**(2), 665–705 (2018)
9. Doran, C.F., Kelly, T., Salerno, A., Sperber, S., Voight, J., Whitcher, U.: Hypergeometric decomposition of symmetric K3 quartic pencils. arXiv:1810.06254
10. Dwork, B.: On the rationality of the zeta function of an algebraic variety. *Am. J. Math.* **82**(3), 631–648 (1960)
11. Dwork, B.: p -adic cycles. *Inst. Hautes Études Sci. Publ. Math.* **37**, 27–115 (1969)
12. Fu, L., Wan, D.: Mirror congruence for rational points on Calabi–Yau varieties. *Asian J. Math.* **10**(1), 1–10 (2006)
13. Furtado Gomida, E.: On the theorem of Artin–Weil. *Soc. Mat. São Paulo* **4**, 267–277 (1951)
14. Gähns, S.: Picard–Fuchs equations of special one-parameter families of invertible polynomials, Ph.D. thesis, Gottfried Wilhelm Leibniz Univ. Hannover, arXiv:1109.3462
15. Greene, B.R., Plesser, M.: Duality in Calabi–Yau moduli space. *Nuclear Phys. B* **338**(1), 15–37 (1990)
16. Kloosterman, R.: The zeta function of monomial deformations of Fermat hypersurfaces. In: *Algebra Number Theory*, vol.1, no. 4. Mathematical Science Publishers, Berkeley (2007)
17. Krawitz, M.: FJRW rings and Landau–Ginzburg Mirror Symmetry. arxiv:0906.0796 (2009)
18. Mukai, S.: Finite groups of automorphisms of K3 surfaces and the Mathieu group. *Invent. Math.* **94**(1), 183–221 (1988)
19. Oguiso, K.: A characterization of the Fermat quartic K3 surface by means of finite symmetries. *Compos. Math.* **141**(2), 404–424 (2005)
20. Wan, D.: Mirror symmetry for zeta functions. In: *Mirror Symmetry. V*, vol. 38. AMS/IP Studies in Advanced Mathematics (2006)

Schwarzian Equations and Equivariant Functions



Abdellah Sebbar

Abstract In this review article we show how the theory of Schwarzian differential equations leads to an interesting class of meromorphic functions on the upper-half plane \mathbb{H} named equivariant functions. These functions have the property that their Schwarz derivatives are weight 4 automorphic forms for a discrete subgroup Γ of $\mathrm{PSL}_2(\mathbb{R})$. It turns out that these functions must satisfy the relation

$$f(\gamma\tau) = \rho(\gamma)f(\tau), \quad \tau \in \mathbb{H}, \quad \gamma \in \Gamma,$$

where ρ is a 2-dimensional complex representation of Γ and the matrix action on both sides is by linear fractional transformation. When ρ is the identity representation $\rho(\gamma) = \gamma$, the equivariant functions are parameterized by scalar automorphic forms, while if ρ is an arbitrary representation they are parameterized by vector-valued automorphic forms with multiplier ρ . If Γ is a modular subgroup we obtain important applications to modular forms for Γ as well as a description in terms of elliptic functions theory. We also prove the existence of equivariant functions for the most general case by constructing a vector bundle attached to the data (Γ, ρ) and applying the Kodaira vanishing theorem.

1 The Schwarz Derivative

Let D be a domain in \mathbb{C} and f a meromorphic function on D . The Schwarz derivative or the Schwarzian of f is defined by

$$\{f, z\} = \left(\frac{f''}{f'}\right)' - \frac{1}{2}\left(\frac{f''}{f'}\right)^2 = \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2.$$

A. Sebbar (✉)

Department of Mathematics and Statistics, University of Ottawa, Ottawa, ON, Canada

e-mail: asebbar@uottawa.ca

It was named after Schwarz by Cayley, however, Schwarz himself pointed out that it was discovered by Lagrange in 1781 [10]. The Schwarz derivative has many interesting properties which are given below. The functions involved are meromorphic functions on a domain D .

- Projective invariance:

$$\left\{ \frac{af + b}{cf + d}, z \right\} = \{f, z\}, \quad a, b, c, d \in \mathbb{C}, \quad ad - bc \neq 0.$$

- Cocycle property: If w is a function of z , then

$$\{f, z\} = \{f, w\}(dw/dz)^2 + \{w, z\}.$$

- $\{f, z\} = 0$ if and only if $f(z) = \frac{az + b}{cz + d}$ for some $a, b, c, d \in \mathbb{C}$.
- If $w = \frac{az + b}{cz + d}$ with $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \text{GL}_2(\mathbb{C})$, then

$$\{f, z\} = \{f, w\} \frac{(ad - bc)^2}{(cz + d)^4}.$$

- For two meromorphic functions f and g on D ,

$$\{f, z\} = \{g, z\} \text{ if and only if } f(z) = \frac{ag(z) + b}{cg(z) + d}, \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \text{GL}_2(\mathbb{C})$$

- If $w(z)$ is a function of z with $w'(z_0) \neq 0$ for some $z_0 \in D$, then in a neighborhood of z_0 , we have

$$\{z, w\} = \{w, z\} (dz/dw)^2.$$

Some of the properties are elementary and the rest follows from the following important connection with the theory of ordinary differential equations:

Let $R(z)$ be a meromorphic function on D and consider the second order differential equation

$$y'' + \frac{1}{2}R(z)y = 0$$

with two linearly independent solutions y_1 and y_2 . Then $f = y_1/y_2$ is a solution to the Schwarz differential equation

$$\{f, z\} = R(z).$$

Conversely, if $f(z)$ is locally univalent and $\{f, z\} = R(z)$, then $y_1 = f/\sqrt{f'}$ and $y_2 = 1/\sqrt{f'}$ are two linearly independent solutions to $y'' + \frac{1}{2}R(z)y = 0$.

The Schwarz derivative plays an important role in the study of the complex projective line, univalent functions, conformal mapping, Teichmüller spaces and most importantly in the theory of modular forms and hypergeometric functions [1, 4, 6, 8, 9, 11].

We now look at the effect of the Schwarz derivative on automorphic functions for a discrete subgroup Γ of $PSL_2(\mathbb{Z})$, that is a Fuchsian group of the first kind acting on the upper half-plane $\mathbb{H} = \{\tau \in \mathbb{C} \mid \Im(\tau) > 0\}$ by linear fractional transformation

$$\gamma\tau = \frac{a\tau + b}{c\tau + d}, \quad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma.$$

Proposition 1.1 ([8]) *If f is an automorphic function for a discrete group Γ , then $\{f, \tau\}$ is a weight 4 automorphic form for Γ that is holomorphic everywhere except at the points where f has a multiple zero or a multiple pole (including at the cusps). Moreover, if Γ is of genus zero and f is a Hauptmodul, then $\{f, \tau\}$ is modular for the normalizer of Γ in $PSL_2(\mathbb{R})$*

As an example, let λ be the Klein modular function for $\Gamma(2)$ given by

$$\lambda(\tau) = \left(\frac{\eta(\tau/2)}{\eta(2\tau)} \right)^8,$$

where η is the Dedekind eta-function given by

$$\eta(\tau) = q^{\frac{1}{24}} \prod_{n \geq 1} (1 - q^n), \quad q = \exp(2\pi i \tau),$$

then

$$\{\lambda, \tau\} = \frac{\pi^2}{2} E_4(\tau),$$

where E_4 is the weight 4 Eisenstein series

$$E_4(\tau) = 1 + 240 \sum_{n \geq 1} \sigma_3(n)q^n,$$

with $\sigma_k(n)$ being the sum of the k -th powers of the positive divisors of n .

If $\Gamma = \Gamma_0(8)$ and we consider the Hauptmodul f_8 for Γ given by

$$f_8(\tau) = \frac{\eta(4\tau)^{12}}{\eta(2\tau)^4\eta(8\tau)^8}$$

then

$$\frac{1}{2\pi^2}\{f_8, \tau\} = \frac{1}{4}(\theta_3^4 + \theta_4^4)^2 = \theta_{D_4 \oplus D_4}(2\tau),$$

where $\theta_{D_4 \oplus D_4}(2\tau)$ is the theta function of two copies of the root lattice D_4 and θ_3 and θ_4 are the Jacobi theta-functions

$$\theta_3(\tau) = \sum_{n \in \mathbb{Z}} q^{\frac{1}{2}n^2}, \quad \theta_4(\tau) = \sum_{n \in \mathbb{Z}} (-1)^n q^{\frac{1}{2}n^2}.$$

For later use, we also give

$$\theta_2(\tau) = \sum_{n \in \mathbb{Z}} q^{\frac{1}{2}(n+1/2)^2}.$$

Finally, if $\Gamma = \Gamma_1(5)$ and f_5 is the Hauptmodul given by

$$f_5(\tau) = q \prod_{n \geq 1} (1 - q^n)^{5\left(\frac{n}{5}\right)}$$

where $\left(\frac{n}{5}\right)$ is the Legendre symbol, then

$$\frac{1}{2\pi^2}\{f_5, \tau\} = \theta_{Q_8(1)}$$

where $Q_8(1)$ is the Icosian or Maass lattice which is the 8-dimensional 5-unimodular lattice with determinant 625 and minimal norm 4.

Notice that in the above three examples, the Schwarz derivatives are all holomorphic as the groups involved are torsion-free and thus their Hauptmoduls do not have multiple zeros or poles.

One may ask if the converse of the above properties is true: Suppose that the Schwarz derivative $\{f, \tau\}$ of a meromorphic function on \mathbb{H} is a weight 4 automorphic, what can be said about f ? Does it have any automorphic properties? The following sections will be devoted to elucidate this question.

2 Equivariant Functions

Suppose that F is a weight 4 automorphic form for a discrete group Γ and f is a meromorphic function on \mathbb{H} such that $\{f, \tau\} = F(\tau)$. Then using the properties of the Schwarz derivative from the previous section we have, for $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$,

$$\begin{aligned} (c\tau + d)^4 F(\tau) &= F\left(\frac{a\tau + b}{c\tau + d}\right) \\ &= \left\{ f\left(\frac{a\tau + b}{c\tau + d}\right), \frac{a\tau + b}{c\tau + d} \right\} \\ &= (c\tau + d)^4 \left\{ f\left(\frac{a\tau + b}{c\tau + d}\right), \tau \right\}. \end{aligned}$$

Therefore,

$$\{f, \tau\} = \left\{ f\left(\frac{a\tau + b}{c\tau + d}\right), \tau \right\}.$$

Hence, there exists $\begin{pmatrix} A & B \\ C & D \end{pmatrix} \in \text{GL}_2(\mathbb{C})$ such that

$$f\left(\frac{a\tau + b}{c\tau + d}\right) = \frac{Af(\tau) + B}{Cf(\tau) + D}.$$

This defines a 2-dimensional representation ρ of Γ in $\text{GL}_2(\mathbb{C})$ such that

$$f(\gamma\tau) = \rho(\gamma) f(\tau) \tag{1}$$

where on both sides the action of the matrices is by linear fractional transformation.

We will distinguish three cases:

1. $\rho = 1$ a constant, in which case f is an automorphic function.
2. $\rho = \text{Id}$, the embedding of Γ in $\text{GL}_2(\mathbb{C})$ or the defining representation of Γ , providing a meromorphic function commuting with the action of Γ which we call an equivariant function for Γ .
3. ρ is a general representation not equal to one in the above cases giving a function f called a ρ -equivariant function for Γ .

We will be interested in the last two cases. A trivial example of an equivariant function for a discrete group Γ is $f(\tau) = \tau$. We will see in the next section that there are infinitely many examples parametrized by automorphic forms for Γ . Furthermore, the set of equivariant functions will have a structure of an infinite dimensional vector space isomorphic to the space of meromorphic sections of the

canonical bundle of the compact Riemann surface $X(\Gamma) = (\Gamma \backslash \mathbb{H})^*$ where the star indicates that we have added the cusps to the quotient space, in other words, the space of meromorphic differential forms on $X(\Gamma)$.

In the general case, we establish the existence of ρ -equivariant functions for an arbitrary representation ρ of Γ . This will include the case when $\rho : \Gamma \rightarrow \mathbb{C}^*$ is a character. Of course, when this character is unitary, then we recover the classical automorphic functions with a character.

3 The Automorphic and Modular Aspects

In this section we focus solely on the equivariant functions, that is when ρ is the defining representation. We have already seen that $f(\tau) = \tau$ is equivariant for every discrete group. It turns out that there are many more nontrivial equivariant functions.

Theorem 3.1 ([5]) *Let Γ be a discrete group. We have*

1. *Let f a nonzero automorphic form of weight k for Γ (even with a character), then*

$$h_f(\tau) := \tau + k \frac{f(\tau)}{f'(\tau)} \tag{2}$$

is an equivariant function for Γ .

2. *Let h be an equivariant function for Γ . Then $h(\tau) = h_f(\tau)$ for some automorphic form with a character for Γ if and only if the poles of $1/(h(\tau) - \tau)$ are simple with rational residues. Moreover, if Γ has genus 0, then we can omit the character from this statement.*

If k is a nonzero integer and c is a nonzero constant, then $h_f = h_{fk} = h_{cf}$ and so the correspondence $f \mapsto h_f$ is not one-to-one. Because of the second part of the theorem, an equivariant functions that arises from an automorphic form as in (2) is called a rational equivariant form. It turns out that for such an equivariant function $h(\tau)$, the residues of $1/(h(\tau) - \tau)$ have bounded denominators, and any common multiple of these denominators can be the weight for an automorphic form f such that $h = h_f$. Moreover, not all equivariant functions are rational. An example of a non-rational equivariant function is given by

$$h(\tau) = \tau + 4 \frac{E_4(\tau)}{E_4'(\tau) + E_6(\tau)},$$

where E_6 is the weight six Eisenstein series

$$E_6(\tau) = 1 - 504 \sum_{n \geq 1} \sigma_5(n)q^n .$$

Indeed, one can show that $1/(h(\tau) - \tau)$ has a simple pole at the cubic root of unity, but with residue $\frac{1}{4} + \frac{\pi i}{6}$.

The following theorem provides two important applications of equivariant functions to modular forms.

Theorem 3.2 ([12]) *Let f be a modular form for a finite index subgroup of $SL_2(\mathbb{Z})$ of nonzero weight, then*

1. *The derivative of f has infinitely many non-equivalent zeros in \mathbb{H} , all but a finite number are simple zeros.*
2. *The q -expansion of f , where q is the uniformizer at ∞ for Γ , cannot have only a finite number of nonzero coefficients.*

This theorem follows from the properties of the equivariant function h_f attached to f as in (2); the most important of which is that h_f takes always real values. This is clear if h_f has a pole in \mathbb{H} as it will take rational values at the orbit of this pole, but if h_f is holomorphic, then we have to apply the theorem of Denjoie-Wolfe applied to the iterates of h_f . Then we prove that h_f has infinitely many non-equivalent poles in \mathbb{H} . To prove that the zeros are all simple except for a finite number of them requires the use the Rankin-Cohen brackets. The second statement is usually proven using the L -function of the modular form, but here it is a simple consequence of the first statement.

We end this section with an interesting connection with the cross-ratio which is defined for four distinct complex numbers $z_i, 1 \leq i \leq 4$ by

$$[z_1, z_2, z_3, z_4] = \frac{(z_1 - z_2)(z_4 - z_3)}{(z_1 - z_3)(z_4 - z_2)}.$$

As it is projectively invariant, the cross-ratio of four distinct equivariant functions for a discrete group Γ is an automorphic function for Γ . As examples, we have

$$[\tau, h_{\theta_2}, h_{\theta_3}, h_{\theta_4}] = \lambda,$$

and

$$[\tau, h_{E_4}, h_{\Delta}, h_{E_6}] = \frac{1}{1728} j,$$

where $\Delta = \eta^{24}$ is the discriminant cusp form and the Dedekind j -function is given by $j = E_4^3/\Delta$.

4 The Elliptic Aspect

The ideas in this section first started in [3] and were developed further in [15] and more recently in full generalization in [2]. Let $L = \omega_1\mathbb{Z} + \omega_2\mathbb{Z}$ be a lattice in \mathbb{C} with $\Im\omega_2/\omega_1 > 0$. Its Weierstrass \wp -function is given by

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega \in L \setminus \{0\}} \left(\frac{1}{(z - \omega)^2} - \frac{1}{\omega^2} \right),$$

and the Weierstrass ζ -function is given by

$$\zeta(z) = \frac{1}{z} + \sum_{\omega \in L \setminus \{0\}} \left(\frac{1}{z - \omega} + \frac{1}{\omega} + \frac{z}{\omega^2} \right).$$

Notice that $\zeta'(z) = -\wp(z)$, and while \wp is L -periodic, ζ is quasi-periodic with respect to L in the sense that for $\omega \in L$ and $z \in \mathbb{C}$, we have

$$\zeta(z + \omega) = \zeta(z) + H_L(\omega)$$

where the quasi-period map depends on the Lattice L . It is \mathbb{Z} -linear and so it is determined by the quasi-periods $\eta_1 = H_L(\omega_1)$ and $\eta_2 = H_L(\omega_2)$. Moreover, H_L is homogeneous of weight -1 in the sense that if $\alpha \in \mathbb{C}^\times$, then

$$H_{\alpha L}(\alpha\omega) = \alpha^{-1}H_L(\omega).$$

The quasi-periods satisfy the Legendre relation

$$\omega_1\eta_2 - \omega_2\eta_1 = 2\pi i.$$

We now suppose that $\omega_1 = 1$ and $\omega_2 = \tau \in \mathbb{H}$. Using the fact that $SL_2(\mathbb{Z})$ acts on L by isomorphisms (by a change of basis) and using the homogeneity of the quasi period map H_L , it is easy to see that

$$h_0(\tau) = \frac{\eta_2}{\eta_1}$$

is equivariant for $SL_2(\mathbb{Z})$ [3].

In fact, from the expression of the Weierstrass ζ -function one can prove that

$$\eta_1 = \frac{\pi^2}{3} E_2(\tau)$$

where E_2 is the weight 2 Eisenstein series

$$E_2(\tau) = 1 - 24 \sum_{n \geq 1} \sigma_1(n)q^n = \frac{1}{2\pi i} \frac{\Delta'(\tau)}{\Delta(\tau)}.$$

Therefore, using the Legendre relation, we get

$$h_0(\tau) = \tau + \frac{6}{i\pi E_2(\tau)} = \tau + 12 \frac{\Delta(\tau)}{\Delta'(\tau)},$$

and thus $h_0 = h_\Delta$ is a rational equivariant function.

Let us put

$$M_\tau = \begin{pmatrix} \tau & \eta_2 \\ 1 & \eta_1 \end{pmatrix}.$$

Then M_τ is invertible as $\det M_\tau = -2\pi i$ by the Legendre relation.

Let Γ be a finite index subgroup of $SL_2(\mathbb{Z})$ and denote by $Eq(\Gamma)$ the set of all equivariant functions for $SL_2(\mathbb{Z})$ excluding the trivial one $h(\tau) = \tau$. Also denote by $M_2(\Gamma)$ be the set of all weight two meromorphic modular forms for $SL_2(\mathbb{Z})$. We have

Theorem 4.1 *The map from $M_2(\Gamma)$ to $Eq(\Gamma)$*

$$f \mapsto M_\tau f$$

is a bijection where $M_\tau f$ is the linear fraction of f given by M_τ .

The above map sends the zero modular form to h_0 which is equivariant for $SL_2(\mathbb{Z})$ and hence for every subgroup. In the meantime, h_0 was built using the quasi-periods of the Weierstrass ζ -function. One might ask: what about the remaining equivariant functions? can they arise also from elliptic objects in the same way h_0 does? In the paper [2], this question is fully answered, and indeed for each equivariant function for Γ , one can construct a generalizations of the Weierstrass ζ -function called elliptic zeta functions which are quasi-periodic maps on the set of lattices such the quotient of two fundamental quasi-periods is an equivariant function. The interesting aspect is that there is a triangular commutative correspondence between the set of these elliptic zeta functions, $M_2(\Gamma)$ and $Eq(\Gamma)$ which encompasses the modular and elliptic nature of equivariant functions.

As for the geometric aspect, because the weight 2 meromorphic modular forms are identified with the meromorphic differential forms on the Riemann surface $X(\Gamma)$, we can thus view the equivariant functions as the global meromorphic sections of the canonical bundle of $X(\Gamma)$.

5 The General Case

In this section, we consider the case of a general discrete group and an arbitrary representation $\rho : \Gamma \rightarrow \text{GL}_2(\mathbb{C})$ and investigate the existence of ρ -equivariant functions for Γ , that is, the meromorphic functions on \mathbb{H} such that

$$f(\gamma\tau) = \rho(\gamma)f(\tau).$$

We will denote the set of such functions by $Eq(\Gamma, \rho)$. Let us first recall the notion of vector-valued automorphic forms for the data (Γ, ρ) . A meromorphic function $F = (f_1, f_2)^t : \mathbb{H} \rightarrow \mathbb{C}^2$ where f_1 and f_2 are two meromorphic functions on \mathbb{H} is called a 2-dimensional vector-valued automorphic form for Γ of multiplier ρ and weight $k \in \mathbb{Z}$ if

$$(c\tau + d)^{-k} F(\gamma\tau) = \rho(\gamma) F(\tau), \quad \tau \in \mathbb{H}, \quad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma,$$

in addition to the usual growth behavior at the cusps. Denote by $V_k(\Gamma, \rho)$ the space of all such forms. They were fairly studied in the last two decades by various authors in different contexts from algebraic, arithmetic, analytic, geometric and theoretical physics points of view, see [14] and the extensive list of references therein. Their existence is well established in the literature for a unitary representation ρ and for Γ being a subgroup of $\text{SL}_2(\mathbb{Z})$ or a genus zero discrete group among other cases. The existence for an arbitrary data (Γ, ρ) has been recently proved in [14] even for Γ being a Fuchsian group of the second kind.

The first result of importance to us is

Theorem 5.1 ([13]) *Let $F = (f_1, f_2)^t$ be a 2-dimensional vector-valued automorphic form of multiplier ρ and arbitrary weight for Γ , then $h_F = f_1/f_2$ is a ρ -equivariant function for Γ .*

This settles the question of the existence of ρ -equivariant functions which is then a consequence of the existence of vector valued automorphic forms. A more interesting result is that every ρ -equivariant functions arises in this way

Theorem 5.2 ([13]) *The map from $V_k(\Gamma, \rho)$ to $Eq(\Gamma, \rho)$ given by*

$$F = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} \mapsto h_F = f_1/f_2 \tag{3}$$

is surjective.

Surprisingly, the proof uses almost all the properties of the Schwarz derivative which lead to the next theorem. If D is a domain in \mathbb{C} , and $R(z)$ is a holomorphic function on D , then we cannot guarantee the existence of two linearly independent global

solutions to the differential equation

$$y'' + R(z)y = 0$$

when D is not simply connected, and all we can hope for are local solutions. However, when $R(z)$ comes from a Schwarz derivative, then we have a different outcome.

Theorem 5.3 *Let D be a domain and f be a meromorphic function on D such that $R(z) = \{f, z\}$ is holomorphic on D . Then the differential equation $y'' + R(z)y = 0$ has two linearly independent global solutions on D .*

It is this important result and the use of the Bol identity that lead to the surjectivity of the map (3).

So far we have established this close connection between ρ -equivariant functions for Γ and 2-dimensional vector valued automorphic forms. All we need is to prove, for arbitrary data (Γ, ρ) , the existence of such automorphic forms. To this end we associate to (Γ, ρ) a vector bundle $\mathcal{E} = \mathcal{E}_{\Gamma, \rho}$ over $X = X(\Gamma)$ constructed as follows:

We choose a covering $\mathcal{U} = (U_i)_{i \in I}$ where I is the set of cusps and elliptic fixed points on X . We then construct holomorphic maps $\psi_i : U_i \rightarrow \text{GL}_2(\mathbb{C})$ having ρ as a factor of automorphy [7]. This is carried out by solving the Riemann-Hilbert problem over U_i with the monodromy ρ . These maps yield a cocycle $(F_{ij}) \in \mathcal{Z}^1(\mathcal{U}, \text{GL}(2, \mathcal{O}))$ to which is associated a rank two holomorphic vector bundle \mathcal{E} over X whose transition functions are the maps F_{ij} on $U_i \cap U_j$.

Now if P is a given point (that can be a cusp) and \mathcal{L} is the line bundle over X corresponding to the divisor $[P]$, then using the Kodaira vanishing theorem, there exists an integer $\mu \geq 0$ such that

$$\dim H^0(X, \mathcal{O}(\mathcal{L}^\mu \otimes \mathcal{E})) \geq 2$$

where $\mathcal{O}(\mathcal{L}^\mu \otimes \mathcal{E})$ is the set of holomorphic sections of the sheaf $\mathcal{L}^\mu \otimes \mathcal{E}$ which can be seen as sections in $H^0(X \setminus \{P\}, \mathcal{O}(\mathcal{E}))$ having a pole at P of order at most μ . Thus we have two linearly independent meromorphic sections of \mathcal{E} with a single pole at P . When lifted to $H \cup \{\text{cusps}\}$ these sections yield two linearly independent vector-valued automorphic functions (of weight 0) attached to (Γ, ρ) with poles at the fiber of P . The full details of the proof can be found for a higher dimension of the representation in [14].

We have therefore established the following:

Theorem 5.4 *For every discrete group Γ and every 2-dimensional representation ρ of Γ , vector-valued Γ -automorphic functions of multiplier ρ exist and so do ρ -equivariant functions for Γ .*

Acknowledgements I would like to thank the organizers of the MATRIX Institute workshop on Hypergeometric motives and Calabi–Yau differential equations for their invitation. The 3 weeks I spent at the MATRIX institute were truly inspiring.

References

1. Ahlfors, L.: Lectures on Quasiconformal Mappings. Van Nostrand, Princeton (1966)
2. Sebbar, A., Al-Shbeil, I.: Elliptic zeta functions and equivariant functions. *Can. Math. Bull.* **61**(2), 376–389 (2018)
3. Brady, M.: Meromorphic solutions of a system of functional equations involving the modular group. *Proc. AMS* **30**(2), 271–277 (1970)
4. Duren, P.L.: Univalent Functions. Grundlehren der Mathematischen Wissenschaften, vol. 259. Springer, New York (1983)
5. Elbasraoui, A., Sebbar, A.: Rational equivariant forms. *Int. J. Number Theory* **8**(4), 963–981 (2012)
6. Ford, L.R.: Automorphic Functions. McGraw-Hill, New York (1929)
7. Forster, O.: Lectures on Riemann Surfaces. Graduate Texts in Mathematics, vol. 81. Springer, New York (1991)
8. McKay, J., Sebbar, A.: Fuchsian groups, automorphic functions and Schwarzians. *Math. Ann.* **318**(2), 255–275 (2000)
9. Nehari, Z.: The Schwarzian derivative and schlicht functions. *Bull. Am. Math. Soc.* **55**, 545–551 (1949)
10. Ovsienko, V., Tabachnikov, S.: What is the Schwarzian derivative? *AMS Not.* **56**(01), 34–36 (2009)
11. Sansone, G., Gerretsen, J.: Lectures on the Theory of Functions of a Complex Variable. II. Geometric Theory. Wolters-Noordhoff Publishing, Groningen (1969)
12. Sebbar, A., Saber, H.: On the critical points of modular forms. *J. Number Theory* **132**(8), 1780–1787 (2012)
13. Sebbar, A., Saber, H.: Equivariant functions and vector-valued modular forms. *Int. J. Number Theory* **10**(4), 949–954 (2014)
14. Saber, H., Sebbar, A.: On the existence of vector-valued automorphic forms. *Kyushu J. Math.* **71**(2), 271–285 (2017)
15. Sebbar, A., Sebbar, A.: Equivariant functions and integrals of elliptic functions. *Geom. Dedicata* **160**(1), 373–414 (2012)

Hypergeometric Functions over Finite Fields



Jenny Fuselier, Ling Long, Ravi Ramakrishna, Holly Swisher,
and Fang-Ting Tu

Abstract We discuss recent work of the authors in which we study the translation of classical hypergeometric transformation and evaluation formulas to the finite field setting.

Our approach is motivated by the desire for both an algorithmic type approach that closely parallels the classical case, and an approach that aligns with geometry. In light of these objectives, we focus on period functions in our construction which makes point counting on the corresponding varieties as straightforward as possible.

We are also motivated by previous work joint with Deines, Fuselier, Long, and Tu in which we study generalized Legendre curves using periods to determine a condition for when the endomorphism algebra of the primitive part of the associated Jacobian variety contains a quaternion algebra over \mathbb{Q} . In most cases this involves computing Galois representations attached to the Jacobian varieties using Greene's finite field hypergeometric functions.

J. Fuselier
High Point University, High Point, NC, USA
e-mail: jfuselie@highpoint.edu

L. Long · F.-T. Tu
Louisiana State University, Baton Rouge, LA, USA
e-mail: llong@math.lsu.edu; tu@math.lsu.edu

R. Ramakrishna
Cornell University, Ithaca, NY, USA
e-mail: ravi@math.cornell.edu

H. Swisher (✉)
Oregon State University, Corvallis, OR, USA
e-mail: swisherh@math.oregonstate.edu

1 Motivation

In this talk we discuss recent work of the authors [9], in which we study the translation of classical hypergeometric transformation and evaluation formulas to the finite field setting. The theory of classical hypergeometric functions and hypergeometric functions over finite fields sits inside the broader framework of hypergeometric motives. Hypergeometric functions over finite fields have been developed by several people, including for example Evans [5, 6], Greene [10], Katz [11], and McCarthy [12], and a number of current developments have been discussed at this workshop including recent work of Roberts et al. [13], Doran et al. [4], and Beukers et al. [2], for example.

Our approach to translation of classical hypergeometric transformation and evaluation formulas to the finite field setting is motivated by two strong desires:

1. An algorithmic type approach that closely parallels the classical case (and does not require particular ingenuity for each example).
2. An approach that aligns with geometry by explicitly interpreting the finite field hypergeometric functions in terms of Galois representations corresponding to associated algebraic varieties.

In light of these objectives, we focus on period functions in our construction which makes point counting on the corresponding varieties as straightforward as possible.

We are also motivated by previous work joint with Deines, Fuselier, Long, and Tu [3] in which we study generalized Legendre curves $y^N = x^i(1-x)^j(1-\lambda x)^k$, using periods to determine for certain N a condition for when the endomorphism algebra of the primitive part of the associated Jacobian variety contains a quaternion algebra over \mathbb{Q} . In most cases this involves computing Galois representations attached to the Jacobian varieties using Greene's finite field hypergeometric functions.

2 Method

From this perspective, the following approach is very natural. We slightly modify the finite field hypergeometric function definition of Greene (or McCarthy) by inductively using the Euler integral representation to construct our analogues. In the classical setting, define the period function ${}_1P_0[a; z] := (1-z)^{-a} = {}_1F_0[a; z]$, and then use the Euler integral formula to define

$${}_2P_1[a, b; c; z] := \int_0^1 t^{b-1}(1-t)^{c-b-1} {}_1P_0[a; zt] dt = B(b, c-b) {}_2F_1[a, b; c; z],$$

where $B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} dt$ is the beta function. To justify calling these period functions we note that Wolfart [15] realized that if the parameters $a, b, c \in \mathbb{Q}$,

and $a, b, a - c, b - c \notin \mathbb{Z}$, then the integrals

$${}_2P_1 \left[\begin{matrix} a & b \\ & c \end{matrix} ; \lambda \right] \text{ and } (-1)^{c-a-b-1} \lambda^{1-c} {}_2P_1 \left[\begin{matrix} 1+b-c & 1+a-c \\ & 2-c \end{matrix} ; \lambda \right]$$

are both periods of a generalized Legendre curve $y^N = x^i(1-x)^j(1-\lambda x)^k$, where $N = \text{lcd}(a, b, c)$ (least common denominator), $i = N \cdot (1 - b)$, $j = N \cdot (1 + b - c)$, and $k = N \cdot a$.

Inductively we define higher period functions ${}_{n+1}P_n$, gathering additional beta function terms in front of the ${}_{n+1}F_n$. We can then use the following ‘‘dictionary’’ which is well-known to experts to translate these period functions to the finite field setting. Here $q = p^e$ for prime p , and $\widehat{\mathbb{F}_q^\times}$ denotes the group of multiplicative characters on \mathbb{F}_q^\times , where each character A is extended to \mathbb{F}_q by defining $A(0) = 0$. Furthermore $N \in \mathbb{N}$, $a, b \in \mathbb{Q}$ with common denominator N , $\eta_N \in \widehat{\mathbb{F}_q^\times}$ has order N , \bar{A} denotes the complex conjugate of A , and ζ_p is a primitive p th root of unity.

$$\begin{aligned} a = \frac{i}{N}, b = \frac{j}{N} &\leftrightarrow A, B \in \widehat{\mathbb{F}_q^\times}, A = \eta_N^i, B = \eta_N^j \\ x^a &\leftrightarrow A(x) \\ -a &\leftrightarrow \bar{A} \\ \int_0^1 dx &\leftrightarrow \sum_{x \in \mathbb{F}} \\ \Gamma(a) &\leftrightarrow g(A) = \sum_{x \in \mathbb{F}_q^\times} \zeta_p^{x+x^p+x^{p^2}+\dots+x^{p^{e-1}}} \\ B(a, b) &\leftrightarrow J(A, B) = \sum_{x \in \mathbb{F}_q} A(x)B(1-x) \end{aligned}$$

Thus we correspondingly define ${}_1P_0[A; \lambda] := \bar{A}(1 - \lambda)$,

$${}_2P_1[A, B; C; \lambda] := \sum_{y \in \mathbb{F}_q} B(y)\bar{B}C(1-y){}_1P_0[A; \lambda y],$$

and define ${}_{n+1}P_n$ inductively. We obtain a nice point counting formula for the hypergeometric variety $X_\lambda : y^N = x_1^{i_1} \dots x_n^{i_n}(1-x_1)^{j_1} \dots (1-x_n)^{j_n}(1-\lambda x_1 \dots x_n)^k$. In particular, we have for $q \equiv 1 \pmod{N}$,

$$\#X_\lambda(\mathbb{F}_q) = 1 + q^n + \sum_{m=1}^{N-1} {}_{n+1}P_n \left[\begin{matrix} \eta_N^{-mk} & \eta_N^{mi_n} & \dots & \eta_N^{mi_1} \\ \eta_N^{mi_n+mj_n} & \dots & \eta_N^{mi_1+mj_1} & \end{matrix} ; \lambda \right].$$

Normalizing the period functions ${}_{n+1}P_n$ by dividing by the appropriate Jacobi sums using the dictionary, gives our finite field analogues to the classical hypergeometric functions which we denote by ${}_{n+1}F_n$. For example when $n = 1$,

$${}_2F_1[A, B; C; \lambda] := \frac{1}{J(B, \bar{C}\bar{B})} {}_2P_1[A, B; C; \lambda].$$

If none of the “top” parameters are the trivial character, or match with one of the “bottom” parameters, we call the ${}_{n+1}\mathbb{P}_n$ or ${}_{n+1}\mathbb{F}_n$ *primitive*. Our definition of ${}_{n+1}\mathbb{F}_n$ has two nice properties that match the classical case: it is 1 when evaluated at 0, and in the primitive case it is symmetric in both the top or bottom parameters.

We note that key properties such as the reflection and multiplication formulas for the Gamma function translate using the dictionary to properties of the Gauss sum. Our method allows that any classical formula proved using these properties, as well as their corollaries such as the Pfaff-Saalchütz formula, can be translated directly (introducing error terms as needed) and thus we can indeed use an algorithmic type approach to translation that closely parallels proofs in the classical case. However, this approach does not work for everything in the classical setting; for example proofs involving a derivative structure cannot be translated in this way.

3 Galois Interpretation

We can interpret the ${}_{n+1}\mathbb{P}_n$ or ${}_{n+1}\mathbb{F}_n$ functions as traces of Galois representations at Frobenius elements via the corresponding hypergeometric algebraic varieties. In the $n = 1$ case we make this explicit in the following theorem.

For a given number field K , denote its ring of integers by \mathcal{O}_K , its algebraic closure by \overline{K} , and set $G_K := \text{Gal}(\overline{K}/K)$. We call a prime ideal \mathfrak{p} of \mathcal{O}_K unramified if it is coprime to the discriminant of K . Fix $\lambda \in \overline{\mathbb{Q}}$. Given a rational number of the form $\frac{i}{m}$, a number field K containing $\mathbb{Q}(\zeta_m, \lambda)$ and a prime ideal \mathfrak{p} of K coprime to the discriminant of K , one can assign a multiplicative character $\iota_{\mathfrak{p}}(\frac{i}{m})$ to the residue field $\mathcal{O}_K/\mathfrak{p}$ of \mathfrak{p} with size $q(\mathfrak{p}) = |\mathcal{O}_K/\mathfrak{p}|$. This assignment, based on the m th power residue symbol, is compatible with the Galois perspective when \mathfrak{p} varies. It is also compatible with field extensions of K . Thus the finite field analogues of classical period (or hypergeometric) functions are viewed as the converted functions over the finite residue fields, unless otherwise specified. We show the following.

Theorem 1 *Let $a, b, c \in \mathbb{Q}$ with least common denominator N such that $a, b, a - c, b - c \notin \mathbb{Z}$ and $\lambda \in \overline{\mathbb{Q}} \setminus \{0, 1\}$. Let K be the Galois closure of $\mathbb{Q}(\lambda, \zeta_N)$ with the ring of integers \mathcal{O}_K , and ℓ any prime. Then there is a 2-dimensional representation $\sigma_{\lambda, \ell}$ of $G_K := \text{Gal}(\overline{K}/K)$ over $\mathbb{Q}_{\ell}(\zeta_N)$, depending on a, b, c , such that for each unramified prime ideal \mathfrak{p} of \mathcal{O}_K for which λ and $1 - \lambda$ can be mapped to nonzero elements in the residue field, $\sigma_{\lambda, \ell}$ evaluated at the arithmetic Frobenius conjugacy class $\text{Frob}_{\mathfrak{p}}$ at \mathfrak{p} is an algebraic integer (independent of the choice of ℓ), satisfying*

$$\text{Tr } \sigma_{\lambda, \ell}(\text{Frob}_{\mathfrak{p}}) = -2\mathbb{P}_1 \begin{bmatrix} \iota_{\mathfrak{p}}(a) & \iota_{\mathfrak{p}}(b) \\ & \iota_{\mathfrak{p}}(c) \end{bmatrix}; \lambda; q(\mathfrak{p})$$

As a corollary to this theorem, given a primitive $2\mathbb{P}_1$ and a prime ℓ we can compute the L-function of the corresponding 2-dimensional Galois representation (ℓ -adic) as a product over good primes of terms involving $2\mathbb{P}_1$ and $(2\mathbb{P}_1)^2$.

4 Examples of Translated Identities

As examples of our techniques we use both our algorithmic type approach as well as the Galois perspective to translate several classical hypergeometric formulas to the finite field setting, including transformations of degree 1, 2, 3, algebraic identities, and evaluation formulas.

For example, we translate a Clausen identity between the square of a ${}_2F_1$ and a ${}_3F_2$. From a differential equations perspective, this identity is indicating that the symmetric square of the 2-dimensional solution space to the corresponding hypergeometric differential equation (HDE) corresponding to the ${}_2F_1$ is the 3-dimensional solution space of the HDE corresponding to the ${}_3F_2$. Translating this identity to the finite field setting yields a finite field hypergeometric transformation due to Greene and Evans [7] which in our notation more closely matches the classical identity. With the representation theoretic perspective, it indicates the fact that the tensor square of a 2-dimensional representation (associated to the ${}_2\mathbb{F}_1$) is its symmetric square (which is a 3-dimensional representation associated to the ${}_3\mathbb{F}_2$) plus its alternating square (which is a linear representation).

For an example of an algebraic type identity, consider the following identity in Slater [14, (1.5.20)] which gives that

$${}_2F_1 \left[\begin{matrix} a & a - \frac{1}{2} \\ 2a \end{matrix} ; z \right] = \left(\frac{1 + \sqrt{1-z}}{2} \right)^{1-2a}.$$

To see its finite field analogue, it is tempting to translate the right hand side into a corresponding character evaluated at $\frac{1+\sqrt{1-z}}{2}$ using the dictionary. However, Theorem 1 implies that one character is insufficient as the corresponding Galois representations should be 2-dimensional. Instead, our translated identity becomes the following. For \mathbb{F}_q of odd characteristic, ϕ the quadratic character, $A \in \widehat{\mathbb{F}_q^\times}$ having order at least 3, and $z \in \mathbb{F}_q$,

$${}_2\mathbb{F}_1 \left[\begin{matrix} A & A\phi \\ A^2 \end{matrix} ; z \right] = \left(\frac{1 + \phi(1-z)}{2} \right) \left(\overline{A}^2 \left(\frac{1 + \sqrt{1-z}}{2} \right) + A^2 \left(\frac{1 - \sqrt{1-z}}{2} \right) \right).$$

The proof is quite straightforward using only translations of Kummer 24 relations as well as the reflection and duplication formulas. This example highlights that the Galois perspective allows us to predict analogues beyond the dictionary alone.

As another example, we use our dictionary technique to translate a quadratic ${}_2F_1$ transformation of Kummer [1, Thm. 3.1.1] which we first show can be proved using only the multiplication and reflection formulas with the Pfaff-Saalschütz identity. The finite field version we obtain is equivalent to a quadratic formula of Greene in [10], but holds for all values in \mathbb{F}_q . Our proof (although it might appear technical on the surface) is very straightforward. In comparison, the approaches of Evans

and Greene to higher order transformation formulas (such as [8, 10]) often involve clever changes of variables. This example also demonstrates that our method has the capacity to produce finite field analogues that are satisfied by all values in \mathbb{F}_q .

As an explicit application of finite field formulas in computing the arithmetic invariants of hypergeometric varieties, we use the finite field quadratic transformation from the previous example to obtain the decomposition of a generically 4-dimensional abelian variety arising naturally from the generalized Legendre curve $y^{12} = x^9(1-x)^5(1-\lambda x)$.

Acknowledgements Many thanks to the International Mathematical Research Institute MATRIX in Australia for hosting the workshop on Hypergeometric Motives and Calabi–Yau Differential Equations where this talk was presented.

References

1. Andrews, G.E., Askey, R., Roy, R.: Special Functions. Encyclopedia of Mathematics and Its Applications, vol. 71. Cambridge University Press, Cambridge (1999)
2. Beukers, F., Cohen, H., Mellit, A.: Finite hypergeometric functions. arXiv: 1505.02900
3. Deines, A., Fuselier, J.G., Long, L., Swisher, H., Tu, F.T.: Generalized legendre curves and quaternionic multiplication. *J. Number Theory* **161**, 175–203 (2016)
4. Doran, C.F., Kelly, T.L., Salerno, A., Sperber, S., Voight, J., Whitcher, U.: Zeta functions of alternate mirror Calabi–Yau families. arXiv: 1612.09249
5. Evans, R.J.: Identities for products of Gauss sums over finite fields. *Enseign. Math. (2)* **27**(3–4), 197–209 (1982)
6. Evans, R.J.: Character sums over finite fields. In: *Finite Fields, Coding Theory, and Advances in Communications and Computing* (Las Vegas, NV, 1991). Lecture Notes in Pure and Applied Mathematics, vol. 141, pp. 57–73. Dekker, New York (1993)
7. Evans, R.J., Greene, J.: Clausen’s theorem and hypergeometric functions over finite fields. *Finite Fields Appl.* **15**(1), 97–109 (2009)
8. Evans, R.J., Greene, J.: A quadratic hypergeometric 2F1 transformation over finite fields. *Proc. Am. Math. Soc.* **145**, 1071–1076 (2017)
9. Fuselier, J.G., Long, L., Ramakrishna, R., Swisher, H., Tu, F.T.: Hypergeometric functions over finite fields. arXiv: 1510.02575
10. Greene, J.: Hypergeometric functions over finite fields. *Trans. Am. Math. Soc.* **301**(1), 77–101 (1987)
11. Katz, N.M.: *Exponential Sums and Differential Equations*. Annals of Mathematics Studies, vol. 124. Princeton University Press, Princeton (1990)
12. McCarthy, D.: Transformations of well-poised hypergeometric functions over finite fields. *Finite Fields Appl.* **18**(6), 1133–1147 (2012)
13. Roberts, D., Rodriguez Villegas, F.: Hypergeometric supercongruences. arXiv:1803.10834
14. Slater, L.J.: *Generalized Hypergeometric Functions*. Cambridge University Press, Cambridge (1966)
15. Wolfart, J.: Werte hypergeometrischer Funktionen. *Invent. Math.* **92**(1), 187–216 (1988)

Supercongruences Occurred to Rigid Hypergeometric Type Calabi–Yau Threefolds



Ling Long, Fang-Ting Tu, Noriko Yui, and Wadim Zudilin

Abstract In this project, we establish the supercongruences for the 14 families of rigid hypergeometric Calabi–Yau threefolds conjectured by Roriguez-Villegas in 2003.

1 Main Result

The talk outlines the proof of the supercongruences for the 14 families of rigid hypergeometric Calabi–Yau threefolds conjectured by Roriguez-Villegas [9].

Theorem 1 Let $d_1, d_2 \in \{1/2, 1/3, 1/4, 1/6\}$ or

$$(d_1, d_2) = (1/5, 2/5), (1/8, 3/8), (1/10, 3/10), (1/12, 5/12).$$

Then for each prime $p > 5$, we have

$${}_4F_3 \left[\begin{matrix} d_1 & 1 - d_1 & d_2 & 1 - d_2 \\ & 1 & 1 & 1 \end{matrix} ; 1 \right]_{p-1} \equiv a_p(f_{d_1, d_2}) \pmod{p^3},$$

where the hypergeometric series on the left-hand side is truncated after $p - 1$ terms and $a_p(f_{d_1, d_2})$ is the p th coefficient of an explicit Hecke eigenform f_{d_1, d_2} of weight 4 associated to the corresponding rigid Calabi–Yau manifold via the modularity theorem.

L. Long · F.-T. Tu (✉)
Louisiana State University, Baton Rouge, LA, USA
e-mail: llong@lsu.edu; ftu@lsu.edu

N. Yui
Queen's University, Kingston, ON, Canada
e-mail: yui@mast.queensu.ca

W. Zudilin
The University of Newcastle, Callaghan, NSW, Australia

2 Motivation

The term supercongruence refers to a congruence which is stronger than what the formal group law implies. In [3] Beukers proved

$$A\left(\frac{p-1}{2}\right) = {}_4F_3\left[\begin{matrix} \frac{1-p}{2} & \frac{1-p}{2} & \frac{p+1}{2} & \frac{p+1}{2} \\ & 1 & 1 & 1 \end{matrix}; 1\right] \equiv a_p(f) \pmod{p},$$

where $A(n)$ are the Apéry numbers

$$A(n) := \sum_{k=0}^n \binom{n}{k}^2 \binom{n+k}{k}^2 = {}_4F_3\left[\begin{matrix} -n & -n & n+1 & n+1 \\ & 1 & 1 & 1 \end{matrix}; 1\right]$$

and $a_p(f)$ is the p th coefficient of the Hecke eigenform $\eta^4(2\tau)\eta^4(4\tau)$. In [3] Beukers also conjectured the supercongruence

$${}_4F_3\left[\begin{matrix} \frac{1-p}{2} & \frac{1-p}{2} & \frac{p+1}{2} & \frac{p+1}{2} \\ & 1 & 1 & 1 \end{matrix}; 1\right] \equiv a_p(f) \pmod{p^2}.$$

This was proved by Ahlgren and Ono [1]. Their key idea is using Green’s hypergeometric function over finite fields to perform point counting on the Calabi–Yau threefold

$$\left\{x + \frac{1}{x} + y + \frac{1}{y} + z + \frac{1}{z} + w + \frac{1}{w} = 0\right\},$$

which is modular. Later, Kilbourn [7] gives an extension of the supercongruence

$$a_p(f) \equiv \sum_{j=0}^{p-1} \binom{2j}{j}^4 2^{-8j} = {}_4F_3\left[\begin{matrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ & 1 & 1 & 1 \end{matrix}; 1\right]_{p-1} \pmod{p^3}, \tag{1}$$

which was conjectured by Van Hamme. Kilbourn’s proof is mainly relying on p -adic tools. Using the techniques similar to the ones given by Ahlgren, Ono and Kilbourn, McCarthy [8] obtained the supercongruence

$${}_4F_3\left[\begin{matrix} \frac{1}{5} & \frac{2}{5} & \frac{3}{5} & \frac{4}{5} \\ & 1 & 1 & 1 \end{matrix}; 1\right]_{p-1} \equiv a_p(f_{1/5,2/5}) \pmod{p^3}, \tag{2}$$

where $f_{1/5,2/5}$ is an explicit Hecke eigenform conjectured by Rodriguez-Villegas. This supercongruence corresponds to the mirror quintic threefold in \mathbb{P}^4 , whose modularity was first established by Schoen [10]. The supercongruences given by

Kilbourn (1) and McCarthy (2) are particular instances of Rodriguez-Villegas’s conjectures.

In this joint project, our main motivation is to study the arithmetic aspect of rigid hypergeometric type Calabi–Yau manifolds. The first step is verifying the supercongruences conjectured by Rodriguez-Villegas coming from the well-known 14 hypergeometric families of Calabi–Yau threefolds whose Picard–Fuchs equations are degree 4 hypergeometric differential equations with solution near 0 of the form

$${}_4F_3 \left[\begin{matrix} d_1 & 1 - d_1 & d_2 & 1 - d_2 \\ & 1 & 1 & 1 \end{matrix} ; z \right],$$

where d_1, d_2 are as in Theorem 1. When $z = 1$, it corresponds to the singularity of the hypergeometric differential equation, which is equivalent to getting a rigid Calabi–Yau threefold in the fibre. Due to Gouvêa and Yui [6], a rigid Calabi–Yau threefold defined over \mathbb{Q} is modular. This means, the L -function associated with the third étale cohomology group of a rigid Calabi–Yau threefold V in the 14 families is equal to the L -function of an explicit Hecke eigenform of weight 4 conjectured by Rodriguez-Villegas.

Very recently, Fuselier and McCarthy [5] establish the case $(d_1, d_2) = (1/2, 1/4)$. In this joint project, we provide a more general method to verify the remaining 11 cases of supercongruences conjectured by Rodriguez-Villegas.

3 Key Ideas and Example

The strategy of our proof is to use hypergeometric motives over \mathbb{Q} to describe the arithmetic background. There are different versions of hypergeometric motives such as given by Katz, Greene and McCarthy. However, for our purposes, the most convenient one is the general version given by Beukers, Cohen and Mellit in [4]. They modify Katz’s finite hypergeometric function $H(\alpha, \beta; \lambda)$ so that their version works for all the primes p . They also give a recipe to realize toric models as hypergeometric motives arising from certain type hypergeometric data $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d)$ and $\beta = (\beta_1, \beta_2, \dots, \beta_d)$, where $\alpha_i, \beta_j \in \mathbb{Q}$. For such case, one can express the number of rational points over finite fields on the given model in terms of $H(\alpha, \beta; \lambda)$.

For example, the multi-sets $\alpha = (1/3, 2/3, 1/3, 2/3)$ and $\beta = (1, 1, 1, 1)$ give the hypergeometric type Calabi–Yau threefold with $d_1 = d_2 = 1/3$. The toric model in this case corresponds to the resolution of singularities on the affine variety given by projective equations

$$W : x_1 + x_2 + x_3 + x_4 = y_1 + y_2 + y_3 + y_4 = 0, (x_1 y_1)^3 = 3^6 x_2 x_3 x_4 y_2 y_3 y_4, x_i, y_j \neq 0.$$

The resulting manifold \overline{W} is the rigid Calabi–Yau threefold labelled as $V_{3,3}$ by Batyrev and van Straten in [2].

In the talk, principal ideas of our proof are illustrated in the case $d_1 = d_2 = 1/3$.

References

1. Ahlgren, S., Ono, K.: A Gaussian hypergeometric series evaluation and Apéry number congruences. *J. Reine Angew. Math.* **518**, 187–212 (2000)
2. Batyrev, V., van Straten, D.: Generalized hypergeometric functions and rational curves on Calabi–Yau complete intersections in toric varieties. *Commun. Math. Phys.* **168**, 493–533 (1995)
3. Beukers, F.: Another congruence for the Apéry numbers. *J. Number Theory* **25**, 201–210 (1987)
4. Beukers, F., Cohen, H., Mellit, A.: Finite hypergeometric functions. *Pure Appl. Math. Q.* **11**, 559–589 (2015)
5. Fuselier, J.G., McCarthy, D.: Hypergeometric type identities in the p -adic setting and modular forms. *Proc. Am. Math. Soc.* **144**, 1493–1508 (2016)
6. Gouvêa, F., Yui, N.: Rigid Calabi–Yau threefolds over \mathbb{Q} are modular. *Expo. Math.* **29**, 142–149 (2011)
7. Kilbourn, T.: An extension of the Apéry number supercongruence. *Acta Arith.* **123**, 335–348 (2006)
8. McCarthy, D.: On a supercongruence conjecture of Rodriguez-Villegas. *Proc. Am. Math. Soc.* **140**, 2241–2254 (2012)
9. Rodriguez-Villegas, F.: Hypergeometric families of Calabi–Yau manifolds. In: *Calabi–Yau Varieties and Mirror Symmetry* (Toronto, ON, 2001). Fields Institute Communications, vol. 38, pp. 223–231. American Mathematical Society, Providence (2003)
10. Schoen, C.: On the geometry of a special determinantal hypersurfaces associated to the Mumford–Horrocks vector bundle. *J. Reine Angew. Math.* **364**, 85–111 (1986)

p -Adic Hypergeometrics



Fernando Rodriguez Villegas

Abstract We study classical hypergeometric series as a p -adic function of its parameters inspired by a problem in the Monthly solved by D. Zagier.

1 Introduction

The classical generalized hypergeometric series is defined by

$${}_rF_{r-1} \left[\begin{matrix} \alpha_1 \dots \alpha_r \\ \beta_1 \dots \beta_{r-1} \end{matrix} \middle| t \right] = \sum_{k \geq 0} \frac{(\alpha_1)_k \dots (\alpha_r)_k}{(\beta_1)_k \dots (\beta_{r-1})_k} \frac{t^k}{k!} \quad (1)$$

for $\alpha_j \in \mathbb{C}$ and $\beta_j \in \mathbb{C} \setminus \{0, -1, -2, \dots\}$. If $\alpha_r = -n$ for n a non-negative integer the series terminates and we have

$${}_rF_{r-1} \left[\begin{matrix} \alpha_1 \dots -n \\ \beta_1 \dots \beta_{r-1} \end{matrix} \middle| t \right] = \sum_{k=0}^n (-1)^k \binom{n}{k} \frac{(\alpha_1)_k \dots (\alpha_{r-1})_k}{(\beta_1)_k \dots (\beta_{r-1})_k} t^k. \quad (2)$$

One can show that for fixed $\alpha_i \in \mathbb{Z}_p, \beta_i \in \mathbb{Z}_p \setminus \{0, -1, -2, \dots\}$ and $|t|_p < 1$ this yields a convergent Mahler series and hence a continuous function f of the variable $x := n$ in \mathbb{Z}_p

$$f(x) := {}_rF_{r-1} \left[\begin{matrix} \alpha_1 \dots -x \\ \beta_1 \dots \beta_{r-1} \end{matrix} \middle| t \right] = \sum_{k \geq 0} (-1)^k \binom{x}{k} \frac{(\alpha_1)_k \dots (\alpha_{r-1})_k}{(\beta_1)_k \dots (\beta_{r-1})_k} t^k. \quad (3)$$

These functions seem very interesting and worthy of further investigation.

F. Rodriguez Villegas (✉)
Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
e-mail: villegas@ictp.it

2 The Problem

It turns out that a special case of these functions appears in the solution of an interesting Monthly problem [3] solved by D. Zagier. The problem is to prove that

$$v_3 \left(\sum_{k=0}^{n-1} \binom{2k}{k} \right) = v_3 \left(n^2 \binom{2n}{n} \right), \tag{4}$$

where v_p denotes the p -adic valuation. Zagier does this by showing that there is a continuous function $f_1 : \mathbb{Z}_3 \rightarrow -1 + 3\mathbb{Z}_3$ which interpolates the values

$$f_1(n) = \frac{\sum_{k=0}^{n-1} \binom{2k}{k}}{n^2 \binom{2n}{n}}, \quad n = 1, 2, \dots \tag{5}$$

Considering the expansion

$$f_1(n) = A + Bn + Cn^2 + \dots$$

he goes further and conjectures, based on numerical evidence, that $B = 0$; moreover, he mentions

Another interesting problem would be to evaluate in closed form the 3-adic number A .

We prove that in fact

$$A = -\frac{3}{2}\zeta_3(2) = 2 + 3 + 2 \cdot 3^2 + 2 \cdot 3^6 + 3^7 + 2 \cdot 3^8 + 2 \cdot 3^9 + O(3^{10}), \tag{6}$$

where $\zeta_3(s)$ is the Kubota-Leopoldt 3-adic zeta function.

3 Periods

The connection with zeta values is perhaps to be expected: in general the Taylor coefficients of the functions of Sect. 1 involve multiple polylogarithms. In the specific case in question we have

$$f(n) = \sum_{k=1}^n \frac{1}{\binom{2k}{k}} \binom{n}{k} (-3)^{k-1}, \quad f(n) = nf_1(n). \tag{7}$$

If we expand in general

$$f(x) := \sum_{k \geq 1} \frac{1}{\binom{2k}{k}} \binom{x}{k} t^{k-1} = \sum_{n \geq 0} b_n(t) x^n,$$

then

$$b_n(t) = \frac{1}{(t+4)} \sum_{0 \leq j_1 < j_2 < \dots < j_n} \frac{\left(\frac{t}{t+4}\right)^{j_n}}{(j_1 + \frac{1}{2})(j_2 + \frac{1}{2}) \cdots (j_n + \frac{1}{2})}. \tag{8}$$

These multiple polylogarithms can be expressed in terms of usual polylogarithms for small n . Trivially $b_0 = 0$. For $n = 1$ we have the following identity of power series in $z = 1 - w$

$$b_1((w - w^{-1})^2) = (w^2 - w^{-2})^{-1} \log(w^2), \quad w = 1 - z. \tag{9}$$

By plugging in a primitive third root of unity $\zeta_3 \in \mathbb{C}_3$ for w it follows that 3-adically we have $b_1(-3) = 0$. This shows that in this case $f(x)$ is divisible by x and we may consider $f_1(x) := f(x)/x$ (see [3]).

With some effort one can prove that as power series in z , with $w = 1 - z$, we have

$$b_2(w - w^{-1})^2 = (w^2 - w^{-2})^{-1} [\text{Li}_2(1 - w^2) - \frac{1}{2} \text{Li}_2(1 - w^4) - \text{Li}_2(1 - w^{-2}) + \frac{1}{2} \text{Li}_2(1 - w^{-4})], \tag{10}$$

where Li_2 is the standard dilogarithm function.

Plugging in $w = \zeta_3 \in \mathbb{C}_3$ into (10) and using a result of Coleman [1] we obtain (6). The identity is the special case $p = 3, r = 1$ of the following. Given a prime $p > 2$ fix $\zeta_p \in \mathbb{C}_p$ a primitive p -th root of unity.

Theorem 1

i) *The following limit exists*

$$A(\zeta_p) := \lim_{s \rightarrow \infty} \frac{1}{\binom{2p^s}{p^s} p^{2s}} \sum_{k=0}^{p^s-1} \binom{2k}{k} (\zeta_p + \zeta_p^{-1})^{2(p^s-1-k)} \tag{11}$$

ii) *Let $\omega : \mathbb{F}_p^\times \rightarrow \mathbb{C}_p^\times$ be the Teichmüller character. For $0 < r < p - 1$ we have*

$$\frac{1}{(\omega^r(4) - 2\omega^r(2))} \sum_{i=1}^{p-1} \omega(i)^{-r} (\zeta_p^{2i} - \zeta_p^{-2i}) A(\zeta_p^i) = L_p(2, \omega^{r-1}), \tag{12}$$

where L_p is Kubota-Leopoldt's p -adic L -function.

We note in passing that

$$\lim_{s \rightarrow \infty} \binom{2p^s}{p^s} = 2 \prod_{k \geq 1} \frac{\Gamma_p(2p^k)}{\Gamma_p(p^k)^2}$$

(see [2, §6.3.4, ex. 16]), where Γ_p denotes the p -adic gamma function.

The beauty of the expressions (9) and (10) is that though their proof were obtained working over the complex numbers they are identities of power series with rational coefficients and hence also hold p -adically in an appropriate domain. Fortunately, this domain includes the point we need to evaluate for Zagier's questions ($w = \zeta_3 \in \mathbb{C}_3$).

For $n = 3$ there is an expression for $b_3(t)$ in terms of polylogarithms valid over the complex numbers, which is much more difficult to obtain. For $n > 3$ we do not expect $b_n(t)$ to reduce to polylogarithms.

However, to apply this expression for b_3 to our p -adic setting requires some form of analytic continuation. This we will achieve by delicate manipulations using Coleman's integration but the details have not yet been fully carried out.

The expectation nevertheless is that for $p = 3$ we should have that $b_3(-3)$ is a simple multiple of $L_3(3, \chi_{-3})$. But $L_3(s, \chi_{-3})$ is identically zero since χ_{-3} is odd! Hence the constant B of Zagier should vanish because it is a special value of an L -function which happens to be identically zero.

4 Speculation

We tested numerically to see if there are any other relations for $b_n(t)$ and p -adic L -values and found only the following likely identities:

$$\mathbb{Q}_3 : \begin{cases} b_4(-3) = -\frac{27}{8}\zeta_3(4) \\ b_6(-3) = -\frac{297}{32}\zeta_3(6) \end{cases} \quad \mathbb{Q}_5 : \begin{cases} b_2(-5) = 0 \\ b_3(-5) = -\frac{25}{12}\zeta_5(3) \end{cases} \quad (13)$$

but we did not attempt to prove these. We pointed out above that $b_4(t)$ and $b_6(t)$ are not expected to be expressible in terms of polylogarithms. Hence the connection of the observed identities for $b_4(-3)$ and $b_6(-3)$ in \mathbb{Q}_3 appear to be less obvious than the others.

References

1. Coleman, R.: Dilogarithms, regulators and p -adic L -functions. *Invent. Math.* **69**, 171–208 (1982)
2. Rodriguez Villegas, F.: *Experimental Number Theory*. Oxford Graduate Texts in Mathematics, vol. 13, xvi+214 pp. Oxford University Press, Oxford (2007). ISBN: 978-0-19-922730-3
3. Zagier, D., Shallit, J., Strauss, N.: Problems and Solutions: 6625. *Am. Math. Mon.* **99**(1), 66–69 (1992)

On p -Adic Unit-Root Formulas



Masha Vlasenko

Abstract For a multivariate Laurent polynomial $f(x)$ with coefficients in a ring R we construct a sequence of matrices with entries in R whose reductions modulo p give iterates of the Hasse–Witt operation for the hypersurface of zeroes of the reduction of $f(x)$ modulo p . We show that our matrices satisfy a system of congruences modulo powers of p . If the Hasse–Witt operation is invertible these congruences yield p -adic limit formulas, which conjecturally describe the Gauss–Manin connection and the Frobenius operator on the slope 0 part of a crystal attached to $f(x)$. We also apply our results on congruences to integrality of formal group laws of Artin–Mazur kind.

1 Hasse–Witt Matrix

Let X/\mathbb{F}_q be a smooth projective variety of dimension n over a finite field with $q = p^a$ elements. The congruence formula due to Katz (see [1]) states that modulo p the zeta function of X is described as

$$Z(X/\mathbb{F}_q; T) \equiv \prod_{i=0}^n \det(1 - T \cdot \mathcal{F}^a | H^i(X, \mathcal{O}_X))^{(-1)^{i+1}} \pmod{p}, \quad (1)$$

where $H^i(X, \mathcal{O}_X)$ is the cohomology of X with coefficients in the structure sheaf \mathcal{O}_X and \mathcal{F} is the Frobenius map, the p -linear vector space map induced by $h \mapsto h^p$ on the structure sheaf (p -linear means $\mathcal{F}(bs + ct) = b^p \mathcal{F}(s) + c^p \mathcal{F}(t)$ for $b, c \in \mathbb{F}_q$ and $s, t \in H^i(X, \mathcal{O}_X)$). When X is a complete intersection the only interesting term in formula (1) is given by $H^n(X, \mathcal{O}_X)$. The action of \mathcal{F} on this space is classically known as the *Hasse–Witt operation*.

M. Vlasenko (✉)

Institute of Mathematics of the Polish Academy of Sciences, Warsaw, Poland

e-mail: m.vlasenko@impan.pl

The following algorithm (see [2, §7.10], [1, Corollary 6.1.13] or [3, §II.1]) can be used to compute the Hasse–Witt matrix of a hypersurface $X \subset \mathbb{P}^{n+1}$ given by a homogeneous equation $f(x_0, \dots, x_{n+1}) = 0$ of degree $d > n + 2$. One extends the Frobenius to a transformation of the exact sequence of sheaves on \mathbb{P}^{n+1} :

$$\begin{array}{ccccccc} 0 & \rightarrow & \mathcal{O}_{\mathbb{P}^{n+1}}(-d) & \xrightarrow{f} & \mathcal{O}_{\mathbb{P}^{n+1}} & \rightarrow & \mathcal{O}_X \rightarrow 0 \\ & & \downarrow f^{p-1}\mathcal{F} & & \downarrow \mathcal{F} & & \downarrow \mathcal{F} \\ 0 & \rightarrow & \mathcal{O}_{\mathbb{P}^{n+1}}(-d) & \xrightarrow{f} & \mathcal{O}_{\mathbb{P}^{n+1}} & \rightarrow & \mathcal{O}_X \rightarrow 0. \end{array}$$

The coboundary in the resulting long exact cohomology sequence allows to identify

$$H^n(X, \mathcal{O}_X) \cong H^{n+1}(\mathbb{P}^{n+1}, \mathcal{O}_{\mathbb{P}^{n+1}}(-d)),$$

so that the Frobenius \mathcal{F} on $H^n(X, \mathcal{O}_X)$ corresponds to the map on $H^{n+1}(\mathbb{P}^{n+1}, \mathcal{O}_{\mathbb{P}^{n+1}}(-d))$ induced by

$$0 \rightarrow \mathcal{O}_{\mathbb{P}^{n+1}}(-d) \xrightarrow{\mathcal{F}} \mathcal{O}_{\mathbb{P}^{n+1}}(-pd) \xrightarrow{f^{p-1}} \mathcal{O}_{\mathbb{P}^{n+1}}(-d) \rightarrow 0.$$

Computing Čech cohomology we find that Laurent monomials $x^{-u} = x_0^{-u_0} \dots x_{n+1}^{-u_{n+1}}$ where u runs through the set

$$U = \{u = (u_0, \dots, u_{n+1}) : u_i \in \mathbb{Z}_{\geq 1}, \sum_{i=0}^{n+1} u_i = d\} \tag{2}$$

form a basis in $H^{n+1}(\mathbb{P}^{n+1}, \mathcal{O}_{\mathbb{P}^{n+1}}(-d))$ and the Hasse–Witt matrix is given in this basis by

$$\mathcal{F}_{u,v \in U} = \text{the coefficient of } x^{pv-u} \text{ in } f(x)^{p-1}. \tag{3}$$

Suppose one starts from a polynomial f in characteristic 0, e.g. with coefficients in \mathbb{Z} . In my talk at the MATRIX institute in Creswick I presented a construction which lifts (3) to a matrix with entries in \mathbb{Z}_p whose characteristic polynomial conjecturally gives the p -adic unit root part of the zeta function attached to the middle cohomology of X . The proofs and a few evidences for the conjecture can be found in [4].

2 Main Results

We study a sequence of matrices which generalize (3). Let R be a commutative characteristic 0 ring, that is the natural map $R \rightarrow R \otimes \mathbb{Q}$ is an embedding. Let $f \in R[x_1^{\pm 1}, \dots, x_N^{\pm 1}]$ be a Laurent polynomial in N variables. If $f(x) = \sum_u a_u x^u$, $a_u \in R$, the *Newton polytope* $\Delta(f) \subset \mathbb{R}^N$ is the convex hull of the finite set $\{u : a_u \neq 0\}$. Consider the set of internal integral points $J = \Delta(f)^o \cap \mathbb{Z}^N$, where $\Delta(f)^o$ denotes the topological interior of the Newton polytope. Let $g = \#J$ be the number of internal integral points in the Newton polytope, which we assume to be positive. Consider the following sequence of $g \times g$ matrices with entries in R whose rows and columns are indexed by the elements of J :

$$(\beta_m)_{u,v \in J} = \text{the coefficient of } x^{(m+1)v-u} \text{ in } f(x)^m. \tag{4}$$

By convention, β_0 is the identity matrix. We shall consider arithmetic properties of the sequence $\{\beta_m; m \geq 0\}$.

Let us fix a prime number p . We restrict our attention to the sub-sequence $\{\alpha_s = \beta_{p^s-1}; s \geq 0\}$. The entries of these matrices are then given by

$$(\alpha_s)_{u,v \in J} = \text{the coefficient of } x^{p^s v-u} \text{ in } f(x)^{p^s-1}.$$

Notice that when R/pR is a finite field and f is a homogeneous polynomial of degree d such that its reduction modulo p defines a smooth hypersurface, then U in (2) coincides with J (with $N = n+2$) and $\alpha_1 = \beta_{p-1}$ modulo p is the Hasse–Witt matrix.

Theorem 1 *Assume that the ring R is endowed with a p th power Frobenius endomorphism, that is a ring endomorphism $\sigma : R \rightarrow R$ satisfying $\sigma(a) \equiv a^p \pmod p$ for all $a \in R$. Then for every s*

$$\alpha_s \equiv \alpha_1 \cdot \sigma(\alpha_1) \cdot \dots \cdot \sigma^{s-1}(\alpha_1) \pmod p. \tag{5}$$

If α_1 is invertible modulo p then for every $s \geq 1$ one has congruences

$$\alpha_{s+1} \cdot \sigma(\alpha_s)^{-1} \equiv \alpha_s \cdot \sigma(\alpha_{s-1})^{-1} \pmod{p^s} \tag{6}$$

and

$$D(\alpha_s) \cdot \alpha_s^{-1} \equiv D(\alpha_{s-1}) \cdot \alpha_{s-1}^{-1} \pmod{p^s} \tag{7}$$

for any derivation $D : R \rightarrow R$.

Congruence (5) shows that $\alpha_s \pmod p$ are iterates of the Hasse–Witt operation whenever the latter is defined. It also implies that when α_1 is invertible modulo p then all α_s are invertible modulo p and hence also modulo p^s for all s . Therefore

statements (6) and (7) make sense. We remark that analogous congruences also hold when one multiplies by the inverse matrices on the left, that is we can prove that $\sigma(\alpha_s)^{-1} \cdot \alpha_{s+1} \equiv \sigma(\alpha_{s-1})^{-1} \cdot \alpha_s$ and $\alpha_s^{-1} \cdot D(\alpha_s) \equiv \alpha_{s-1}^{-1} \cdot D(\alpha_{s-1}) \pmod{p^s}$.

Our results are related to the topic of the workshop because when $\Delta(f)$ is a reflexive polytope (in this case $g = 1$), the toric hypersurface of zeroes of f can be compactified to a Calabi–Yau variety. Congruence (6) then generalizes so called Dwork’s congruences (see [5, 6]) and (7) seems to be new even in the Calabi–Yau case.

Theorem 1 implies existence of the p -adic limits

$$F = \lim_{s \rightarrow \infty} \alpha_{s+1} \cdot \sigma(\alpha_s)^{-1} \tag{8}$$

and

$$\nabla_D = \lim_{s \rightarrow \infty} D(\alpha_s) \cdot \alpha_s^{-1} \quad \text{for every derivation } D \in \text{Der}(R). \tag{9}$$

These $g \times g$ matrices have entries in the p -adic closure $\widehat{R} = \varprojlim R/p^s R$. Note that $F \equiv \alpha_1 \pmod{p}$. We are currently working on identifying the limiting matrices (8) and (9) with the Frobenius and Gauss–Manin connection on the slope 0 part of a crystal attached to the Laurent polynomial f . This fact was conjectured in [4] based on several examples and analogy with the congruences for expansion coefficients of differential forms stated in [7]. The progress in this project is due to our collaboration with Frits Beukers, which started at the MATRIX institute. I am also grateful to Frits for the series of extremely helpful lectures on Dwork cohomology which he gave during the first week of the program.

Matrices (4) showed up in [8] as coefficients of the logarithms of explicit coordinatizations of the Artin–Mazur formal group laws of projective hypersurfaces and complete intersections. Under certain conditions (e.g. R is the ring of integers of the unramified extension of \mathbb{Q}_p of degree a and f is a homogeneous polynomial whose reduction modulo p defines a non-singular hypersurface X/\mathbb{F}_{p^a}) one can combine (6) with the generalized Atkin and Swinnerton-Dyer congruences in [9], which yields that the eigenvalues of $\Phi = F \cdot \sigma(F) \cdot \dots \cdot \sigma^{a-1}(F)$ are p -adic unit eigenvalues of the Frobenius operator on the middle crystalline cohomology of X (see [4, Section 5]).

Our second result is the following integrality theorem for formal group laws attached to a Laurent polynomial. Its proof is based on explicit congruences (similar to those in Theorem 1) and Hazewinkel’s functional equation lemma (see [4, Section 4]).

Theorem 2 *Let J be either the set $\Delta(f) \cap \mathbb{Z}^N$ of all integral points in the Newton polytope of f or the subset of internal integral points $\Delta(f)^\circ \cap \mathbb{Z}^N$. Assume that J is non-empty and let $g = \#J$. Consider the sequence of matrices $\beta_m \in \text{Mat}_{g \times g}(R)$, $m \geq 0$ given by formula (4) and define a g -tuple of formal powers series*

$l(\tau) = (l_u(\tau))_{u \in J}$ in g variables $\tau = (\tau_v)_{v \in J}$ as

$$l(\tau) = \sum_{m=1}^{\infty} \frac{1}{m} \beta_{m-1} \tau^m .$$

Consider the g -dimensional formal group law $G_f(\tau, \tau') = l^{-1}(l(\tau) + l(\tau'))$ with coefficients in $R \otimes \mathbb{Q}$.

Let p be a prime number. If R can be endowed with a p th power Frobenius endomorphism then G_f is p -integral, that is $G_f \in R_{(p)}[[\tau, \tau']]$ where $R_{(p)} = R \otimes_{\mathbb{Z}_{(p)}} \mathbb{Q}$ is the subring of $R \otimes \mathbb{Q}$ formed by elements without p in the denominator.

Note that if one can define a Frobenius endomorphism on R for every prime p then Theorem 2 implies that $G_f \in R[[\tau, \tau']]$ because the subring $\cap_p R_{(p)} \subset R \otimes \mathbb{Q}$ coincides with R . For example, rings \mathbb{Z} and $\mathbb{Z}[t]$ are of this type: one can take the Frobenius endomorphism to be the identity on \mathbb{Z} and $h(t) \mapsto h(t^p)$ on $\mathbb{Z}[t]$.

Acknowledgements My work is supported by the National Science Centre of Poland, grant UMO-2016/21/B/ST1/03084. I am grateful to the European Mathematical Society and the MATRIX institute for sponsoring my participation in the program. I would like to mention that hospitality of the staff of the institute and MATRIX Family Fund made it possible for me to bring to Creswick my 3 month old daughter Helena.

References

1. Katz, N.: Une formule de congruence pour la fonction ζ , S.G.A. 7 II. Lecture Notes in Mathematics, vol. 340. Springer, Berlin (1973)
2. Dwork, B.: On the zeta function of a hypersurface II. Ann. Math. **80**, 227–299 (1964)
3. Koblitz, N.: p -adic variation of the zeta function over families of varieties defined over finite fields. Compos. Math. **31**(f. 2), 119–218 (1975)
4. Vlasenko, M.: Higher Hasse–Witt matrices. Indag. Math. **29**, 1411–1424 (2018)
5. Mellit, A., Vlasenko, M.: Dwork congruences for the constant terms of powers of a Laurent polynomial. Int. J. Number Theory **12**(2), 313–321 (2016)
6. Samol, K., van Straten, D.: Dwork congruences and reflexive polytopes. Ann. Math. Qué. **39**(2), 185–203 (2015)
7. Katz, N.: Internal reconstruction of unit-root F-crystals via expansion-coefficients. Annales scientifiques de l' É.N.S. 4e série, **18**(2), 245–285 (1985)
8. Stienstra, J.: Formal group laws arising from algebraic varieties. Am. J. Math. **109**(5), 907–925 (1987)
9. Stienstra, J.: Formal groups and congruences for L-functions. Am. J. Math. **109**(6), 1111–1127 (1987)

Triangular Modular Curves



John Voight

Abstract We consider certain generalizations of modular curves arising from congruence subgroups of triangle groups.

1 Triangle Groups

Let $a, b, c \in \mathbb{Z}_{\geq 2} \cup \{\infty\}$ satisfy $a \leq b \leq c$. Consider the triangle T with angles $\pi/a, \pi/b, \pi/c$ (with $\pi/\infty = 0$) in the space H , where H is the sphere, Euclidean plane, or hyperbolic plane according as the quantity $\chi(a, b, c) = 1/a + 1/b + 1/c - 1$ is positive, zero, or negative. Let τ_a, τ_b, τ_c be reflections in the sides of T and let $\Delta = \Delta(a, b, c)$ be the subgroup of orientation-preserving isometries in the group generated by the reflections: then Δ is generated by

$$\delta_a = \tau_b \tau_c, \quad \delta_b = \tau_c \tau_a, \quad \delta_c = \tau_a \tau_b$$

and has a presentation

$$\Delta = \langle \delta_a, \delta_b, \delta_c \mid \delta_a^a = \delta_b^b = \delta_c^c = \delta_a \delta_b \delta_c = 1 \rangle.$$

We call Δ a *triangle group*. The quotient

$$X = X(a, b, c; 1) = \Delta(a, b, c) \backslash H$$

is a complex Riemannian 1-orbifold of genus zero; it has as many punctures as occurrences of ∞ among a, b, c .

Example 1 We have $\Delta(2, 3, 3) \simeq A_4$, and the other spherical triangle groups (i.e., those with $\chi(a, b, c) > 0$) correspond to the Platonic solids. The Euclidean triangle

J. Voight (✉)
Dartmouth College, Hanover, NH, USA

groups are the familiar tessellation of the plane by triangles. We have $\Delta(2, 3, \infty) \simeq \text{PSL}_2(\mathbb{Z})$; and $\Delta(\infty, \infty, \infty) \simeq \Gamma(2)$, the free abelian group on two generators.

A uniformizer for X is expressed by an explicit ratio of ${}_2F_1$ -hypergeometric functions, with parameters given in terms of a, b, c . As a consequence, containments of triangle groups imply relations between ${}_2F_1$ -hypergeometric functions, with arguments given by Belyi maps. Moreover, the quotient is a moduli space for certain abelian varieties, often called *hypergeometric abelian varieties*: the values of the hypergeometric functions are periods of the *generalized Legendre curve*

$$y^N = x^A(1 - x)^B(1 - tx)^C$$

for certain integers A, B, C, N again given explicitly in terms of a, b, c .

The triangle group Δ is *arithmetic* if and only if it is commensurable with the units of reduced norm 1 in an order in a quaternion algebra over a number field (necessarily defined over a totally real field and ramified at all but one real place). There are only 85 arithmetic triangle groups, the list given by Takeuchi [4]; for these groups, the corresponding curve X is a Shimura curve.

2 Triangular Modular Curves

For the remaining nonarithmetic triangle groups, there is still a quaternion algebra! This observation was used by Cohen–Wolfart [2] in their work on transcendence of values of hypergeometric functions. This relationship can be interpreted geometrically: there is a finite map $X \rightarrow V$ where V is a quaternionic Shimura variety, a moduli space for abelian varieties with quaternionic multiplication, suitably interpreted. The dimension $\text{adim}(a, b, c)$ of V is given in terms of a, b, c ; we call it the *arithmetic dimension* of (a, b, c) . Nugent–Voight [3] have proven that for every t , the set $\{(a, b, c) : \text{adim}(a, b, c) = t\}$ is finite and effectively computable. For example, there are $148 + 16 = 164$ triples with arithmetic dimension 2.

Like with the modular curves, we now add level structure: we take a congruence subgroup $\Gamma(\mathfrak{P}) \leq \Gamma$ of the uniformizing group Γ for V , and we intersect

$$\Delta(\mathfrak{p}) = \Gamma(\mathfrak{P}) \cap \Delta.$$

By pullback, this gives a cover

$$\phi : X(\mathfrak{p}) = \Delta(\mathfrak{p}) \backslash H \rightarrow X(1);$$

this corresponds geometrically to adding level structure to the family of hypergeometric abelian varieties. Clark–Voight [1] have proven that the cover ϕ has Galois group $\text{PSL}_2(\mathbb{F}_\mathfrak{p})$ or $\text{PGL}_2(\mathbb{F}_\mathfrak{p})$ (cases distinguished by a Legendre symbol); moreover, the minimal field of definition of ϕ is explicitly given as an at most

quadratic extension of an explicitly given totally real abelian number field with controlled ramification.

We call these curves $X(\mathfrak{p})$ *triangular modular curves* as generalizations of the classical modular curves, and we expect that their study will be as richly rewarding for arithmetic geometers as the classical case.

Acknowledgement The author was supported by an NSF CAREER Award (DMS-1151047).

References

1. Clark, P.L., Voight, J.: Algebraic curves uniformized by congruence subgroups of triangle groups. *Trans. Am. Math. Soc.* **371**(1), 33–82 (2019)
2. Cohen, P., Wolfart, J.: Modular embeddings for some non-arithmetic Fuchsian groups. *Acta Arith.* **56**, 93–110 (1990)
3. Nugent, S., Voight, J.: On the arithmetic dimension of triangle groups. *Math. Comput.* **86**(306), 1979–2004 (2017)
4. Takeuchi, K.: Arithmetic triangle groups. *J. Math. Soc. Jpn.* **29**(1), 91–106 (1977)

Jacobi Sums and Hecke Grössencharacters



Mark Watkins

Abstract We give an extended abstract regarding our talk, and the associated Magma implementation of Jacobi sums and Hecke Grössencharacters. This builds upon seminal work of Weil (Trans Am Math Soc 73:487–495, 1952), and makes his construction explicitly computable, inherently relying on his upper bound for the conductor. Moreover, we can go slightly further than Weil by additionally allowing Kummer twists of the Jacobi sums. We also note the correspondence of these (twisted) Jacobi sums to tame prime information for hypergeometric motives.

Although our viewpoint and notation is derived from later work of Anderson, we do not use his formalism in any substantial way, and indeed the main thrust of all we do is already in Weil's work.

Let $\theta = \sum_j n_j \langle x_j \rangle \in \mathbf{Z}[\mathbf{Q}/\mathbf{Z}]^0$ be an integral linear combination of nonzero elements $x_j \in \mathbf{Q}/\mathbf{Z}$ as a formal sum, with $\sum_j n_j x_j = 0$. We put m for the least common multiple of the denominators of the x_j , and write $K_\theta \subseteq \mathbf{Q}(\zeta_m)$ for the subfield corresponding by Galois theory to modding out $(\mathbf{Z}/m\mathbf{Z})^*$ by those u for which the scaling $u \circ \theta = \sum_j n_j \langle ux_j \rangle$ is equal to θ . Letting α be a nontrivial additive character modulo p and recalling the Gauss sum of a multiplicative character ψ on \mathbf{F}_p^\times as

$$G_\alpha(\psi) = - \sum_{x \in \mathbf{F}_p^\times} \psi(x) \alpha(\mathrm{Tr} x),$$

for ideals \mathfrak{p} of $\mathbf{Q}(\zeta_m)$ we define the Jacobi sum

$$J_\theta(\mathfrak{p}) = \prod_j G_\alpha(\chi_{\mathfrak{p}}^{m x_j})^{n_j}$$

M. Watkins (✉)

University of Sydney, School of Mathematics and Statistics, Sydney, NSW, Australia
e-mail: watkins@maths.usyd.edu.au

where this is independent of the choice of α and $\chi_{\mathfrak{p}}$ is the power residue symbol

$$\chi_{\mathfrak{p}}(x) = \left(\frac{x}{\mathfrak{p}}\right)_m \equiv x^{(q-1)/m} \pmod{\mathfrak{p}}.$$

where q is the norm of \mathfrak{p} . One then has a partial L -function

$$L_{\theta}^*(s) = \prod_{\mathfrak{p}} \left(1 - \frac{J_{\theta}(\mathfrak{p})}{q^s}\right)^{-1},$$

where the product is over $\mathfrak{p} \nmid m$ in K_{θ} .

In a 1952 paper [3], Weil associates a Grössencharacter to such a Jacobi sum L -function, and in particular gets an upper bound on the modulus. This gives us an algorithm in principle to compute said Grössencharacter, which has been implemented in the Magma computer algebra system [1, 2]. Briefly, one first determines the field of definition K_{θ} of the Grössencharacter as above, and then the ∞ -type in a similar manner. The upper bound on the modulus then makes it a finite problem to recognize the correct twist in the Hecke character group (the dual of the ray class group), and by computing $J_{\theta}(\mathfrak{p})$ at sufficiently many primes of small norm we can isolate the desired twist. The possibility of including Kummer twists of the θ was not considered directly by Weil, but fits easily into the above framework.

The resulting Jacobi sum machinery also helps explain the tame prime behavior of hypergeometric motives, in particular giving the Euler factors when the inertia corresponding to such primes is in fact trivialized. As an example, for the quintic 3-fold at (say) $t = t_0 \cdot p^5$ with $p \equiv 1 \pmod{5}$, the Euler factor corresponds to a Grössencharacter over $\mathbf{Q}(\zeta_5)$, with the precise twist varying with t_0 .

This is joint work with David Roberts and Fernando Rodriguez Villegas.

References

1. Bosma, W., Cannon, J.J., Fieker, C., Steel, A. (eds.): Handbook of Magma functions, Edition 2.22, Chapter 132 (Hypergeometric Motives) (2016)
2. Watkins, M.: Computing with Hecke Grössencharacters. Publications mathématiques de Besançon. **2011**, 119–135 (2011)
3. Weil, A.: Jacobi Sums as “Grössencharaktere”. Trans. Am. Math. Soc. **73**, 487–495 (1952)

Special Values of Hypergeometric Functions and Periods of CM Elliptic Curves



Yifan Yang

Abstract Let $X = X_0^6(1)/W_6$ be the quotient of the Shimura curve $X_0^6(1)$ by all the Atkin-Lehner involutions. By realizing modular forms on X in two ways, one in terms of hypergeometric functions and the other in terms of Borcherds forms, and using Schofer's formula for values of Borcherds forms at CM-points, we obtain special values of certain hypergeometric functions in terms of periods of elliptic curves over $\overline{\mathbb{Q}}$ with complex multiplication.

Let $X_0^D(N)$ be the Shimura curve associated to an Eichler order of level N in an indefinite quaternion algebra of discriminant D over \mathbb{Q} . When $D = 1$, the Shimura curve $X_0^1(N)$ is just the classical modular curve $X_0(N)$ and there are many different constructions of modular forms on $X_0(N)$ in literature, such as Eisenstein series, Dedekind eta functions, Poincare series, theta series, and etc. These explicit constructions provide practical tools for solving problems related to classical modular curves. On the other hand, when $D \neq 1$, because of the lack of cusps, most of the methods for classical modular curves cannot possibly be extended to the case of general Shimura curves. However, in recent years, there have been several methods for Shimura curve emerging in literature, such as the method of Yang [9] realizing modular forms on a Shimura curve of genus zero in terms of solutions of its Schwarzian differential equation, the method of Voight and Willis [7] for computing power series expansions of modular forms, the method of Nelson [4] for computing values of modular forms using explicit Shimizu lifting [8], and the method of Elkies [1] via $K3$ surfaces. Finally, there is a powerful method that realizes modular forms on Shimura curves as Borcherds forms. To make the method of Borcherds forms useful in practice, one would employ Schofer's formula [5] for values of Borcherds forms at CM-points (see [2] for sample computation). In [3],

Y. Yang (✉)

Department of Applied Mathematics, National Chiao Tung University, Hsinchu, Taiwan

National Center for Theoretical Sciences, Hsinchu, Taiwan

e-mail: yfyang@math.nctu.edu.tw

we developed a systematic method to construct Borcherds forms and determined the equations of all hyperelliptic Shimura curves $X_0^D(N)$ using Schofer’s formula.

In [11], by combining the method of Schwarzian differential equations and the method of Borcherds forms, we obtain some intriguing evaluations of hypergeometric functions, such as

$${}_2F_1\left(\frac{1}{24}, \frac{5}{24}; \frac{3}{4}; -\frac{3^7 \cdot 7^4}{2^{10} \cdot 5^6}\right) = \frac{1}{2} \sqrt[4]{10} \sqrt{7 + \sqrt{43}} \frac{\omega_{-43}}{\omega_{-4}} \tag{1}$$

and

$${}_3F_2\left(\frac{1}{3}, \frac{1}{2}, \frac{2}{3}; \frac{3}{4}, \frac{5}{4}; -\frac{3^7 \cdot 7^4}{2^{10} \cdot 5^6}\right) = \frac{100}{21} \omega_{-43}^2, \tag{2}$$

where for a negative fundamental discriminant d , we let

$$\omega_d = \frac{1}{\sqrt{|d|}} \prod_{a=1}^{|d|-1} \Gamma\left(\frac{a}{|d|}\right)^{\chi_d(a)\mu_d/4h_d}$$

be the Chowla-Selberg period. Here χ_d is the Kronecker character associated to $\mathbb{Q}(\sqrt{d})$, μ_d is the number of roots of unity in $\mathbb{Q}(\sqrt{d})$, and h_d is the class number of $\mathbb{Q}(\sqrt{d})$. Note that if E is an elliptic curve over $\overline{\mathbb{Q}}$ with complex multiplication by $\mathbb{Q}(\sqrt{d})$, then its periods are algebraic multiples of $\sqrt{\pi}\omega_d$. We now explain the origin of such evaluations.

Assume that $t(\tau)$ is a modular function on $X_0^D(N)$ that takes algebraic values at all CM-points. Then according to Shimura [6, Theorem 7.1] and Yoshida [12, Theorem 1.2 and (1.4) of Chapter 3], the value of $t'(\tau)$ at a CM-point of discriminant d is an algebraic multiple of ω_d^2 . Here we choose $t(\tau)$ to be the Hauptmodul of $X = X_0^6(1)/W_6$, the quotient of $X_0^6(1)$ by all the Atkin-Lehner involutions, that takes values 0, 1, and ∞ at the CM-points of discriminants -4 , -24 , and -3 , respectively. Now the Schwarzian differential equation of X is essentially a hypergeometric differential equation (see [9]), which means that all (meromorphic) modular forms on X can be expressed in terms of hypergeometric functions. In particular, we have

$$t'(\tau) = \frac{2t^{1/4}(1-t)^{1/2}}{Ci} \left({}_2F_1\left(\frac{1}{24}, \frac{5}{24}; \frac{3}{4}; t\right) - C_2 {}_2F_1\left(\frac{7}{24}, \frac{11}{24}; \frac{5}{4}; t\right) \right)^2,$$

where $C = -1/\sqrt[4]{12}\omega_{-4}^2$ (see Lemma 8 of [10]). Manipulating this identity and recalling the result of Shimura and Yoshida above, we find that at a CM-point τ_d of discriminant d , we have

$${}_2F_1\left(\frac{1}{24}, \frac{5}{24}; \frac{3}{4}; t(\tau_d)\right) \in \frac{\omega_d}{\omega_{-4}} \cdot \overline{\mathbb{Q}}$$

and

$${}_3F_2\left(\frac{1}{3}, \frac{1}{2}, \frac{2}{3}; \frac{3}{4}, \frac{5}{4}; t(\tau_d)\right) \in \omega_d^2 \cdot \overline{\mathbb{Q}}.$$

This explains the algebraicity of values of hypergeometric functions at singular moduli. To determine actual values, we use theory of Borcherds forms.

In [9], we find that the one-dimensional space of modular forms of weight 8 on X is spanned by

$$\left({}_2F_1\left(\frac{1}{24}, \frac{5}{24}; \frac{3}{4}; t\right) + \frac{1}{\sqrt[4]{12}\omega_{-4}^2} t^{1/4} {}_2F_1\left(\frac{7}{24}, \frac{11}{24}; \frac{5}{4}; t\right) \right)^8.$$

On the other hand, we can construct a Borcherds form Ψ of weight 8. As the space of modular forms has dimension 1, these two modular forms must be scalar multiples of each other. Evaluating Ψ at the CM-point of discriminant -4 using Schofer’s formula, we can determine the ratio of the two modular forms. Then evaluating at other CM-points, we obtain the special values of hypergeometric functions.

References

1. Elkies, N.D.: Shimura curve computations via $K3$ surfaces of Néron-Severi rank at least 19. In: Algorithmic Number Theory. Lecture Notes in Computer Science, vol. 5011, pp. 196–211. Springer, Berlin (2008)
2. Errthum, E.: Singular moduli of Shimura curves. *Can. J. Math.* **63**(4), 826–861 (2011)
3. Guo, J.W., Yang, Y.: Equations of hyperelliptic Shimura curves. *Compos. Math.* **153**, 1–40 (2017)
4. Nelson, P.D.: Evaluating modular forms on Shimura curves. *Math. Comput.* **84**(295), 2471–2503 (2015)
5. Schofer, J.: Borcherds forms and generalizations of singular moduli. *J. Reine Angew. Math.* **629**, 1–36 (2009)
6. Shimura, G.: Automorphic forms and the periods of abelian varieties. *J. Math. Soc. Jpn.* **31**(3), 561–592 (1979)
7. Voight, J., Willis, J.: Computing power series expansions of modular forms. In: Computations with Modular Forms. *Contrib. Math. Comput. Sci.* vol. 6, pp. 331–361. Springer, Cham (2014)
8. Watson, T.C.: Rankin triple products and quantum chaos. ProQuest LLC, Ann Arbor, MI (2002). Thesis (Ph.D.), Princeton University
9. Yang, Y.: Schwarzian differential equations and Hecke eigenforms on Shimura curves. *Compos. Math.* **149**(1), 1–31 (2013)
10. Yang, Y.: Ramanujan-type identities for Shimura curves. *Isr. J. Math.* **214**(2), 699–731 (2016)
11. Yang, Y.: Special values of hypergeometric functions and periods of CM elliptic curves. *Trans. Am. Math. Soc.* **370**, 6433–6467 (2018)
12. Yoshida, H.: Absolute CM-periods. *Mathematical Surveys and Monographs*, vol. 106. American Mathematical Society, Providence (2003)

CY-Operators and L-Functions



Duco van Straten

Abstract This is a write up of a talk given at the MATRIX conference at Creswick in 2017 (to be precise, on Friday, January 20, 2017). It reports on work in progress with P. CANDELAS and X. DE LA OSSA. The aim of that work is to determine, under certain conditions, the local Euler factors of the L-functions of the fibres of a family of varieties *without recourse to the equations* of the varieties in question, but solely from the associated *Picard–Fuchs equation*.

1 Introduction

It is very honourable to speak the last words in this nice conference; surely these words are not the last on hypergeometrics, but rather represents a further exploration into *Transhypergeometria*, the unknown land of our dreams. I will report on joint work in progress with CANDELAS and DE LA OSSA [9]. I will start with some motivation.

2 Elliptic Curves Versus Rigid Calabi–Yau Threefolds

Elliptic curves and rigid Calabi–Yau manifolds share many common features. As a topological space, an elliptic curve is isomorphic to $S^1 \times S^1$ and a rigid Calabi–Yau threefold is a bit like $S^3 \times S^3$, at least what its third cohomology is concerned. On the arithmetic level, an elliptic curve E defined over \mathbb{Q} determines a two dimensional motive $H^1(E)$ and in a similar way a rigid Calabi–Yau threefold X defined over \mathbb{Q} produces a two dimensional motive $H^3(X)$. There are Hodge and

D. van Straten (✉)
Johannes Gutenberg-Universität Mainz, Mainz, Germany
e-mail: straten@mathematik.uni-mainz.de

p -adic realisations, giving rise to L -functions that come from classical modular forms for some $\Gamma_0(N)$.

Space	Motive	Hodge	Frobenius	Weil	Hecke
E/\mathbb{Q}	$H^1(E)$	0 1 1 0	$T^2 - a_p T + p$	$ a_p \leq p^{1/2}$	$L(H^1(E)) = L(f), f \in S_2(\Gamma_0(N))$
X/\mathbb{Q}	$H^3(E)$	1 0 0 1	$T^2 - a_p T + p^3$	$ a_p \leq p^{3/2}$	$L(H^3(X)) = L(f), f \in S_4(\Gamma_0(N))$

By the great theorem of WILES [35, 36] we know that all elliptic curves over \mathbb{Q} are modular, and by further development of these methods, it was shown that rigid Calabi–Yau threefolds defined over \mathbb{Q} are also modular [14, 17].

However, there are also big differences between these two cases. Elliptic curves depend on a single modulus and form nice families. Classical normal forms are provided by the *Legendre family*

$$L_\lambda : y^2 = x(x - 1)(x - \lambda)$$

or the *Hesse family*

$$H_\lambda : x^3 + y^3 + z^3 + \lambda xyz = 0,$$

where λ is the parameter.

On the other hand, as by definition $h^{1,2} = 0$, rigid Calabi–Yau spaces do not admit any non-trivial deformations, and their occurrence is sporadic. No general description or construction is known for them. We refer to [23, 37] for an overview of the exciting bestiary.

Question

Which weight four cusps forms appear as modular form of rigid Calabi–Yau manifolds?

For example, as can be seen from consulting [23], there are many different rigid Calabi–Yau varieties leading to the weight four cusp form for $\Gamma_0(6)$, but I do not know of any rigid Calabi–Yau threefold realising the weight four cusp form for $\Gamma_0(7)$.

2.1 How Can Rigid Varieties Appear in a Pencil?

Let us look at an example. The famous *Schoen quintic* X_1 studied in [29] is the degree 5 hypersurface in \mathbb{P}^4 given by the equation

$$X_1 : x_1^5 + x_2^5 + x_3^5 + x_4^5 + x_5^5 = 5x_1x_2x_3x_4x_5.$$

It is easily seen to have the 125 points

$$x_i^5 = 1, \quad x_1x_2x_3x_4x_5 = 1$$

as nodal singularities. There exists a small resolution $\pi : X \rightarrow X_1$ that replaces each node by a projective line \mathbb{P}^1 . X is a rigid Calabi–Yau threefold: the infinitesimal deformations of X can be identified with the infinitesimal deformations of X_1 for which the nodes lift, which are none. For small prime numbers the Euler factors of the L -function can be determined counting points of X_1 and correcting these counts to get the numbers of points of the resolved manifold X . As the Galois representation is determined by finitely many Euler factors, it was found that the $L(H^3(X_1)) = L(f)$ for some $f \in S_4(\Gamma_0(25))$, which was identified by C. SCHOEN.

Now note that the quintic X_1 (and not X) is a member of the even more famous *Dwork pencil*

$$X_\psi : x_1^5 + x_2^5 + x_3^5 + x_4^5 + x_5^5 = 5\psi x_1 x_2 x_3 x_4 x_5$$

that stands at the beginning of the mirror symmetry story, for which we refer to [8, 11, 24, 33]. The third cohomology of X_ψ is the direct sum of two pieces

$$H^3(X_\psi) = V \oplus F.$$

Here the part F has Hodge numbers 0 100 100 0, and the part V has Hodge numbers 1 1 1 1. The Picard–Fuchs equation for this part leads to the hypergeometric differential equation

$$\mathcal{D} := \Theta^4 - 5^5 t (\Theta + \frac{1}{5})(\Theta + \frac{2}{5})(\Theta + \frac{3}{5})(\Theta + \frac{4}{5}), \quad t = 1/(5\psi)^5, \quad \Theta = t \frac{d}{dt},$$

which describes a *variation of Hodge structures* (VHS) over $S := \mathbb{P}^1 \setminus \{0, 1/5^5, \infty\}$. At the three singular points these Hodge structures degenerate into *mixed Hodge structures* (MHS) [30]. We refer to [25] for a detailed account of (mixed) Hodge theory. Quite generally, the Jordan structure of the local monodromy determines the weight filtration. At $t = 0$ we have a so-called MUM-point, the monodromy has a maximal Jordan block. The mixed Hodge diamond looks like

$$\begin{array}{ccccccc} & & & & & & 1 \\ & & & & & & 0 & 0 \\ & & & & & & 0 & 1 & 0 \\ & & & & & & 0 & 0 & 0 & 0 \\ & & & & & & 0 & 1 & 0 \\ & & & & & & 0 & 0 \\ & & & & & & 1 \end{array}$$

(The weight is equal to the height in the diagram, counted by putting lowest row at height zero; the operator N shifts two steps downwards.) The limiting mixed Hodge structure is an iterated extension of Tate Hodge structures and it leads to

the extension data described in [13] that are equivalent to the so-called *instanton numbers* computed in [8].

At $t = 1/5^5$ there is a single Jordan block of size 2 (a *C*-point in the terminology of [31]). The mixed Hodge-diamond for H^3 looks like:

$$\begin{array}{ccccccc}
 & & & & 0 & & \\
 & & & & 0 & 0 & \\
 & & & & 0 & 1 & 0 \\
 & & & 1 & 0 & 0 & 1 \\
 & & & 0 & 1 & 0 & \\
 & & & 0 & 0 & & \\
 & & & 0 & & &
 \end{array}$$

So we see that the motive $Gr_3^W H^3$ is like that of a rigid Calabi–Yau.

There is one further possible degeneration of a $(1, 1, 1, 1)$ -VHS, that does not appear in this family, namely where there are two Jordan blocks of size 2 (a *K*-point in the terminology of [31]). The mixed Hodge diamond for H^3 now looks like

$$\begin{array}{ccccccc}
 & & & & 0 & & \\
 & & & & 0 & 0 & \\
 & & & & 1 & 0 & 1 \\
 & & & 0 & 0 & 0 & 0 \\
 & & & 1 & 0 & 1 & \\
 & & & 0 & 0 & & \\
 & & & 0 & & &
 \end{array}$$

So $Gr_2^W H^3$ is a $(1, 0, 1)$ -Hodge structure that looks like the one appearing for *K* 3-surfaces with Picard number 20.

One of the motivations to look at general motivic $(1, 1, 1, 1)$ -variations over $S = \mathbb{P}^1 \setminus \Sigma$ is the natural appearance of weight four and weight three cusp forms for $\Gamma_0(N)$ at the boundary points $\Sigma \subset \mathbb{P}^1$. Such motivic $(1, 1, 1, 1)$ -variations are expected to arise from *Calabi–Yau operators*.

3 Calabi–Yau Operators

Calabi–Yau operators, as understood in [3] and [31], are operators ‘like’ \mathcal{P} . First of all, they are *fourth order Fuchsian* differential operators

$$\mathcal{P} \in \mathbb{C}[t, \Theta], \quad \Theta = t \frac{d}{dt}$$

that are *symplectic* and have 0 as a MUM-point. If we look at it from the point of view of differential operators, it is rather easy to satisfy these conditions, for example by looking at operators of the form

$$\mathcal{P} = \Theta^2 P \Theta^2 + \Theta Q \Theta + R,$$

where P, Q, R are any polynomials with $P(0) = 1$. In order to classify as a *Calabi–Yau operator*, one has to complement these easy conditions with further arithmetical conditions that are supposed to hold if the operator is a Picard–Fuchs operator of a 1-parameter family of Calabi–Yau varieties defined over \mathbb{Q} . In [3] the following *integrality conditions* were put forward and used to define Calabi–Yau operators.

I. The holomorphic solution $\phi_0(t)$ has an integral power-series expansion:

$$\phi_0(t) \in \mathbb{Z}[[t]].$$

II. The q -coordinate has an integral power series expansion

$$q(t) \in \mathbb{Z}[[t]].$$

III. The normalised instanton numbers become integral

$$n_0 := 1, \quad n_1, \quad n_2, \dots, n_d, \dots$$

after multiplication by a common denominator.

Furthermore, the case where all $n_d = 0, d \geq 1$ is considered as *trivial*, as in that case \mathcal{P} is the third symmetric power of a second order operator. In fact, it is more natural to have coefficients in $\mathbb{Z}[\frac{1}{N}]$, so to allow denominators involving a finite set of *bad primes*. Currently more than 500 operators are known that seem to satisfy these three conditions (see [2, 4, 10]), but condition III is not proven to hold in a single case. The first condition should already imply that the operator is of geometric origin, see [5]. There are many examples of operators that satisfy I, but not II. In a good number of cases integrality of the q -coordinate have been proven [12, 21]. For some time it was expected that condition III was implied by I and II, until MICHAEL BOGNER [6] found an operator that satisfies I and II, but for which III appears to fail. There exists an unpublished paper [34] in which it is claimed that Picard–Fuchs operators coming from families of Calabi–Yau varieties indeed satisfy these three arithmetical conditions.

Of course, one can also look at differential operators of order different from four, and try to single out a particular nice sub-class of Calabi–Yau operators of arbitrary order. For an account, we refer to [1, 6] and [7].

A particular nice example is operator **AESZ 34**

$$\Theta^4 - t(35 \Theta^4 + 70 \Theta^3 + 63 \Theta^2 + 28 \Theta + 5) + t^2(\Theta + 1)(259 \Theta^2 + 518 \Theta + 235) - 5^3 t^3(\Theta + 1)^2(\Theta + 2)^2$$

that was reported to us long ago by VERRILL [32]. It turned up prominently at this conference, as it is associated to the five-fold banana FEYNMAN graph. As such, it is part of a very nice series of Calabi–Yau operators that exist for all orders. Its *Riemann symbol* (see [18, 19]) is

$$\left\{ \begin{array}{cccccc} 0 & 1/25 & 1/9 & 1 & \infty \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 2 \\ 0 & 2 & 2 & 2 & 2 \end{array} \right\}.$$

and the holomorphic solution has an expansion of the form

$$\phi_0(t) = \sum_{n=0}^{\infty} A_n t^n, \quad A_n := \sum_{i+j+k+l+m=n} \left(\frac{n!}{i!j!k!l!m!} \right)^2.$$

As for all Calabi–Yau operators, there is a unique *Frobenius basis* of solutions around 0 of the form

$$\begin{aligned} \phi_0(t) &= f_0(t) \\ \phi_1(t) &= \log(t)\phi_0(t) + f_1(t) \\ \phi_2(t) &= \log(t)^2\phi_0(t) + 2\log(t)\phi_1(t) + f_2(t) \\ \phi_3(t) &= \log(t)^3\phi_0(t) + 3\log(t)^2\phi_1(t) + 3\log(t)\phi_2(t) + f_3(t) \end{aligned}$$

where $f_0(t) \in \mathbb{Z}[[t]]$, $f_i(t) \in t\mathbb{Q}[[t]]$ ($i = 1, 2, 3$).

The points $1/25, 1/9, 1$ are *C*-points: there appears a single logarithm ‘between’ the two equal exponents. The point ∞ is a *K*-point: there are two logarithms, again between the two pairs of equal exponents. At each of the conifold points should appear a weight four modular form of some level, at ∞ there is a weight three modular form.

4 Euler Factors from Picard–Fuchs Operators

It has been known from the work of DWORK [15, 16] that there is a very tight link between the Frobenius operator and the Picard–Fuchs operator in a family of varieties. For the sake of concreteness, let us consider as before a family Y_t of Calabi–Yau threefolds defined over \mathbb{Q} with a MUM-point at 0 and let us fix a prime p . Then the Frobenius operator

$$F := F_p \in \text{Aut}(H^3(Y_t))$$

has a characteristic polynomial $P(T) = \det(T - F)$ of the form

$$T^4 + aT^3 + bpT^2 + ap^3T + p^6 \in \mathbb{Z}[T],$$

where

$$a = a_p(t) = \text{Tr}(F), \quad b = b_p(t) = (\text{Tr}(F^2) - \text{Tr}(F)^2)/2p.$$

from which we get the local Euler factor

$$1 + ap^{-s} + bp^{1-2s} + ap^{3-3s} + p^{6-4s}$$

for the L -function of $H^3(Y_t)$.

4.1 Unit Root Method

Let us suppose that the Frobenius polynomial is irreducible, but factors over \mathbb{Z}_p as

$$(T - u)(T - v)(T - p^3/v)(T - p^3/u) \in \mathbb{Z}_p[T]$$

with $\text{ord}_p(u) = 0, \text{ord}_p(v) = 1$. Then u is called the *unit-root* and according to DWORK [16], this unit root $u = u(t)$ can be computed from the holomorphic solution $\phi_0(t)$ using p -adic analytic continuation of

$$\frac{\phi_0(t)}{\phi_0(t^p)}$$

and evaluation at Teichmüller lift \tilde{t} of $t \in \mathbb{P}^1$ (avoiding singular and supersingular values of t .) Dwork's unit-root method has been clarified by KATZ [20] by formulating it in terms of *crystals*. In her thesis, SAMOL [26] used this method to compute Euler factors for many families of Calabi–Yau varieties, using only the Picard–Fuchs equation. One of the important discoveries she made was that in many cases the method even worked at the singular points of the differential equation, and thus managed to determine weight four forms attached to C -points of Calabi–Yau operators [27]. The explicit control of the p -adic analytic continuation can sometimes be obtained from *Dwork congruences* on the coefficients A_n of the holomorphic solution. In the context of Calabi–Yau varieties defined by Laurent polynomials such Dwork congruences can be shown to hold [22, 28].

4.2 Deformation Method

The type of crystals we are considering are defined over a ring R , which is a certain two-dimensional regular local sub-ring of $\mathbb{Z}_p[[t]]$. On R there are two operations: the derivation

$$\Theta : R \longrightarrow R, \quad a \mapsto t \frac{\partial a}{\partial t}$$

and the lifted Frobenius map

$$\sigma : R \longrightarrow R, \quad a(t) \mapsto a(t^p).$$

One has

$$\Theta \circ \sigma = p \sigma \circ \Theta.$$

We will consider a free R -module of rank four H , a non-degenerate symplectic pairing

$$\langle -, - \rangle : H \times H \longrightarrow R$$

and two operations

$$\nabla : H \longrightarrow H, \quad F : H \longrightarrow H$$

that we call the *Gauss-Manin* and *Frobenius*. The operator ∇ a connection, so is supposed to satisfy the appropriate Leibniz rule, whereas F is σ -linear. These three structures are required to satisfy the following compatibilities

- (i) $\Theta \langle x, y \rangle = \langle \nabla x, y \rangle + \langle x, \nabla y \rangle$.
- (ii) $p^3 \langle x, y \rangle = \langle Fx, Fy \rangle$.
- (iii) $\nabla F = p F \nabla$.

Furthermore, we will have a Hodge-filtration

$$Fil^3 \subset Fil^2 \subset Fil^1 \subset Fil^0 = H$$

with

$$\nabla(Fil^i) \subset Fil^{i-1}, \quad F(Fil^i) \subset p^i H.$$

The first part of the structure may be called a *polarised F -crystal*, including the filtration makes us speak about a *polarised divisible Hodge F -crystals*

(Fontaine-Lafaille crystals), we will call it a *CY-crystal* for short. Let us try to associate such a structure to a differential operator of the form

$$\mathcal{P} := \Theta^4 + tP_1(\Theta) + t^2P_2(\Theta) + \dots + t^rP_r(\Theta).$$

For this, we write everything out in MATRIX-form. We let

$$H := \sum_{i=0}^3 R\phi_i,$$

where the ϕ_i are abstract basis vectors, that behave with respect to differentiation as the Frobenius basis of \mathcal{P} . Writing out the action of Θ on them, we can construct the companion matrix $A(t)$ for the connection ∇ on H corresponding to \mathcal{P} :

$$\nabla = t \frac{d}{dt} - A(t),$$

where $A(t)$ is of the form

$$A(t) = \begin{pmatrix} 0 & 0 & 0 & * \\ 1 & 0 & 0 & * \\ 0 & 1 & 0 & * \\ 0 & 0 & 1 & * \end{pmatrix} = A_0 + A_1t + A_2t^2 + \dots + A_rt^r \in \mathbb{Q}[t]^{4 \times 4}.$$

Because of the MUM-condition, we have

$$A_0 = N = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

The matrix Σ of the symplectic form at $t = 0$ can be taken to be of the form

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix}.$$

We now write the Frobenius matrix in a series

$$F = F(t) = F_0 + F_1t + F_2t^2 + \dots$$

The above conditions, especially the Griffiths transversality and divisibility, lead to a very specific form for the constant term F_0 :

$$F_0 = \begin{pmatrix} \xi & 0 & 0 & 0 \\ p\alpha & \xi p & 0 & 0 \\ p^2\beta & p^2\alpha & \xi p^2 & 0 \\ p^3\gamma & p^3\beta & p^3\alpha & \xi p^3 \end{pmatrix},$$

where $\xi^2 = 1$ and $\xi\beta = \alpha^2/2$. One give an explicit formula for the series $F(t)$ as

$$F(t) = E(t^p)^{-1} F_0 E(t) \in \mathbb{Q}[t]^{4 \times 4},$$

where the matrix $E(t)$ is a modification of the *fundamental matrix* for the differential equation

$$\tilde{E}_{jk} = \Theta^k \phi_j = \begin{pmatrix} \phi_0 & \Theta(\phi_0) & \Theta^2(\phi_0) & \Theta^3(\phi_0) \\ \phi_1 & \Theta(\phi_1) & \Theta^2(\phi_1) & \Theta^3(\phi_1) \\ \phi_2 & \Theta(\phi_2) & \Theta^2(\phi_2) & \Theta^3(\phi_2) \\ \phi_3 & \Theta(\phi_3) & \Theta^2(\phi_3) & \Theta^3(\phi_3) \end{pmatrix} \in \mathbb{Q}[[t]][\log t]^{4 \times 4}.$$

This matrix reduces mod t to $E_{jk}^0 := \Theta^k \log^j(t)/j!$ and we set

$$E := (E^0)^{-1} \tilde{E} = \text{“}\tilde{E}\Big|_{\log(t)=0}\text{”}.$$

In all examples we have computed so far, we could make the following

Observations

- All terms of the series $F(t)$ are p -adically integral (depending linearly on α, β, γ .)
- One can write

$$F(t) = \frac{\varphi(t)}{\Delta(t)^{p-1}} \text{ mod } p^3,$$

where $\varphi(t) \in (\mathbb{Z}/p^3)[t]^{4 \times 4}$ is a polynomial matrix and $\Delta(t)$ is the discriminant of the operator \mathcal{P} .

- The poles cancel at all singularities of \mathcal{P} , except for the apparent singularities. So if \mathcal{P} does not have apparent singularities, the matrix $F(t) \text{ mod } p^3$ is in fact polynomial.
- We can ‘trivially’ read off

$$a(t) = -\text{Tr} F(t) \text{ mod } p^3, \quad b(t) = (\text{Tr}(F(t)^2) - \text{Tr}(F(t))^2)/2p \text{ mod } p^3$$

and these do not depend on the choice of α, β, γ (this was already observed in [26].) This suffices to determine the local Euler factor at p for $p \geq 5$.

Using this, we can compute Euler factors even at the singular points, as long as they are not apparent singularities. In particular, it works at the conifold points and we do not have to care about super-singular behaviour. For example, for the above mentioned operator **AESZ 34** one finds characteristic polynomials of Frobenius of the form

$$T(T - p\chi(p))(T^2 - a_p T + p^3)$$

for some character χ . We find

	1/25	1/9	1
a_7	32	-16	-16
a_{11}	-60	12	12
a_{13}	-34	38	38
a_{17}	42	-126	-126

So we recognise, using the table in [23], the weight four cusp forms 6/1 for $\Gamma_0(6)$ at $t = 1$ and $t = 1/9$, and the form 30/1 for $\Gamma_0(30)$ at $t = 1/25$.

4.3 Lifting to Higher Order

Let us set $\alpha = \beta = 0$ and $\xi = 1$, but keep γ as a parameter. It appears that there is a *unique choice* for $\gamma \pmod p$ for which

$$F(t) = \frac{\varphi(t)_{m(p)}}{\Delta(t)^{\ell(p)}} \pmod{p^4}$$

where $\ell(p)$ is a small slope linear functions of p and $\varphi(t)_{m(p)}$ is a matrix-polynomial of small degree $m(p)$ linear in p . For all other choices of γ this structure seems to get lost. By playing the same game modulo p^5, p^6, p^7 , etc, we can determine a number γ modulo p^2, p^3 , etc. Continuing this way, we obtain a well-defined p -adic number γ that goes into the Frobenius matrix at the MUM-point:

$$F_0 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & p & 0 & 0 \\ 0 & 0 & p^2 & 0 \\ \gamma & 0 & 0 & p^3 \end{pmatrix}$$

For the quintic and $p = 11$ one finds

$$\gamma = 2 + 2 \cdot 11 + 3 \cdot 11^2 + 7 \cdot 11^3 + 5 \cdot 11^4 + 5 \cdot 11^5 + 6 \cdot 11^6 + \dots$$

Recall the relation between the p -adic $\zeta(3)$ and the p -adic gamma function:

$$-2\zeta_p(3) = \log \Gamma_p'''(0) = \Gamma_p'''(0) - \Gamma_p'(0)^3.$$

The following marvellous miracle seems to take place:

Observation

- $\gamma = r \cdot \zeta_p(3)$.
- $r = c_3(X)/d$, where d is the degree of the mirror manifold.

For the quintic $r = 200/5 = 40$. This is reminiscent of a very similar matrix describing the hermitian form $\langle x, \bar{y} \rangle$, where $\bar{\cdot}$ is the Frobenius at ∞ , that is, complex conjugation, and the real $\zeta(3)$ appears at the place of $\zeta_p(3)$!

This is the end of the talk and of the conference, but I feel it is the beginning of something great.

During the conference we have seen some *amazing maths*, we had a *great taam*, it was really a *naas workshop*.

Acknowledgement A great thank to the organisers Masha, Ling and Wadim!

References

1. Almkvist, G.: The Art of Finding Calabi–Yau Differential Equations. Dedicated to the 90-th Birthday of Lars Gårding. Gems in Experimental Mathematics. Contemporary Mathematics, vol. 517, pp. 1–18. American Mathematical Society, Providence (2010)
2. Almkvist, G., van Straten, D.: Update on Calabi–Yau operators, in preparation
3. Almkvist, G., Zudilin, W.: Differential equations, mirror maps and zeta values. In: Mirror symmetry. V, 481515, AMS/IP Studies in Advanced Mathematics, vol. 38. American Mathematical Society, Providence (2006)
4. Almkvist, G., van Enckevort, C., van Straten, D., Zudilin, W.: Tables of Calabi–Yau operators (2010). arXiv:math/0507430
5. André, Y.: G-functions and geometry. Aspects of Mathematics, E13. Friedr. Vieweg & Sohn, Braunschweig (1989)
6. Bogner, M.: On differential operators of Calabi–Yau type, Thesis, Mainz (2012)
7. Bogner, M.: Algebraic characterization of differential operators of Calabi–Yau type. arXiv:math.AG/1304.5434
8. Candelas, P., de la Ossa, X., Green, P., Parkes, L.: An exactly soluble superconformal theory from a mirror pair of Calabi–Yau manifolds. Phys. Lett. B **258**(1–2), 118–126 (1991)
9. Candelas, P., de la Ossa, X., van Straten, D.: Local Euler factors from Picard–Fuchs equations, in preparation
10. Calabi–Yau Database, Version 2.0 (<http://www2.mathematik.uni-mainz.de/CYequations/db/>), Version 3.0 (<http://cydb.mathematik.uni-mainz.de>)

11. Cox, D., Katz, S.: Mirror symmetry and algebraic geometry. *Mathematical Surveys and Monographs*, vol. 68. American Mathematical Society, Providence (1999)
12. Delaygue, E.: A criterion for the integrality of the Taylor coefficients of mirror maps in several variables. *Adv. Math.* **234**, 414–452 (2013)
13. Deligne, P.: Local Behavior of Hodge Structures at Infinity, Mirror Symmetry II. *Studies in Advanced Mathematics*, vol. 1, pp. 683–699. American Mathematical Society/International Press, Providence (1997)
14. Dieulefait, L.: On the modularity of rigid Calabi–Yau threefolds: epilogue. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)* **377**, 44–49 (2010), *Issledovaniya po Teorii Chisel.* 10, 44–49, 241; translation in *J. Math. Sci. (N.Y.)* 171 (2010), no. 6, 725–727
15. Dwork, B.: On the zeta function of a hypersurface, III. *Ann. Math. (2)* **83**, 457–519 (1966)
16. Dwork, B.: p -adic cycles. *Inst. Hautes Études Sci. Publ. Math.* **37**, 27–115 (1969)
17. Gouvêa, F., Yui, N.: Rigid Calabi–Yau threefolds over \mathbb{Q} are modular. *Expo. Math.* **29**(1), 142–149 (2011)
18. Gray, J.: Fuchs and the theory of differential equations. *Bull. Am. Math. Soc. (New series)* **10**, 1–26 (1984)
19. Ince, E.: *Ordinary Differential Equations*. Longmans, Green and Co. Ltd., New York (1927)
20. Katz, N.: *Travaux de Dwork (French) Séminaire Bourbaki, 24ème année (1971/1972)*, Exp. No. 409. *Lecture Notes in Mathematics*, vol. 317, pp. 167–200. Springer, Berlin (1973)
21. Krattenthaler, C.: On the integrality of Taylor coefficients of mirror maps. *Duke Math. J.* **151**, 175–218 (2010)
22. Mellit, A., Vlasenko, M.: Dwork’s congruences for the constant terms of powers of a Laurent polynomial. *Int. J. Number Theory* **12**(2), 313–321 (2016)
23. Meyer, C.: *Modular Calabi–Yau threefolds*. *Fields Institute Monographs*, vol. 22, American Mathematical Society, Providence (2005)
24. Morrison, D.R.: *Geometric Aspects of Mirror Symmetry. Mathematics Unlimited 2001 and Beyond*, pp. 899–918. Springer, Berlin (2001)
25. Peters, C., Steenbrink, J.: *Mixed Hodge Structures. Ergebnisse der Mathematik und ihre Grenzgebiete, 3. Folge*, Springer, Berlin (2008)
26. Samol, K.: *Frobenius Polynomial for Calabi–Yau Equations*, Thesis, Mainz (2010)
27. Samol, K., van Straten, D.: Frobenius polynomials for Calabi–Yau equations. *Commun. Number Theory Phys.* **2**(3), 537–561 (2008)
28. Samol, K., van Straten, D.: Dwork congruences and reflexive polytopes. *Ann. Math. Qué.* **39**, 185–203 (2015)
29. Schoen, C.: On the geometry of a special determinantal hypersurface associated to the Mumford-Horrocks vector bundle. *J. Reine Angew. Math.* **364**, 85–111 (1986)
30. Steenbrink, J.: Limits of Hodge structures. *Invent. Math.* **31**, 229–257 (1976)
31. van Straten, D.: Calabi–Yau operators. arXiv:1704.00164 [math.AG]
32. Verrill, H.: Root lattices and pencils of varieties. *J. Math. Kyoto Univ.* **36**(2), 423–446 (1996)
33. Voisin, C.: *Symmetrie miroir. Panorama et synthèse*, Soc. Math. France, translated as: *Mirror Symmetry*. American Mathematical Society, Providence (1999)
34. Vologodsky, V.: On the N -Integrality of instanton numbers (2008). arXiv:0707.4617 [math.AG]
35. Wiles, A.: Modular elliptic curves and Fermat’s last theorem. *Ann. Math. (2)* **141**(3), 443–551 (1995)
36. Wiles, A.: Modular forms, elliptic curves, and Fermat’s last theorem. In: *Proceedings of the International Congress of Mathematicians (Zürich, 1994)*, vols. 1, 2, pp. 243–245. Birkhäuser, Basel (1995)
37. Yui, N.: *Modularity of Calabi–Yau Varieties: 2011 and Beyond. Arithmetic and Geometry of K3 Surfaces and Calabi–Yau Threefolds*. *Fields Institute Communications*, vol. 67, pp. 101–139. Springer, New York (2013)

A Matrix Theoretic Derivation of the Kalman Filter



Johnathan M. Bardsley

Abstract The Kalman filter is a data analysis method used in a wide range of engineering and applied mathematics problems. This paper presents a matrix-theoretic derivation of the method in the linear model, Gaussian measurement error case. Standard derivations of the Kalman filter make use of probabilistic notation and arguments, whereas we make use, primarily, of methods from numerical linear algebra. In addition to the standard Kalman filter, we derive an equivalent variational (optimization-based) formulation, as well as the extended Kalman filter for nonlinear problems.

1 Introduction

We start with the standard linear model with Gaussian measurement error:

$$\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{e}, \quad (1)$$

where $\mathbf{b} \in \mathbb{R}^m$ is measured data; $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a known observation matrix; $\mathbf{x} \in \mathbb{R}^n$ is the unknown parameter vector to be estimated; $\mathbf{e} \in \mathbb{R}^m$ is a zero-mean Gaussian random vector with covariance matrix \mathbf{C}_e , which we denote by $\mathbf{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_e)$; and $\mathbf{x} \in \mathbb{R}^n$ is the unknown vector that is to be estimated.

The standard technique for estimating \mathbf{x} , known as *least squares estimation*, was developed by Gauss in his study of planetary motion [1]. The extension of least squares estimation to the case when the unknown \mathbf{x} is also assumed to be a Gaussian random vector, which will be the case for us, is known as *minimum variance estimation* [4].

J. M. Bardsley (✉)

Department of Mathematical Sciences, University of Montana, Missoula, MT, USA

e-mail: bardsleyj@mso.umt.edu

© Springer Nature Switzerland AG 2019

D. R. Wood et al. (eds.), 2017 *MATRIX Annals*, MATRIX Book Series 2,

https://doi.org/10.1007/978-3-030-04161-8_44

505

In the study of time varying phenomena, it is natural to generalize (1) as follows:

$$\mathbf{x}_k = \mathbf{M}_k \mathbf{x}_{k-1} + \mathbf{E}_k, \quad (2)$$

$$\mathbf{b}_k = \mathbf{A}_k \mathbf{x}_k + \mathbf{e}_k, \quad (3)$$

where Eq. (3) is defined analogous to (1) for each k ; and in (2), $\mathbf{M}_k \in \mathbb{R}^{n \times n}$ is the known evolution matrix, \mathbf{x}_{k-1} is a Gaussian random vector, and $\mathbf{E}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{\mathbf{E}_k})$. The Kalman filter [2, 5] is the extension of minimum variance (and hence least squares) estimation to the problem of sequentially estimating $\{\mathbf{x}_1, \mathbf{x}_2, \dots\}$ given data $\{\mathbf{b}_1, \mathbf{b}_2, \dots\}$ arising from the model in (2), (3). For the interested reader, a discussion of the progression of ideas from Gauss to Kalman is the subject of the excellent paper [3].

This paper is organized as follows. First, in Sect. 2, we present the basic statistical definitions and results that we will need in our later discussion. In Sect. 3, we define the minimum variance estimator, which we then apply to (2), (3) to derive the Kalman filter in Sect. 4. Finally, we present an equivalent formulation of the Kalman filter, which we call the variational Kalman filter, as well as the extended Kalman filter for the case when (2), (3) contain nonlinear evolution and/or observation operators.

2 Statistical Preliminaries

Let $\mathbf{x} = (x_1, \dots, x_n)^T$ be a random vector with $E(x_i)$ the mean of x_i and $E((x_i - \mu_i)^2)$, where $\mu_i = E(x_i)$, its variance. The mean of \mathbf{x} is then defined $E(\mathbf{x}) = (E(x_1), \dots, E(x_n))^T$, while the $n \times n$ covariance matrix of \mathbf{x} is defined

$$[\text{cov}(\mathbf{x})]_{ij} = E((x_i - \mu_i)(x_j - \mu_j)), \quad 1 \leq i, j \leq n.$$

Note that the diagonal of $\text{cov}(\mathbf{x})$ contains the variances of x_1, \dots, x_n , while the off diagonal elements contain the covariance values. Thus if x_i and x_j are independent $[\text{cov}(\mathbf{x})]_{ij} = 0$ for $i \neq j$.

The $n \times m$ cross correlation matrix of the random n -vector \mathbf{x} and m -vector \mathbf{y} , which we will denote $\mathbf{\Gamma}_{\mathbf{xy}}$, is defined

$$\mathbf{\Gamma}_{\mathbf{xy}} = E(\mathbf{xy}^T), \quad (4)$$

where $[E(\mathbf{xy}^T)]_{ij} = E(x_i y_j)$. If \mathbf{x} and \mathbf{y} are independent, then $\mathbf{\Gamma}_{\mathbf{xy}}$ is the zero matrix. Furthermore,

$$E(\mathbf{x}) = \mathbf{0} \quad \text{implies} \quad \mathbf{\Gamma}_{\mathbf{xx}} = \text{cov}(\mathbf{x}). \quad (5)$$

Finally, given an $m \times n$ matrix \mathbf{A} and a random n -vector \mathbf{x} , it is not difficult to show that

$$\text{cov}(\mathbf{Ax}) = \mathbf{A}\text{cov}(\mathbf{x})\mathbf{A}^T. \quad (6)$$

We end these preliminary comments with the probability density function of primary interest to us in this paper, the Gaussian distribution. If \mathbf{b} is an $n \times 1$ Gaussian random vector, then its probability density function has the form

$$p_{\mathbf{b}}(\mathbf{b}; \boldsymbol{\mu}, \mathbf{C}) = \frac{1}{\sqrt{(2\pi)^n \det(\mathbf{C})}} \exp\left(-\frac{1}{2}(\mathbf{b} - \boldsymbol{\mu})^T \mathbf{C}^{-1}(\mathbf{b} - \boldsymbol{\mu})\right), \quad (7)$$

where $\boldsymbol{\mu} \in \mathbb{R}^n$ is the mean of \mathbf{b} ; \mathbf{C} is an $n \times n$ symmetric positive definite covariance matrix of \mathbf{b} ; and $\det(\cdot)$ denotes matrix determinant. As above, we will use the notation $\mathbf{b} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ in this case. For more details on introductory mathematical statistics, see one of many introductory mathematics statistics texts.

3 Minimum Variance Estimation

First, we consider model (1). When \mathbf{x} is assumed to be deterministic, it is a standard exercise to show that if \mathbf{A} has full column rank, the least squares estimator is given by

$$\mathbf{x}^{ls} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}.$$

However, we are interested in the case in which $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_x)$. We assume, furthermore, that \mathbf{x} and \mathbf{e} are independent random variables. We now define the minimum variance estimator of \mathbf{x} .

Definition 1 Suppose \mathbf{b} is defined as in (1), $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_x)$, and \mathbf{e} and \mathbf{x} independent random vectors. Then the *minimum variance estimator* of \mathbf{x} given \mathbf{b} has the form

$$\mathbf{x}^{est} = \hat{\mathbf{B}}\mathbf{b},$$

where $\hat{\mathbf{B}} \in \mathbb{R}^{n \times m}$ solves the optimization problem

$$\hat{\mathbf{B}} = \arg \min_{\mathbf{B} \in \mathbb{R}^{n \times m}} E \left(\|\mathbf{B}\mathbf{b} - \mathbf{x}\|_2^2 \right).$$

Because our model (1) is a linear model with Gaussian measurement error, $\hat{\mathbf{B}}$ has an elegant closed form, as described in the following theorem.

Theorem 1 *If $\Gamma_{\mathbf{bb}}$ is invertible, then the minimum variance estimator of \mathbf{x} from \mathbf{b} is given by*

$$\mathbf{x}^{est} = (\Gamma_{\mathbf{xb}}\Gamma_{\mathbf{bb}}^{-1})\mathbf{b}.$$

Proof First, we note that

$$\begin{aligned} E(\|\mathbf{Bb} - \mathbf{x}\|_2^2) &= \text{trace}\left(E[(\mathbf{Bb} - \mathbf{x})(\mathbf{Bb} - \mathbf{x})^T]\right), \\ &= \text{trace}\left(\mathbf{B}E[\mathbf{bb}^T]\mathbf{B}^T - \mathbf{B}E[\mathbf{bx}^T] - E[\mathbf{xb}^T]\mathbf{B}^T + E[\mathbf{xx}^T]\right). \end{aligned}$$

Then, using the distributive property of the trace function and the identity

$$\frac{d}{d\mathbf{B}}\text{trace}(\mathbf{B}^T\mathbf{C}) = \left(\frac{d}{d\mathbf{B}}\text{trace}(\mathbf{BC})\right)^T = \mathbf{C},$$

we see that $dE(\|\mathbf{Bb} - \mathbf{x}\|_2^2)/d\mathbf{B} = \mathbf{0}$ when

$$\hat{\mathbf{B}} = \Gamma_{\mathbf{xb}}\Gamma_{\mathbf{bb}}^{-1},$$

which establishes the result.

In the context of (1), and given our assumptions stated above, we can obtain a more concrete form for the minimum variance estimator. In particular, we note that since \mathbf{x} and \mathbf{e} are assumed to be independent, $\Gamma_{\mathbf{xe}} = \Gamma_{\mathbf{ex}} = \mathbf{0}$. Hence, using (1), we obtain

$$\begin{aligned} \Gamma_{\mathbf{xb}} &= E[\mathbf{x}(\mathbf{Ax} + \mathbf{e})^T], \\ &= \Gamma_{\mathbf{xx}}\mathbf{A}^T. \end{aligned}$$

Similarly,

$$\begin{aligned} \Gamma_{\mathbf{bb}} &= E[(\mathbf{Ax} + \mathbf{e})(\mathbf{Ax} + \mathbf{e})^T], \\ &= \mathbf{A}\Gamma_{\mathbf{xx}}\mathbf{A}^T + \Gamma_{\mathbf{ee}}. \end{aligned}$$

Thus, since $\Gamma_{\mathbf{xx}} = \mathbf{C}_x$ and $\Gamma_{\mathbf{ee}} = \mathbf{C}_e$, the minimum variance estimator has the form

$$\begin{aligned} \mathbf{x}^{est} &= \hat{\mathbf{B}}\mathbf{b} \\ &= \mathbf{C}_x\mathbf{A}^T(\mathbf{A}\mathbf{C}_x\mathbf{A}^T + \mathbf{C}_e)^{-1}\mathbf{b}, \\ &= (\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{A} + \mathbf{C}_x^{-1})^{-1}\mathbf{A}^T\mathbf{C}_e^{-1}\mathbf{b}. \end{aligned} \tag{8}$$

We note, in passing, that (8) can also be expressed as

$$\mathbf{x}^{est} = \arg \min_{\mathbf{x}} \left\{ \|\mathbf{Ax} - \mathbf{b}\|_{\mathbf{C}_e^{-1}}^2 + \|\mathbf{x}\|_{\mathbf{C}_x^{-1}}^2 \right\}, \tag{9}$$

where “arg min” denotes “argument of the minimum” and $\|\mathbf{x}\|_{\mathbf{C}}^2 \stackrel{\text{def}}{=} \mathbf{x}^T \mathbf{C} \mathbf{x}$. This establishes a clear connection between minimum variance estimation and generalized Tikhonov regularization [4]. Note in particular that if $\mathbf{C}_e = \sigma_1^2 \mathbf{I}$ and $\mathbf{C}_x = \sigma_2^2 \mathbf{I}$, problem (9) can be equivalently expressed as

$$\mathbf{x}^{est} = \arg \min_{\mathbf{x}} \left\{ \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + (\sigma_1^2/\sigma_2^2) \|\mathbf{x}\|_2^2 \right\},$$

which has classical Tikhonov form. This formulation is also equivalent to maximum a posteriori (MAP) estimation.

4 The Kalman Filter

In the previous section, we considered the stationary linear model (1), but suppose our model now has the form (2), (3). Equation (2) is the equation of evolution for \mathbf{x}_k with \mathbf{M}_k the $n \times n$ linear evolution matrix, and $\mathbf{E}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{E_k})$. In Eq. (3), \mathbf{b}_k denotes the $m \times 1$ observed data, \mathbf{A}_k the $m \times n$ linear observation matrix, and $\mathbf{e}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_{e_k})$. In both equations, k denotes the time index.

The problem is to estimate \mathbf{x}_k at time k from \mathbf{b}_k and an estimate \mathbf{x}_{k-1}^{est} of the state at time $k - 1$. We assume $\mathbf{x}_{k-1}^{est} \sim \mathcal{N}(\mathbf{x}_{k-1}, \mathbf{C}_{k-1}^{est})$. To facilitate a more straightforward application of the result of Theorem 1, we rewrite (2), (3). First, define

$$\mathbf{x}_k^a = \mathbf{M}_k \mathbf{x}_{k-1}^{est} \tag{10}$$

$$\mathbf{z}_k = \mathbf{x}_k - \mathbf{x}_k^a, \tag{11}$$

$$\mathbf{r}_k = \mathbf{b}_k - \mathbf{A}_k \mathbf{x}_k^a. \tag{12}$$

Then, subtracting (10) from (2) and $\mathbf{A}_k \mathbf{x}_k^a$ from both sides of (3), and dropping the k dependence for notational simplicity, we obtain the stochastic linear equations

$$\mathbf{z} = \mathbf{M}(\mathbf{x} - \mathbf{x}^{est}) + \mathbf{E}, \tag{13}$$

$$\mathbf{r} = \mathbf{A}\mathbf{z} + \mathbf{e}. \tag{14}$$

The minimum variance estimator of \mathbf{z} from \mathbf{r} given (13), (14) is then given, via Theorem 1 (note that \mathbf{z} is a zero mean Gaussian random vector), by

$$\mathbf{z}^{est} = \mathbf{\Gamma}_{z\mathbf{r}} \mathbf{\Gamma}_{\mathbf{r}\mathbf{r}}^{-1} \mathbf{r}.$$

We assume that $\mathbf{x} - \mathbf{x}^{est}$ is independent of \mathbf{E} , and that $\mathbf{z} = \mathbf{x} - \mathbf{x}^a$ is independent of \mathbf{e} . Then, from (4), (5), (6), (13) and (14), we obtain

$$\begin{aligned}\Gamma_{zz} &= \mathbf{M}\mathbf{C}^{est}\mathbf{M}^T + \mathbf{C}_E \stackrel{\text{def}}{=} \mathbf{C}^a, \\ \Gamma_{zr} &= \mathbf{C}^a\mathbf{A}^T, \\ \Gamma_{rr} &= \mathbf{A}\mathbf{C}^a\mathbf{A}^T + \mathbf{C}_e.\end{aligned}\tag{15}$$

where \mathbf{C}^{est} and \mathbf{C}^a are the covariance matrices for \mathbf{x}^{est} and \mathbf{x}^a , respectively. Thus, finally, the minimum variance estimator of \mathbf{z} is given by

$$\mathbf{z}^{est} = \mathbf{C}^a\mathbf{A}^T(\mathbf{A}\mathbf{C}^a\mathbf{A}^T + \mathbf{C}_e)^{-1}\mathbf{r},\tag{16}$$

From (16) and (11) we then immediately obtain the Kalman Filter estimate of \mathbf{x} given by

$$\mathbf{x}_+^{est} = \mathbf{x}^a + \mathbf{H}(\mathbf{b} - \mathbf{A}\mathbf{x}^a),\tag{17}$$

where

$$\mathbf{H} = \mathbf{C}^a\mathbf{A}^T(\mathbf{A}\mathbf{C}^a\mathbf{A}^T + \mathbf{C}_e)^{-1}\tag{18}$$

is known as the Kalman Gain matrix.

Finally, in order to compute the covariance of \mathbf{x}_+^{est} , we note that by (17) and (3),

$$\mathbf{x}_+^{est} = (\mathbf{I} - \mathbf{H}\mathbf{A})\mathbf{x}^a + \mathbf{H}\mathbf{e} + \mathbf{H}\mathbf{A}\mathbf{x},$$

where \mathbf{x} is the true state. Given our assumptions and using (6), the covariance then takes the form

$$\mathbf{C}_+^{est} = (\mathbf{I} - \mathbf{H}\mathbf{A})\mathbf{C}^a(\mathbf{I} - \mathbf{H}\mathbf{A})^T + \mathbf{H}\mathbf{C}_e\mathbf{H}^T,$$

which can be rewritten, using the identity $\mathbf{H}\mathbf{C}_e\mathbf{H}^T = (\mathbf{I} - \mathbf{H}\mathbf{A})\mathbf{C}^a\mathbf{A}^T\mathbf{H}^T$, in the simplified form

$$\mathbf{C}_+^{est} = \mathbf{C}^a - \mathbf{H}\mathbf{A}\mathbf{C}^a.\tag{19}$$

Incorporating the k dependence again leads directly to the Kalman filter iteration.

The Kalman Filter Algorithm

Step 0: Select initial guess \mathbf{x}_0^{est} and covariance \mathbf{C}_0^{est} , and set $k = 1$.

Step 1: Compute the evolution model estimate and covariance:

- A. Compute $\mathbf{x}_k^a = \mathbf{M}_k \mathbf{x}_{k-1}^{est}$;
- B. Compute $\mathbf{C}_k^a = \mathbf{M}_k \mathbf{C}_{k-1}^{est} \mathbf{M}_k^T + \mathbf{C}_{E_k} := \mathbf{C}_k^a$.

Step 2: Compute the Kalman filter estimate and covariance:

- A. Compute the Kalman Gain $\mathbf{H}_k = \mathbf{C}_k^a \mathbf{A}_k^T (\mathbf{A}_k \mathbf{C}_k^a \mathbf{A}_k^T + \mathbf{C}_e)^{-1}$;
- B. Compute the estimate $\mathbf{x}_k^{est} = \mathbf{x}_k^a + \mathbf{H}_k (\mathbf{b}_k - \mathbf{A}_k \mathbf{x}_k^a)$;
- C. Compute the estimate covariance $\mathbf{C}_k^{est} = \mathbf{C}_k^a - \mathbf{H}_k \mathbf{A}_k \mathbf{C}_k^a$.

Step 3: Update $k := k + 1$ and return to Step 1.

4.1 A Variational Formulation of the Kalman Filter

As in the stationary case (see (8), (9)), we can rewrite Eq. (16) in the form

$$\mathbf{z}^{est} = (\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A} + (\mathbf{C}^a)^{-1})^{-1} \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{r},$$

which, yields, using (11), the Kalman filter estimate

$$\begin{aligned} \mathbf{x}_+^{est} &= \mathbf{x}^a + [\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A} + (\mathbf{C}^a)^{-1}]^{-1} \mathbf{A}^T \mathbf{C}_e^{-1} (\mathbf{b} - \mathbf{A} \mathbf{x}^a), \\ &= \arg \min_{\mathbf{x}} \left\{ \ell(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{2} (\mathbf{b} - \mathbf{A} \mathbf{x})^T \mathbf{C}_e^{-1} (\mathbf{b} - \mathbf{A} \mathbf{x}) + \frac{1}{2} (\mathbf{x} - \mathbf{x}^a)^T (\mathbf{C}^a)^{-1} (\mathbf{x} - \mathbf{x}^a) \right\}. \end{aligned}$$

It can be shown using a Taylor series argument that

$$\mathbf{x}_+^{est} = \mathbf{x}^a - \nabla^2 \ell(\mathbf{x}^a)^{-1} \nabla \ell(\mathbf{x}^a), \quad (20)$$

where $\nabla \ell$ and $\nabla^2 \ell$ denote the gradient and Hessian of ℓ respectively, and are given by

$$\begin{aligned} \nabla \ell(\mathbf{x}) &= \mathbf{A}^T \mathbf{C}_e^{-1} (\mathbf{b} - \mathbf{A} \mathbf{x}) + (\mathbf{C}^a)^{-1} (\mathbf{x} - \mathbf{x}^a), \\ \nabla^2 \ell(\mathbf{x}) &= \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A} + (\mathbf{C}^a)^{-1}. \end{aligned}$$

By the matrix inversion lemma, we have

$$(\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A} + (\mathbf{C}^a)^{-1})^{-1} = \mathbf{C}^a - \mathbf{C}^a \mathbf{A}^T (\mathbf{A} \mathbf{C}^a \mathbf{A}^T + \mathbf{C}_e)^{-1} \mathbf{A} \mathbf{C}^a.$$

Then from Eqs. (18) and (19), we obtain the interesting fact that

$$\mathbf{C}_+^{est} = \nabla^2 \ell(\mathbf{x})^{-1}. \quad (21)$$

This allows us to define the following equivalent formulation of the Kalman filter, which we call the variational Kalman filter.

The Variational Kalman Filter Algorithm

Step 0: Select initial guess \mathbf{x}_0^{est} and covariance \mathbf{C}_0^{est} , and set $k = 1$.

Step 1: Compute the evolution model estimate and covariance:

- A. Compute $\mathbf{x}_k^a = \mathbf{M}_k \mathbf{x}_{k-1}^{est}$;
- B. Compute $\mathbf{C}_k^a = \mathbf{M}_k \mathbf{C}_{k-1}^{est} \mathbf{M}_k^T + \mathbf{C}_{E_k} := \mathbf{C}_k^a$.

Step 2: Compute the Kalman filter estimate and covariance:

- A. Compute the estimate $\mathbf{x}_k^{est} = \arg \min_{\mathbf{x}} \ell(\mathbf{x})$;
- C. Compute the estimate covariance $\mathbf{C}_k^{est} = \nabla^2 \ell(\mathbf{x})^{-1}$.

Step 3: Update $k := k + 1$ and return to Step 1.

A natural question is, what is the use of this equivalent formulation of the Kalman filter? Theoretically there is no benefit gained in using the variational Kalman filter if the estimate and its covariance are computed exactly. However, with the variational approach, the filter estimate, and even its covariance, can be computed approximately using an iterative minimization method, such as conjugate gradient. This is particularly important for large-scale problems where the exact Kalman filter is prohibitively expensive to compute.

4.2 The Extended Kalman Filter

The extended Kalman filter is the extension of the Kalman filter when (2), (3) are replaced by

$$\mathbf{x}_k = \mathcal{M}(\mathbf{x}_{k-1}) + \mathbf{E}_k, \quad (22)$$

$$\mathbf{b}_k = \mathcal{A}(\mathbf{x}_k) + \mathbf{e}_k, \quad (23)$$

where \mathcal{M} and \mathcal{A} are (possibly) nonlinear functions. The extended Kalman filter is obtained by the following simple modification of either of the above algorithms: in Step 1, A use, instead, $\mathbf{x}_k^a = \mathcal{M}(\mathbf{x}_{k-1}^{est})$, and define

$$\mathbf{M}_k = \frac{\partial \mathcal{M}(\mathbf{x}_{k-1}^{est})}{\partial \mathbf{x}}, \quad \text{and} \quad \mathbf{A}_k = \frac{\partial \mathcal{A}(\mathbf{x}_k^a)}{\partial \mathbf{x}}, \quad (24)$$

where $\frac{\partial f}{\partial \mathbf{x}}$ denotes the Jacobian of f .

5 Conclusions

We have presented a derivation of the Kalman filter that utilizes matrix analysis techniques as well as the Bayesian statistical approach of minimum variance estimation. In addition, we presented an equivalent variational formulation, which we call the variational Kalman filter, as well as the extended Kalman filter for nonlinear problems.

References

1. Gauss, K.F.: *Theory of Motion of the Heavenly Bodies*. Dover, New York (1963)
2. Kalman, R.E.: A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng. Ser. D* **82**, 35–45 (1960)
3. Sorenson, H.W.: Least squares estimation: from Gauss to Kalman. *IEEE Spectr.* **7**, 63–68 (1970)
4. Vogel, C.R.: *Computational Methods for Inverse Problems*. SIAM, Philadelphia (2002)
5. Welch, G., Bishop, G.: An Introduction to the Kalman Filter. UNC Chapel Hill, Dept. of Computer Science Tech. Report, TR 95-041

Approximate Bayesian Computational Methods for the Inference of Unknown Parameters



Yuqin Ke and Tianhai Tian

Abstract Recent advances in biology, economics, engineering and physical sciences have generated a large number of mathematical models for describing the dynamics of complex systems. A key step in mathematical modelling is to estimate model parameters in order to realize experimental observations. However, it is difficult to derive the analytical density functions in the Bayesian methods for these mathematical models. During the last decade, approximate Bayesian computation (ABC) has been developed as a major method for the inference of parameters in mathematical models. A number of new methods have been designed to improve the efficiency and accuracy of ABC. Theoretical studies have also been conducted to investigate the convergence property of these methods. In addition, these methods have been applied to a wide range of deterministic and stochastic models. This chapter gives a brief review of the main ABC algorithms and various improvements.

1 Introduction

Since more and more natural and social science problems involve the uncertainty in observations, statistical models and parameter inference play an important role in the development of mathematical methods for studying real-world problems. In particular, the era of big data has generated huge amount of data whose volume is increasing at a very fast speed. Mathematical and statistical models are becoming more and more complex in terms of the network size and regulatory relationships. Thus effective and efficient methods are strongly needed to infer unknown parameters in these models in order to reduce the simulation errors against the experimental data.

There are two major types of inference methods, namely the optimization methods and Bayesian statistical methods. The optimization methods are designed

Y. Ke · T. Tian (✉)

School of Mathematical Sciences, Monash University, Clayton, VIC, Australia

e-mail: yuqin.ke@monash.edu; tianhai.tian@monash.edu

© Springer Nature Switzerland AG 2019

D. R. Wood et al. (eds.), *2017 MATRIX Annals*, MATRIX Book Series 2,

https://doi.org/10.1007/978-3-030-04161-8_45

515

to minimize an objective function by searching for parameters within a given parameter space in a directed manner. The inferred set of parameters produces the best fit to the experimental data [30]. A variety of effective approaches have been developed in recent years. Among them, the genetic algorithm is a popular and effective approach and has been widely applied to various models [49]. These methods all share two main ingredients: a cost function for specifying the distance between simulated data and experimental data and an optimization algorithm for searching for parameters in order to optimize the cost function. However, when the landscape of the cost function is complex, it is difficult for these methods to find the global optimum. To tackle this challenge, the global optimization methods have been proposed to explore the complex surfaces as widely as possible. Comparison studies have been conducted to examine the efficiency of several global optimization algorithms for the test models [21].

Compared with the optimization methods, the Bayesian inference methods can estimate the probability distributions of parameters by using the Bayes' rule to update the prior probability estimates. In addition, Bayesian methods are more robust in dealing with stochastic models and/or experimental data with noise [22, 54]. In recent years Bayesian methods have been successfully used in a diverse range of fields and provide the promise to applications [47]. The recent advances in approximate Bayesian computation (ABC) provide effective methods without any restriction on the requirement of the likelihood function [7, 16, 32, 41, 51]. This chapter provides a brief review for the recent development in ABC, including the rejection ABC, regression ABC, Markov chain Monte Carlo (MCMC) ABC and sequential Monte Carlo (SMC) ABC. We also discuss the relevant improvements and extensions of these methods, such as the choice of summary statistics.

2 Principle of Bayesian Inference

For the Bayesian inference problems, model parameters are treated as random quantities along with the observation data. The Bayesian inference involves the estimation of the posterior probability

$$p(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)} \propto p(y|\theta)\pi(\theta), \quad (1)$$

where y is the observation data and the parameter vector of the model is θ ($\theta \in \Theta \subseteq \mathbb{R}^q$, $q \geq 1$). In addition, $\pi(\theta)$ is the prior distribution representing the prior beliefs about the parameters under investigation, and $p(y|\theta)$ is a likelihood function of parameter θ . The marginal distribution, defined by

$$p(y) = \int_{\theta \in \Theta} p(y|\theta)\pi(\theta)d\theta \quad (2)$$

often involves a high-dimensional integral, and $p(\theta|y)$ is the posterior probability distribution which expresses the uncertainty regarding θ conditional on the observed experimental data y . All the Bayesian inference about θ will be based on the estimated $p(\theta|y)$. However, the integrations which produce the Bayesian quantities of interest (such as marginal posteriors, marginal moments, and probability intervals) can only be performed analytically when the density function $p(y|\theta)$ is available, which can be achieved only for relatively simple cases.

When the density function $p(y|\theta)$ is available, the classic Metropolis-Hasting algorithm is applied to find a Markov chain of the parameters, which is given below.

Algorithm 1 Metropolis-Hasting algorithm

Given the observation data y_{obs} , proposal distribution $q(\cdot)$, and an initial sample from the prior distribution $\theta^{(0)} \sim \pi(\theta)$.

At iteration $i \geq 0$

1. Generate a sample from the proposal distribution $\theta' \sim q(\theta|\theta^{(i)})$.
2. Draw a sample from the uniform distribution $\mu \sim U(0, 1)$ and calculate the ratio

$$\alpha = \min\left(1, \frac{\pi(\theta')p(y_{obs}|\theta')q(\theta^{(i)}|\theta')}{\pi(\theta^{(i)})p(y_{obs}|\theta^{(i)})q(\theta'|\theta^{(i)})}\right). \quad (3)$$

3. If $\mu \leq \alpha$, accept the sample as $\theta^{(i+1)} = \theta'$; otherwise reject the sample.
 4. Repeat steps 1 ~ 3 until the required number of posterior samples is obtained.
-

Based on the classic Metropolis-Hasting algorithm, a number of more sophisticated methods have been designed, such as the Markov chain Monte Carlo (MCMC), the importance sampling (IS), and the sequential Monte Carlo (SMC) [20, 42]. The MCMC sampling methods usually break a high-dimensional problem into a number of smaller dimensional problems and generate a sample of dependent or correlated draws which can be treated as a realization of a Markov chain with equilibrium distribution equal to $p(\theta|y)$. Once the convergence to $p(\theta|y)$ occurs, any subsequent simulated value can be viewed as a sample from $p(\theta|y)$ and all these samples are used to estimate the posterior quantities of interest.

An alternative approach is the Gibbs sampling if the marginal distribution of each parameter is available. In the basic version, the Gibbs sampling is a special case of the Metropolis-Hastings algorithm. However, in its extended versions, these methods can be considered as a framework for sampling each variable (or more generally, each group of variables) from a number of variables in turn. It can also be incorporated into the Metropolis-Hastings algorithm (or other methods) to implement in one or more sampling steps. The detail of this algorithm is given below.

Algorithm 2 Gibbs sampling

Given the observation data y_{obs} , and an initialize sample from the prior distribution $\theta^{(0)} = (\theta_1^{(0)}, \dots, \theta_p^{(0)})^T \sim \pi(\theta)$.

At iteration $i \geq 0$

1. Generate a sample for $\theta_1^{(i+1)}$ using

$$\theta_1^{(i+1)} \sim \pi(\theta_1 | \theta_2^{(i)}, \dots, \theta_p^{(i)}, y_{obs}).$$

2. For $j = 2, \dots, p - 1$, sample for $\theta_j^{(i+1)}$ using

$$\theta_j^{(i+1)} \sim \pi(\theta_j | \theta_1^{(i+1)}, \dots, \theta_{j-1}^{(i+1)}, \theta_{j+1}^{(i)}, \dots, \theta_p^{(i)}, y_{obs}).$$

3. Generate a sample for $\theta_p^{(i+1)}$ using

$$\theta_p^{(i+1)} \sim \pi(\theta_p | \theta_1^{(i+1)}, \theta_2^{(i+1)}, \dots, \theta_{p-1}^{(i+1)}, y_{obs}).$$

4. Repeat steps 1 ~ 3 until the required number of posterior samples is obtained.
-

Although these Bayesian inference methods are effective, they are based on the availability of the likelihood function. However, it may be difficult to derive the likelihood function directly for many complex models. For example, the analytical density function may not be available, or it may be expensive to calculate the likelihood. In some cases, the observed experimental data are insufficient to obtain a tractable likelihood. This intractability prohibits the direct implementation of a generic MCMC algorithm.

3 Rejection ABC Method

To deal with complex models without analytical likelihood, a number of algorithms have been developed during the past two decades, which are referred to as the likelihood-free inference or Approximate Bayesian Computation (ABC). The ABC method is based on the following intuition: namely if a sample of the unknown parameter produces the simulation that matches the observed dataset, this sample should be close to the exact value of the parameter. Conversely, if the simulated dataset differs from the observed data substantially, this sample should not be considered as the estimate of the parameter. Thus the method strongly relies on the metric to determine the distance between simulated dataset and observed dataset. In the late 90s, ABC was first introduced as a rejection technique bypassing the computation of the likelihood function [48]. Later, Pritchard et al. proposed a generalisation based on an approximation of the target [40]. In recent years, the ABC methods have been proposed with various improvements and have been applied to a

wide range of application fields, such as population genetics, ecology, epidemiology and systems biology [2, 3, 17, 25, 27].

The ABC spirit is based on the following algorithm [43].

Algorithm 3 Likelihood-free rejection sampling

Given the observation data y_{obs} , and prior distribution $\pi(\theta)$.

1. Generate a sample from the prior distribution $\theta' \sim \pi(\theta)$.
 2. Simulate the model using θ' to get a dataset $x \sim p(x|\theta')$.
 3. Accept the sample θ' if $x = y_{obs}$, otherwise reject it.
 4. Repeat the above steps until the required number of posterior samples is obtained.
-

In this paper Rubin just exhibited this algorithm as an intuitive way to understand the posterior distributions from a frequentist perspective rather than using it for inferring models where the likelihood function was not available [43]. Then Tavaré et al. proposed an implementation of the rejection algorithm for the Bayesian inference of parameters in population genetics. When the data are discrete and of low dimension, this algorithm is effective. However, the probability of acceptance for a sample is usually very low.

As mentioned earlier, the rejection algorithm is dependent on a metric to measure the distance between the simulation and observation data. For inference problems with continuous distributions, or the datasets are high dimensional, it may be necessary to use summary statistics to reduce the dimensionality. Pritchard et al. suggested the prototype rejection-ABC algorithm as follows in a population genetics setting [40].

Algorithm 4 Rejection ABC method

Given the observation data y_{obs} , prior distribution $\pi(\theta)$, summary statistics $s(\cdot)$, tolerance level $\varepsilon > 0$, and distance function $\rho(\cdot, \cdot)$

1. Generate a sample from the prior distribution $\theta' \sim \pi(\theta)$.
2. Simulate the model using θ' to get a dataset $x \sim p(x|\theta')$.
3. Calculate the distance between the simulation and experimental data $\rho(s(x), s_{obs})$ based on the given summary statistics $s(\cdot)$.
4. Accept the sample θ' if

$$\rho(s(x), s_{obs}) < \varepsilon.$$

Otherwise reject the sample.

5. Repeat steps 1~4 until the required number of posterior samples is obtained
-

The basic idea of ABC is to use summary statistics with a small tolerance to produce a good proximation of the posterior distribution. The output is the samples of parameters from the distribution $p(\theta|\rho(s(x), s_{obs}) \leq \varepsilon)$. The choice of summary statistics is very important which we will discuss later. In addition, the

tolerance ε in Algorithm 4 may determine the efficiency of ABC. The basic ABC rejection algorithm may result in long computing time when a prior distribution is far away from posterior distribution. In addition, there is no learning process in this algorithm; and thus no information could be obtained from the previous accepted samples of parameters. When the search space is complex, the convergence rate of this algorithm may be very slow.

4 Regression ABC

To improve the efficiency of the rejection-ABC algorithm, Beaumont et al. [6] introduced the regression approach by explicitly modeling the discrepancy between the simulated summary statistics and that of the observed data through the following algorithm.

Algorithm 5 Regression ABC

Given the observation data y_{obs} , prior distribution $\pi(\theta)$, summary statistics $s(\cdot)$, tolerance level ε , and distance function $\rho(\cdot)$.

1. Generate a sample from the prior distribution $\theta^{(i)} \sim \pi(\theta)$.
2. Simulate the model using $\theta^{(i)}$ to get a dataset $x^{(i)} \sim p(x|\theta^{(i)})$ and compute the summary statistics $s^{(i)} = s(x^{(i)})$.
3. Repeat steps 1 and 2, until N pairs $\{\theta^{(i)}, s^{(i)}\}$ are obtained.
4. Associate each pair $(\theta^{(i)}, s^{(i)})$ with a weight $\omega^{(i)} \propto K_\varepsilon(\rho(s^{(i)} - s_{obs}))$. The weighted kernel can be selected as:

$$K_\varepsilon(t) = \begin{cases} \varepsilon^{-1}(1 - (t/\varepsilon)^2) & t \leq \varepsilon, \\ 0 & t > \varepsilon. \end{cases}$$

5. Apply a regression model to the n points, which have nonzero weights to obtain an estimate of $E(\theta|s(x) = s^{(i)})$, denoted as $\hat{m}(s^{(i)})$.
6. Adjust each sample to

$$\theta^{*(i)} = \hat{m}(s_{obs}) + (\theta^{(i)} - \hat{m}(s^{(i)})).$$

7. Use $\{\theta^{*(i)}, \omega^{(i)}\}$ to approximate the posterior distribution.
-

Here the samples $\theta^{(i)}$ are adjusted with weights $\omega^{(i)} > 0$ to account for the difference between simulated summary statistics and that of the observed data. Beaumont et al. [6] suggested a local linear model in the region of s_{obs} , given by

$$\begin{aligned} \theta^{(i)} &= m(s^{(i)}) + e^{(i)}, \\ m(s^{(i)}) &= \alpha + \beta^T (s^{(i)} - s_{obs}), \end{aligned}$$

where $e^{(i)}$ are zero-mean random variates with common variance, $m(s)$ is the conditional expectation of θ given s .

In this approach, the choice of ε involves a bias-variance trade-off, namely the increase of ε will reduce the variance because of a larger sample size for fitting the regression. However, this will also increase bias arising from the departure from the linearity and homoscedasticity [8].

When the number of samples is not very large due to the computational constraints, the homoscedastic assumption is no longer valid, because the neighbourhood of samples where $\omega^{(i)} \neq 0$ is too large. Thus Blum et al. [9, 10] extended this algorithm to a nonlinear and heteroscedastic model, given by

$$\theta^{(i)} = m(s^{(i)}) + \sigma(s^{(i)})e^{(i)},$$

where $\sigma(s^{(i)}) = Var(\theta|s^{(i)})$ denotes the conditional variance. The variance is then estimated by using a second regression model for the logarithm of the squared residuals, given by

$$\log(\theta^{(i)} - \hat{m}(s^{(i)}))^2 = \log(\sigma(s^{(i)})) + \eta^{(i)},$$

where $\eta^{(i)}$ are independent, zero-mean variates with common variance. The parameter adjustment then can be performed as follows:

$$\theta^{*(i)} = \hat{m}(s_{obs}) + (\theta^{(i)} - \hat{m}(s^{(i)})) \times \frac{\hat{\sigma}(s_{obs})}{\hat{\sigma}(s^{(i)})}, \tag{4}$$

where $\hat{\sigma}(s)$ denotes the estimator of $\sigma(s)$. Here e plays the same role as for homoscedastic model, but it has more flexibility on deviations from homoscedasticity.

5 MCMC-ABC Algorithm

In the Rejection-ABC and Regression-ABC algorithms, parameter values are sampled from the prior distribution. Thus the acceptance rate may be low if the prior and posterior distributions are quite different. In fact, using samples from a non-informative prior is very inefficient because this scheme does not account for the data at the proposal stage and thus may lead to proposed values located in low posterior probability regions. To address this issue, Marjoram et al. [33] introduced the following MCMC-ABC algorithm.

Algorithm 6 MCMC-ABC algorithm

Given the observation data y_{obs} , summary statistics $s(\cdot)$, tolerance level ε , distance function $\rho(\cdot)$, and proposal distribution $q(\cdot)$.

Initialize the first sample from the prior distribution $\theta^{(0)} \sim \pi(\theta)$.

At iteration $i \geq 0$

1. Generate a sample from the proposal distribution $\theta' \sim q(\theta|\theta^{(i)})$.
2. Simulate the model using θ' to get a dataset $x \sim p(x|\theta')$.
3. Draw a sample from the uniform distribution $\mu \sim U(0, 1)$, and calculate the ratio

$$\alpha = \min(1, \frac{\pi(\theta')q(\theta^{(i)}|\theta')}{\pi(\theta^{(i)})q(\theta'|\theta^{(i)})} \times I(\rho(s(x), s_{obs}) \leq \varepsilon)).$$

Here $I(A)$ is an indicator function.

4. If $\mu \leq \alpha$, accept the sample $\theta^{(i+1)} = \theta'$; otherwise $\theta^{(i+1)} = \theta^{(i)}$.
 5. Repeat steps 1~4 until the required number of posterior samples is obtained.
-

This algorithm has a similar structure as that of the standard MCMC. Both algorithms use a proposal distribution and prior distribution to calculate the ratio. The difference is that the density function is used in MCMC for computing the ratio, while in MCMC-ABC we treat the ratio of density function as one if the simulation error satisfies the criterion. Thus the performance of MCMC-ABC strongly depends on the selection of proposal distribution and prior distribution.

A potential drawback of MCMC-ABC is the selection of tolerance level ε and proposal distribution $q(\theta|\theta^{(i)})$ that may lead to expensive pilot runs [26, 44]. The convergence property of the generated chain $(\theta^{(1)}, \dots, \theta^{(n)})$ is important because MCMC algorithm may suffer if the proposal distribution is poorly chosen [14]. A potential issue is that the chain may get stuck in a low probability region of the posterior and lead to a poor approximation [18]. Since the proposed sample θ' must meet two criteria, the rejection rate of the MCMC ABC may be extremely high.

6 SMC ABC

To tackle the challenges in MCMC-ABC, sequential Monte Carlo sampling techniques have been introduced to ABC. Sequential Monte Carlo sampling differs from the MCMC approach by using the technique of particle filtering. Rather than drawing one candidate sample θ' at a step, this algorithm considers a pool with a large number of samples $(\theta'_1, \dots, \theta'_N)$ simultaneously and treats each sample as a particle. Sisson et al. [45] proposed a method which embed ABC simulation steps in Sequential Monte Carlo algorithm based on the theoretical work in [15]. This method generates sample from a sequence of approximate ABC posteriors under successively smaller acceptance tolerances [4, 46, 50]. SMC-ABC concentrates on simulating a dataset from the parameter regions with relatively high acceptance

probabilities and can adapt tuning choices such as acceptance tolerances during the computation, which has potential advantages over the Rejection-ABC or MCMC-ABC. Here we illustrate the algorithm of Beaumont et al. [4]:

Algorithm 7 SMC-ABC

Given the observation data y_{obs} , summary statistics $s(\cdot)$, distance function $\rho(\cdot)$, tolerance thresholds $\varepsilon_1 \geq \dots \geq \varepsilon_T$, and density kernel $K(\cdot)$.

1. At iteration $t = 1$,
 - a. For $i = 1, \dots, N$, repeat:
 - i. Sample sample $\theta_i^{(1)} \sim \pi(\theta)$, and simulate dataset $x \sim p(x|\theta_i^{(1)})$.
 - ii. Accept $\theta_i^{(1)}$ if $\rho(s(x), s_{obs}) \leq \varepsilon_1$, otherwise reject this sample.
 - iii. Set weight $\omega_i^{(1)} = 1/N$.
 - b. Take τ_2^2 as twice the empirical variance of the $\theta_i^{(1)}$ s
2. At iteration $2 \leq t \leq T$
 - a. For $i = 1, \dots, N$, repeat:
 - i. Pick θ_i^* from $\theta_j^{(t-1)}$ s with probabilities $\omega_j^{(t-1)}$
 - ii. Generate sample $\theta_i^{(t)} \sim K(\theta|\theta_i^*, \tau_t^2)$, and simulate a dataset $x \sim p(x|\theta_i^{(t)})$.
 - iii. Accept $\theta_i^{(t)}$ if $\rho(s(x), s_{obs}) \leq \varepsilon_t$, otherwise reject this sample.
 - iv. Set the weight of this accepted particle as

$$\omega_i^{(t)} \propto \frac{\pi(\theta_i^{(t)})}{\sum_{j=1}^N \omega_j^{(t-1)} K(\theta_i^{(t)}|\theta_j^{(t-1)}, \tau_t^2)}.$$

- b. Take τ_{t+1}^2 as twice the weighted empirical variance of the $\theta_i^{(t)}$ s.
-

At the first iteration, this algorithm draws samples from the prior distribution $\pi(\theta)$, simulates the model using the sample, calculate summary statistics, and select N samples that satisfy the error criterion. This step actually is the rejection-ABC algorithm. However, at the subsequent iterations, samples are drawn from a density kernel $K(\theta)$ based on the previous particle population. A Gaussian kernel is used in Beaumont et al. [4], given by

$$K(\theta_i^{(t)}|\theta_j^{(t-1)}, \tau_t^2) = \varphi\{\tau_t^{-1}(\theta_i^{(t)} - \theta_j^{(t-1)})\},$$

where $\varphi(\cdot)$ is the density of a normal distribution. This algorithm effectively performs the repeated importance sampling technique, which is also known as population Monte Carlo [12]. Similar algorithms have been proposed by using different formulas to calculate the weights and different kernel functions [4, 46, 50].

SMC-ABC has addressed a potential drawback of the rejection and regression approaches. If the data are informative, the posterior distribution may be very narrow

compared with the prior, then the rejection and regression algorithms may become inefficient. Thus repeatedly sampling from a gradually improving approximation of the posterior will make the distribution of summary statistics become closer to the posterior distribution, and increase the density of samples whose summary statistics is located in the vicinity of the target [5].

7 Choice of Summary Statistics

As discussed in previous sections, the posterior distribution of dataset $p(\theta|y_{obs})$ is approximated by

$$p(\theta|s_{obs}) \propto p(s_{obs}|\theta)\pi(\theta),$$

where s_{obs} is the summary statistics which usually has lower dimension than that of the data y_{obs} . If s_{obs} is sufficient,

$$p(\theta|s_{obs}) = p(\theta|y_{obs}).$$

When s_{obs} is highly informative, $p(\theta|s_{obs}) \approx p(\theta|y_{obs})$ is a good approximation. However, for many practical problems, it is hard to derive sufficient statistics or even a highly informative statistics. An appropriate choice of summary statistics is required to balance the informativeness and low-dimensionality. In some application fields, there has been a history of the development of summary statistics within a model-based framework in recent years. However, it is also possible that empirical summaries can be used without any strong theory to support them. Thus the selection of informative summary statistics is one of the important steps in the application of ABC. In recent years a number of methods have been proposed regarding the selection of summary statistics [19, 38].

Joyce and Marjoram [24] first proposed the ε -sufficiency concept and score of statistics for selecting an additional summary statistic s_k from the candidate set, when the model already has summary statistics s_1, \dots, s_{k-1} . Later three methods regarding the choice of summary statistics have been used in application [31], namely

1. selection of a subset of the summary statistics that maximizes prespecified criteria such as the Akaike Information Criterion [11] or the entropy of a distribution [35];
2. partial least square regression to get linear combinations of the original summary statistics that are maximally decorrelated and highly correlated with the parameters [53]; and
3. summary statistics are chosen by minimizing a loss function under the assumption of a statistical model between parameters and transformed statistics of simulated data [1, 19].

Blum et al. [11] provided a comprehensive review of the principal methods. However, this topic still remains as a challenging problem in Bayesian inference.

8 Early Rejection ABC

To reduce the simulation time, a number of inference methods have been proposed based on the idea of early rejection. For example, the delayed ABC divides a method into two stages [13]. In the first stage, a sample of parameters may be rejected or accepted by using an approximated posterior distribution. If it is accepted, a standard ABC method will be applied in the second stage to evaluate the discrepancy between the observation data and simulation. This idea has been used in the MCMC-ABC for inferring stochastic differential equation models, in which the prior distribution and proposal distribution are used in the first stage for early rejection [37]. Based on the MCMC-ABC [37], a sample is rejected if the following ratio is less than a sample $\omega \sim U(0, 1)$ by using the same notations in Eq. (3)

$$\omega > \frac{\pi(\theta^*)\pi(\theta_i|\theta^*)}{\pi(\theta_r)\pi(\theta^*|\theta_i)}. \tag{5}$$

In this approach, the kernel density function $p(y|\theta)$ is removed from the ratio above. Thus the performance of this early-rejection technique is fully dependent on the choice of the proposal density function $\pi(\theta^*|\theta_i)$.

A recently published approach is the Lazy ABC, which proposes a random stopping rule to abandon simulations with unsatisfactory accuracy [39]. This method makes ABC more scalable to applications where simulation is expensive. The detailed algorithm is given below

Algorithm 8 Lazy ABC

Input: prior density $\pi(\theta)$ and importance density $g(\theta)$, observation data y_{obs} , summary statistics $s(\cdot)$, tolerance level ϵ , distance function $\rho(\cdot, \cdot)$, proposal distribution $q(\cdot)$, and a continuous probability function $\alpha(\theta, x)$.

At iteration $i = 1 : N$

1. Generate a sample from importance sampling $\theta^* \sim g(\theta)$.
2. Simulate the model to get a dataset $x^* \sim p(x|\theta^*)$ and let $\alpha^* = \alpha(\theta^*, x^*)$.
3. With probability α^* continue to step 4. Otherwise perform early rejection: namely let $I^* = 0$ and go to step 6.
4. Simulate the model to get dataset $Y^* \sim p(Y|\theta^*, x^*)$.
5. Set $I_{ABC}^* = \mathbb{1}[d(s(y^*), s(y_{obs})) < \epsilon]$ and $I^* = I_{ABC}^*/\alpha^*$.
6. Set $w^* = I^*\pi(\theta^*)/g(\theta^*)$.
7. Repeat steps 1~ 6 until the required number of posterior samples is obtained.

Output: A set of N pairs of (θ^*, w^*) values.

The detailed information of Lazy importance sampling and multiple stopping decision can be found in [39].

9 ABC Software Packages

A number of computer software packages have been designed in recent years to implement ABC in different platforms using various computer languages. A software package, BioBayes, provides a framework for Bayesian parameter estimation and evidential model ranking over models of biochemical systems using ordinary differential equations. This package is extensible allowing additional modules to be included [52]. A Python package, ABC-SysBio, implements parameter inference and model selection for dynamical systems in the ABC framework [29]. This package combines three algorithms: ABC rejection sampler, SMC ABC for parameter inference, and SMC ABC for model selection. It is designed to work with models written in Systems Biology Markup Language (SBML). Deterministic and stochastic models can be analyzed in ABC-SysBio. In addition, a computational tool SYSBIONS has been designed for model selection and parameter inference using nested sampling [23]. Using a data-based likelihood function, this package calculates the evidence of a model and the corresponding posterior parameter distribution. This is a C-based, GPU-accelerated implementation of nested sampling that is designed for biological applications.

Also in the R platform, a number of software packages have been designed. Among them, package `abc` implements Rejection ABC with many methods of regression post-processing; while `EasyABC` implements a wide suite of ABC algorithms but not post-processing [36]. Package `abctools` has been designed to complement the existing software provision of ABC algorithms by focusing on tools for tuning them. It implements many previous unavailable methods from literature and makes them easy available to the research community [36]. In addition, there are also two ABC packages implemented as MATLAB toolbox. `EP-ABC` has been designed for state space models and related models, and `ABC-SDE` for inferring parameters in stochastic differential equations [37]. There are still some other software packages that have been reviewed in [36], including `ABCreg`, `ABCtoolbox`, `Bayes SSC`, `DIY-ABC`, and `PopABC`.

10 Conclusion

In this chapter, we have reviewed a number of algorithms of ABC, together with the relevant improvements, from choice of summary statistics to early rejection, aiming at increasing the statistical accuracy and computational efficiency. In addition, we give a few of the widely used software packages for the practical use of ABC algorithms. In recently years, the ABC methods have been applied to a wide range of

inference problems in biology, economics, engineering and physical sciences. These applications have also raised more challenging questions for parameter inference, such as high-dimensional data [28, 34] and stochastic modeling [55], which provides interesting topics for future research.

References

1. Aeschbacher, S., Beaumont, M. A., Futschik, A.: A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics* **192**, 1027–1047 (2012)
2. Barthelmé, S., Chopin, N.: Expectation propagation for likelihood-free inference. *J. Am. Stat. Assoc.* **109**, 315–333 (2014)
3. Bazin, E., Dawson, K.J., Beaumont, M.A.: Likelihoodfree inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics* **185**, 587–602 (2010)
4. Beaumont, M.A.: Adaptive approximate Bayesian computation. *Biometrika* **96**, 983–990 (2009)
5. Beaumont, M.A.: Approximate Bayesian computation in evolution and ecology. *Annu. Rev. Ecol. Evol. Syst.* **41**, 379–406 (2010)
6. Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002)
7. Biau, G., Cérou, F., Guyader, A.: New insights into approximate Bayesian computation. *Ann. I. H. Poincaré B* **51**, 376–403 (2015)
8. Blum, M.G.B.: Approximate Bayesian computation: a nonparametric perspective. *J. Am. Stat. Assoc.* **105**, 1178–1187 (2010)
9. Blum, M.G.B.: Regression approaches for approximate Bayesian computation (2017). arXiv:1707.01254v1
10. Blum, M.G.B., François, O.: Non-linear regression models for approximate Bayesian computation. *Stat. Comput.* **20**, 63–73 (2010)
11. Blum, M.G.B., Nunes, M.A., Prangle, D., Sisson, S.A.: A comparative review of dimension reduction methods in approximate Bayesian computation. *Stat. Sci.* **28**(2), 189–208 (2013)
12. Cappé, O., Guillin, A., Marin, J.-M., Robert, C.P.: Population Monte Carlo. *J. Comput. Graph. Stat.* **13**(4), 907–929 (2004)
13. Christen, J.A., Fox, C.: Markov chain Monte Carlo using an approximation. *J. Comput. Graph. Stat.* **14**, 795–810 (2005)
14. Csilléry, K., Blum, M.G.B., Gaggiotti, O., François, O.: Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* **25**(7), 410–418 (2010)
15. Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. *J. R. Stat. Soc. Ser. B* **68**, 411–436 (2006)
16. Del Moral, P., Doucet, A., Jasra, A.: An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Stat. Comput.* **22**(5), 1009–1020 (2012)
17. Deng, Z., Tian, T.: A continuous optimization approach for inferring parameters in mathematical models of regulatory networks. *BMC Bioinform.* **15**, 256 (2014)
18. Drovandi, C.C., Pettitt, A.N.: Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics* **67**(1), 225–233 (2011)
19. Fearnhead, P., Prangle, D.: Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B* **74**, 419–474 (2012)
20. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.: *Bayesian Data Analysis*. Chapman and Hall/CRC Press, London (2003)
21. Goel, G., Chou, I.C., Voit, E.O.: System estimation from metabolic time-series data. *Bioinformatics* **24**(21), 2505–2511 (2008)

22. Green, P.J., Łatuszyński, K., Pereyra, M., Robert, C.P.: Bayesian computation: a summary of the current state, and samples backwards and forwards. *Stat. Comput.* **25**, 835–862 (2015)
23. Johnson, R., Kirk, P., Stumpf, M.P.H.: SYSBIONS: nested sampling for systems biology. *Bioinformatics* **31**(4), 604–605 (2015)
24. Joyce, P., Marjoram, P.: Approximately sufficient statistics and Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* **7**(1), Article 26 (2008)
25. Kousathanas, A., Leuenberger, C., Helfer, J., Quinodoz, M., Foll, M., Wegmann, D.: Likelihood-free inference in high-dimensional models. *Genetics* **203**, 893–904 (2016)
26. Kypraios, T., Neal, P., Prangle, D.: A tutorial introduction to Bayesian inference for stochastic epidemic models using approximate Bayesian computation. *Math. Biosci.* **287**, 42–53 (2016)
27. Lenormand, M., Jabot, F., Deffuant, G.: Adaptive approximate Bayesian computation for complex models. *Comput. Stat.* **28**(6), 2777–2796 (2013)
28. Li, J., Nott, D.J., Fan, Y., Sisson, S.A.: Extending approximate Bayesian computation methods to high dimensions via a Gaussian copula model. *Comput. Stat. Data Anal.* **106**, 77–89 (2017)
29. Liepe, J., et al.: A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat. Protoc.* **9**(2), 439–456 (2014)
30. Lillacci, G., Khammash, M.: Parameter estimation and model selection in computational biology. *PLoS Comput. Biol.* **6**(3), e1000696 (2010)
31. Lintusaari, J., Gutmann, M., Dutta, R., Kaski, S., Corander, J.: Fundamentals and recent developments in approximate Bayesian computation. *Syst. Biol.* **66**(1), e66–e82 (2017)
32. Marin, J.-M., Pudlo, P., Robert, C.P., Ryder, R.J.: Approximate Bayesian computational methods. *Stat. Comput.* **22**, 1167–1180 (2012)
33. Marjoram, P., Molitor, J., Plagnol, V., Tavaré, S.: Markov chain Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. U. S. A.* **100**(26), 15324–15328 (2003)
34. Nott, D.J., Fan, Y., Marshall, L., Sisson, S.A.: Approximate Bayesian computation and Bayes's linear analysis: toward high-dimensional ABC. *J. Comput. Graph. Stat.* **23**(1), 65–86 (2014)
35. Nunes, M.A., Balding, D.J.: On optimal selection of summary statistics for approximate Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* **9**, Article 34 (2010)
36. Nunes, M.A., Prangle, D.: abctools: an R package for tuning approximate Bayesian computation analyses. *R J.* **7**(2), 189–205 (2015)
37. Picchini, U.: Inference for SDE models via approximate Bayesian computation. *J. Comput. Graph. Stat.* **23**(4), 1080–1100 (2014)
38. Prangle, D.: Summary statistics in approximate Bayesian computation (2015). arXiv:1512.05633
39. Prangle, D.: Lazy ABC. *Stat. Comput.* **26**, 171–185 (2016)
40. Pritchard, J.K., Seielstad, M.T., Perez-Lezaun, A., Feldman, M.W.: Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**(12), 1791–1798 (1999)
41. Robert, C.P.: Approximate Bayesian Computation: A Survey on Recent Results. *Monte Carlo and Quasi-Monte Carlo Methods*, pp. 185–205. Springer, Cham (2016)
42. Robert, C.P., Casella, G.: *Monte Carlo Statistical Methods*. Springer, New York (2004)
43. Rubin, D.B.: Bayesianly justifiable and relevant frequency calculations for the applied statistics. *Ann. Stat.* **12**(4), 1151–1172 (1984)
44. Sisson, S.A., Fan, Y.: Likelihood-Free MCMC. *Handbook of Markov Chain Monte Carlo*, pp. 313–335. CRC Press, Boca Raton (2011)
45. Sisson, S.A., Fan, Y., Tanaka, M.M.: Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. U. S. A.* **104**(6), 1760–1765 (2007)
46. Sisson, S.A., Fan, Y., Tanaka, M.M.: Sequential Monte Carlo without likelihoods. *Proc. Natl. Acad. Sci. U. S. A.* **106**(39), 16889 (2009)
47. Sunnaker, M., et al.: Approximate Bayesian computation. *PLoS Comput. Biol.* **9**(1), e1002803 (2013)
48. Tavaré, S., Balding, D., Griffith, R., Donnelly, P.: Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997)

49. Tian, T., Smith-Miles, K.: Mathematical modeling of GATA-switching for regulating the differentiation of hematopoietic stem cell. *BMC Syst. Biol.* **8**(Suppl 1), S8 (2014)
50. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.H.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**, 187–202 (2009)
51. Turner, B.M., Van Zandt, T.: A tutorial on approximate Bayesian computation. *J. Math. Psychol.* **56**(2), 69–85 (2012)
52. Vyshemirsky, V., Girolami, M.: BioBayes: a software package for Bayesian inference in systems biology. *Bioinformatics* **24**(17), 1933–1934 (2008)
53. Wegmann, D., Leuenberger, C., Excoffier, L.: Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* **182**, 129–141 (2009)
54. Wilkinson, D.J.: Bayesian methods in bioinformatics and computational systems biology. *Brief Bioinform.* **8**(2), 109–116 (2007)
55. Wu, Q., Smith-Miles, K., Tian, T.: Approximate Bayesian computation schemes for parameter inference of discrete stochastic models using simulated likelihood density. *BMC Bioinform.* **15**, S3 (2014)

The Loop-Weight Changing Operator in the Completely Packed Loop Model

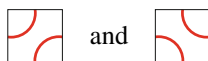


Bernard Nienhuis and Kayed Al Qasimi

Abstract Loop models are statistical ensembles of closed paths on a lattice. The most well-known among them has a variety of names such as the dense $O(n)$ loop model, the Temperley-Lieb (TL) model. This note concerns the model in which the weight of the loop $n = 1$, and a local operator which changes the weight of all the loops that surround the position of the operator to some other value. A conjecture of the expectation value of the one-point function of this operator was formulated 15 years ago. In this note we sketch the proof.

1 Introduction

It has long been recognized that loop models can represent many different local spin models in statistical mechanics. The model we deal with in this note, was introduced [1] as a representation of the Potts model. It has a free parameter, the weight of a loop, which is the square root of the number of states of the Potts model. The case that this weight is unity corresponds to the bond percolation model, and is the case we deal with here. The configurations of the model are a tiling of the square lattice, in which each face of the lattice is covered with one of two tiles



Thus in every configuration the red arcs in the tiles form paths on the lattice, which are either closed (hence loops), or terminating at the boundary, if there is one. The partition sum of the model is trivial, it is the product over all faces of

B. Nienhuis (✉)
Institute of Physics, Amsterdam, The Netherlands
e-mail: b.nienhuis@uva.nl

K. Al Qasimi
Korteweg de Vries Institute, Amsterdam, The Netherlands
e-mail: s.k.s.k.s.alqasemi@uva.nl

the sum of the two weights of the faces. Non-trivial are observables, which give weights to the configurations according to some specific properties of the paths. The loop-weight changing operator (LWCO) inserted at a given vertex, gives the loops surrounding that vertex a new weight w possibly different from that of the other loops, i.e. from one. The expectation value of the one-point function of this operator is the generating function of the probabilities of having a specific number of loops surrounding the point of insertion. For brevity we will use the short-hand LWCO also for one-point function of this operator, trusting the context will make it unambiguous.

A new approach to the study of this model originates in the work of Razumov and Stroganov [2] who found a connection between the XXZ model and combinatoric problems as Alternating Sign Matrices (ASM) and Plane Partitions (PP) [3]. A connections with loop models followed quickly [4] and led to the famous Razumov-Stroganov conjecture featuring a connection between two types of loop models on different geometries [5]. It was proven by Cantini and Sportiello [6]. These connections led to a wealth of explicit formulae for expectation values of observables and of indicator functions, some proven others conjectured. The value of these is that the formulae are (supposedly) exact with finite distances and geometries, rather than only in the scaling limit. One of these observables is the LWCO that turns the loop weight of surrounding loops into w , on a cylinder of infinite length and circumference L . Mitra and Nienhuis [7] conjectured the value of its one-point function $P(L, w) = F(L, w)/F(L, 1)$, with

$$F(L, a^2 + a^{-2}) = (a + a^{-1})^{-(L \bmod 2)} \det_{r,s=0}^{L-1} a^{-1} \binom{r+s}{s} + a \delta_{r,s} \quad (1)$$

This expression was not based on any theoretical understanding, let alone a derivation or proof. It was completely guessed from the recognition of the coefficients in the polynomial, and subsequently verified for large but finite L . We remark that the symmetry for $a \rightarrow 1/a$ is manifest in the LHS, but not in the RHS of Eq. (1).

A discussion with Christian Hagendorf at the Matrix workshop Statistical Mechanics, Combinatorics and Conformal Field Theory in 2017, eventually led to a proof, which we will sketch in this note. The line of argument is as follows. We will first generalize the model to be inhomogeneous, thus introducing a number of variables on which the LWCO depends. We will show that the LWCO is a rational function of these variables. A family of recursion relations in the size of the system can be used to fix the value of the numerator and the denominator, for a number of values of one of the variables. The number of values suffices to completely determine these functions, by the polynomial interpolation formula. A publication of Hagendorf and Morin-Duchesne [8] suggested the inhomogeneous generalization of Eq. (1). It then remained to show that this expression satisfies the same recursion relation as the LWCO.

2 Inhomogeneous TL Model

An important step towards a partial proof of the RS conjecture by Di Francesco and Zinn-Justin [9], was making the TL model inhomogeneous, by associating a variable to each column and each row of faces in the lattice. We imagine the axis of the cylinder to be vertical, so the columns run along the length of the cylinder, and the rows form rings around the cylinder. These variables are often called rapidities due to their role in the relativistic field theory which is the scaling limit of the model. The Boltzmann weights at a particular face then depend on the two rapidities associated with the column and row the face is in. Specifically the Boltzmann weight of a face can be written as

$$R(w, z) = \frac{qz - q^{-1}w}{qw - q^{-1}z} \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} + \frac{z - w}{qw - q^{-1}z} \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} \tag{2}$$

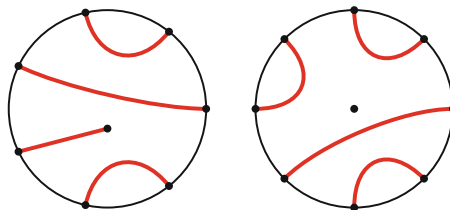
where z and w are the variables associated with the column and row, respectively, that the face belongs to, and $q = e^{2\pi i/3}$. The two coefficients add up to one, and are real positive when $z/w = e^{i\phi}$ with $\phi \in (0, 2\pi/3)$. The transfer matrix can be written as

$$T(w, \mathbf{z}) \equiv T(w; z_1, z_2, \dots, z_L) = \prod_{i=1}^L R(z_i, w), \tag{3}$$

where we use \mathbf{z} as a shorthand for $\{z_1, z_2, \dots, z_L\}$. A term in the expansion of this product corresponds to the graph



This transfer matrix acts as a stochastic matrix in the space of so-called link patterns. In each link pattern the edges cut by the rim of the cylinder are connected pairwise, by paths that do not intersect, as in the following example for $L = 7$ and $L = 8$



For odd L the link pattern includes an unpaired edge, which is the connected to a path all along the half-infinite cylinder. For even L we mark the disk with a puncture, to distinguish if a path between two edges at the rim is along one side of the cylinder or the other. When two edges that are connected in a link pattern are reconnected by the transfer matrix or another operator, the resulting loop is removed.

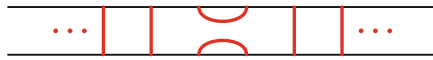
Due to the Yang-Baxter equation $[T(w, \mathbf{z}), T(v, \mathbf{z})] = 0$, and consequently the eigenvectors of $T(w, \mathbf{z})$ do not depend on w . With $\Psi(\mathbf{z})$ we denote the ground state, i.e. the eigenvector with eigenvalue 1. In the regime where all transfer matrix elements are non-negative this is the largest (Perron-Frobenius) eigenvalue. Because the transfer matrix is a rational function of the variables z_j , also $\Psi(\mathbf{z})$ is rational, and with suitable normalization polynomial. As shown in [9] $\Psi(\mathbf{z})$ satisfies

$$\check{R}_i(z_i, z_{i+1}) \Psi(z_1, \dots, z_i, z_{i+1}, \dots, z_L) = \Psi(z_1, \dots, z_{i+1}, z_i, \dots, z_L), \tag{4}$$

where the operator

$$\check{R}(w, z) = \frac{qz - q^{-1}w}{qw - q^{-1}z} \mathbb{1} + \frac{z - w}{qw - q^{-1}z} e_i \tag{5}$$

and the operator e_i acts on position i and $i + 1$ of a link pattern as



i.e. connecting the partners of position i and $i + 1$, and creating an arc connecting i and $i + 1$ themselves. These equations (4) are called quantum Knizhnik-Zamolodchikov (qKZ) equations as they are analogous to q -deformed versions of the Knizhnik-Zamolodchikov equations [10] on correlation functions in conformal field theory.

We write the ground state vector

$$\Psi(\mathbf{z}) = \sum_{\alpha} \psi_{\alpha}(\mathbf{z}) \alpha \tag{6}$$

where α is a link pattern, and the sum is over all link patterns of a given size (i.e. circumference of the cylinder). For the weights ψ_{α} of link patterns α in which the positions i and $i + 1$ are *not connected* by a small (minimal) arc, Eq. (4) leads to

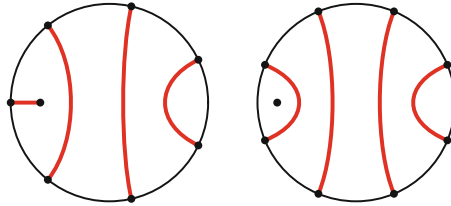
$$(qz_{i+1} - q^{-1}z_i) \psi_{\alpha}(\dots, z_i, z_{i+1}, \dots) = (qz_i - q^{-1}z_{i+1}) \psi_{\alpha}(\dots, z_{i+1}, z_i, \dots) \tag{7}$$

For the polynomial solution this implies that $\psi_{\alpha}(\dots, z_i, z_{i+1}, \dots)$ must contain the factor $(qz_i - q^{-1}z_{i+1})$, and is otherwise symmetric for interchange of z_i and z_{i+1} .

If this argument is used recursively on a link pattern not containing a minimal arc in a sequence of edges $\{k, \dots, n\}$, it must contain the factor

$$\prod_{i=k}^{n-1} \prod_{j=i+1}^n (qz_j - q^{-1}z_i)$$

and be otherwise symmetric in the variables $\{z_k, \dots, z_n\}$. For the most nested link pattern, μ , (again for $L = 7$ and $L = 8$)



in which the small arc not containing the puncture connects the position 1 and L , contains the factor $\prod_{i=1}^{L-1} \prod_{j=i+1}^L (qz_j - q^{-1}z_i)$. Because all other weights can be derived from this one with the qKZ equations (4), from which functions symmetric in z_i and z_{i+1} can be factored out, the weight of the most nested link pattern μ is in fact given by

$$\psi_\mu(\mathbf{z}) = \prod_{i=1}^{L-1} \prod_{j=i+1}^L (qz_j - q^{-1}z_i) \tag{8}$$

with no further factors symmetric in \mathbf{z} . Clearly $\psi_\mu(\mathbf{z})$ is homogeneous and of joint degree $L(L - 1)/2$. As polynomial of a single z_i it is of degree $L - 1$. These properties transcend to all ψ_α , as they are conserved by Eq. (4).

3 Recursions in System Size

Reference [9] also shows that (4) implies a recursion relation between the ground states of systems different in size by 2. For this it is useful to introduce operators that mediate between link patterns of different sizes. The operator σ_i introduces two additional positions between positions $i - 1$ and i , and a small arc that connects them. Conversely, the operator τ_i connects (the partners of) $i - 1$ and i , and then removes the positions themselves. Thus, the operators σ_i acting on a link patterns of size L results in a link pattern of size $L + 2$, and τ_i results in link pattern of size $L - 2$. These operators satisfy

$$\sigma_i \tau_i = e_i \quad \text{and} \quad \tau_i \sigma_i = \mathbb{1} \tag{9}$$

and acting on the ground state vectors gives

$$\psi_{\sigma_i \alpha}(\dots, z_{i-1}, zq, zq^2, z_i, \dots) = \psi_{\alpha}(\dots, z_{i-1}, z_i, \dots) (q^{-1} - q) z \prod_{i=1}^L -(z - z_i)^2 \quad (10)$$

and

$$\tau_i \Psi(\dots, z_{i-1}, zq^2, zq, z_i, \dots) = \Psi(\dots, z_{i-1}, z_i, \dots) (q^{-1} - q) z \prod_{i=1}^L -(z - z_i)^2 \quad (11)$$

These relations for ground state elements can be read as recursion relations in the system size, as the LHS refers to a system of size $L + 2$, and the RHS has size L . We call these relation the *fusion recursion relations*.

Later Di Francesco et al. [11] introduced another recursion in the system size, not by fixing the ratio between two variables, but by sending one to zero. Since we extend their results we treat this in some more detail. When one of the variables z_i is zero, the Boltzmann weights in the corresponding column are from Eq. (2)

$$R(w, 0) = -q \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} - q^{-1} \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} \quad (12)$$

We wish to relate a system with size L to a system with size $L + 1$, with the same variables, and one additional variable equal to zero. We will refer to this recursion relation as the braid recursion relation, as the \check{R} -operator reduces to a so-called braid operator that satisfies the Reidemeister moves of the braid group.

Consider a path in the size- L system, that crosses the location of the rapidity-zero column, wanders around and crosses back on a face adjacent to the other crossing. In the size- $(L + 1)$ system, we consider the same configuration, combined with all possible configurations in the rapidity-zero column. For the weight of these two cases we get the following equation:

$$\begin{array}{l} \text{---} \\ \text{---} \end{array} = q^2 \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} + 1 \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} + 1 \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} + q^2 \begin{array}{|c|} \hline \text{---} \\ \hline \end{array} \quad (13)$$

The green line is the line with variable 0, and the red curves are the paths, while the two faces where the crossing occurs are indicated with a black box. We see that in the system of size $L + 1$, there are only two possible connectivities in four configurations. In one case, with weight 1, the path continues as in the size- L system, but avoiding the D-tour, while the paths entering the two boxes vertically are connected along the D-tour that is avoided by the original path. In the other cases, the branches of the path connect up or down to the vertical. The total weight of the last connectivity is $1 + q^2 + q^{-2} = 0$, so it can be disregarded.

If this applies to paths which cross in two adjacent faces, the same must apply to paths which cross the zero rapidity line at more distant faces, as all the crossings in between these two faces, form a nested set of double crossings.

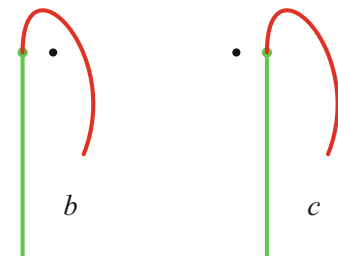
In conclusion one can say that any path that crosses the rapidity-zero column once and back, has the same connectivity between the systems of size L and that of $L + 1$.

When L is odd, the system has one path (the defect path) that is not closed, but runs along the infinite cylinder from one end to the other. The system with size $L + 1$ then does not contain such path, so when the system is extended with a rapidity zero, the defect path must join up with the additional rapidity zero column. However, this can be done in two ways, as the puncture can be placed on either side of the path. Figure 1 shows the two possibilities. Since there is no way to exclude one, we accept both, with weights b if the puncture is to the right of the juncture, and c if it is to the left as indicated in the figure. The result is that the puncture is “inside” the path with weight b and “outside” with weight c .

The defect line can intersect the zero rapidity column any number of times, even or odd. Therefore an additional crossing should not make any difference in the weights. Figure 2 shows the configurations if the defect path crosses one more time before it joins up with the zero-rapidity line. Now we see that the puncture is inside with weight $-cq^{-1}$ and outside with weight $-bq - cq - bq^{-1}$. This is consistent only when both

$$c q^{-1} = -b \quad \text{and} \quad b q + c q + b q^{-1} = -c \tag{14}$$

Fig. 1 The puncture relative to the path formed by the unmatched path and the zero rapidity line. The defect path is shown as red, the zero rapidity line is indicated by green, and the puncture is black



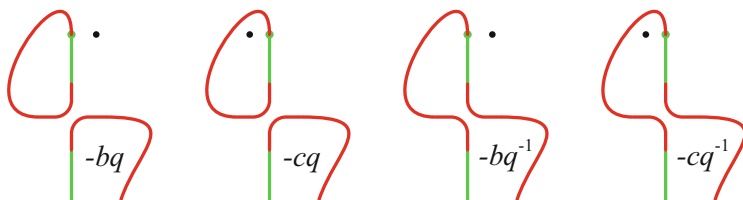


Fig. 2 The configurations formed by placing the puncture on either side of the path and uncrossing the intersection between the path and the rapidity-zero line in the two possible ways. As in Fig. 1 the green line represents the (unresolved) zero-rapidity line, and the red curves represent paths

These (overdetermined) equations are solved by

$$b = (-q)^{-1/2} \quad \text{and} \quad c = (-q)^{1/2} \tag{15}$$

where we chose $c = b^*$. If the defect path is connected to the zero-rapidity line in the opposite direction, the same solution is found. In conclusion, since an intersection of the defect line with the zero-rapidity line does not affect the weight of a configuration, we may argue without loss of generality as if the defect line never crosses the zero-rapidity line.

4 Recursion Relations for the LWCO

In the previous section we showed that the qKZ equations induce recursion relations in the system size for the elements of the ground state vector, and for the probability of certain events. In this section we will present how these recursions lead to recursion relations for the LWCO in the inhomogeneous TL model on an infinite cylinder. First we remind the reader that the elements of the ground state vector, that is the relative configurational weight of a half-infinite cylinder, are polynomials of degree $L - 1$ in each of the rapidities. Thus the configurational weights of the infinite cylinder, made up of two half-infinite cylinder has degree $2(L - 1)$. Therefore the LWCO is a rational function of degree $2(L - 1)$ for both the numerator and denominator. We can determine these polynomials completely, from knowing their value for $2L - 1$ values of one of the variables. Let z_i to be the variable of choice. We can apply the fusion recursion relation when $z_i = q^{-1} z_{i-1}$ or $z_i = q z_{i+1}$, and the braid recursion relation when $z_i = 0$. However, the qKZ equations (4) ensure that both the numerator and denominator are symmetric functions of the variables. Therefore we can use the fusion recursion relation for $z_i = q^{-1} z_j$ or $z_i = q z_j$ for any $j \neq i$. Together with the braid recursion relation this gives precisely enough values.

To establish some notation, let us use $\Phi(w, \mathbf{z})$ for the (one-point function of the) LWCO, with altered loop weight w , and $\Phi_n(w, \mathbf{z})$ and $\Phi_d(\mathbf{z})$ for its polynomial numerator and denominator:

$$\Phi(w, \mathbf{z}) = \frac{\Phi_n(w, \mathbf{z})}{\Phi_d(\mathbf{z})} \tag{16}$$

Because $\Phi_d(\mathbf{z}) = \Phi_n(1, \mathbf{z})$ it suffices to study Φ_n . To denote the recursions, we will use \mathbf{z} for the list $\{z_1, z_2, \dots, z_L\}$ as before, $(\mathbf{z} \mid z_i \rightarrow v)$ to indicate that z_i takes a specific value v , and $(\mathbf{z} \setminus i)$ for the list \mathbf{z} from which z_i is omitted, and similarly $(\mathbf{z} \setminus i, j)$ from which both z_i and z_j are omitted. The system size is implicit as the length of the last argument. We choose \mathbf{z} to have length L always, so that e.g. $(\mathbf{z} \setminus i)$ has length $(L - 1)$.

The weight of a specific combination of link patterns in the two half-infinite cylinders, is the product of two ground state elements. But the dependence on the variables is different, as the order of one is reversed relative to the other. The list of variable in reversed order is denoted as $\rho\mathbf{z}$. For example the sum of weights of all link patterns for both halves of the cylinder, should be equal to $\Phi_d(\mathbf{z})$, so

$$\Phi_d(\mathbf{z}) = \sum_{\alpha, \beta} \psi_\alpha(\mathbf{z}) \psi_\beta(\rho\mathbf{z}) = \left(\sum_{\alpha} \psi_\alpha(\mathbf{z}) \right)^2 \tag{17}$$

The order of the variables is immaterial because the qKZ equations ensure that the sum of all elements of the ground state vector is a symmetric function.

4.1 The Fusion Recursion Relation

From Eqs. (10) and (11) it is clear that

$$\Phi_d(\mathbf{z} \mid z_i \rightarrow qz_j) = \Phi_d(\mathbf{z} \setminus i, j) (-3q) z_j^2 \prod_{k \neq i, j}^L -(q^{-1}z_j - z_k)^4, \tag{18}$$

and

$$\Phi_d(\mathbf{z} \mid z_i \rightarrow q^{-1}z_j) = \Phi_d(\mathbf{z} \setminus i, j) (-3q^{-1}) z_j^2 \prod_{k \neq i, j}^L -(qz_j - z_k)^4. \tag{19}$$

In order to see what the analogous recursion is for Φ_n , we first use the symmetry of Φ_n for permutation of the \mathbf{z} to place z_i and z_j adjacent. As shown in [9], the corresponding double column simply connects the paths on the left to those on the right in the same row. This implies that the topology of the paths in the system of size L and that of size $L - 2$ is the same. As long as the LWCO is not inserted between the two adjacent columns carrying the variables z_i and z_j , the number of loops surrounding the point of insertion is the same in the two models. This implies that Φ_n satisfies precisely the same fusion recursion relations (18) and (19) as Φ_d , irrespective of the value of the altered loop weight. Clearly, the difference between the two functions Φ_n and Φ_d must come from the braid recursion relation.

4.2 The Braid Recursion Relation

We consider an inhomogeneous TL model on a cylinder with perimeter $L - 1$ and another one on a cylinder with perimeter L , with the same variables supplemented with one variable equal to zero. In this we consider a contractible loop in the size- $(L - 1)$ system, that intersects the position of the zero rapidity column in two adjacent faces, and study how this configuration is resolved in the size- L system. This is shown in Fig. 3. We observe that if the LWCO is *not* inserted in the original loop, the weight in the size- L system is equal to that in the size- $(L - 1)$ system, multiplied with the sum of the four weights, that is $1 + 1 + q^2 + q^{-2} = 1$. If the LWCO is inserted in the original loop, to the left of the zero-rapidity line, then it sits in a loop in the upper-right figure, and not in the remaining configurations, whose weights add up as $1 + q^2 + q^{-2} = 0$. In the surviving (upper-right) configuration, the two ends of the zero rapidity line are connected by a path. Likewise if the operator is inserted inside the loop, to the right of the zero rapidity line, it sits in a loop in the lower-left figure, and again not in the remaining figures. As far as this type of configuration is concerned, the weight of the LWCO is the same between the system of size L and that of size $(L - 1)$.

When a contractible loop intersects the zero-rapidity line in two arbitrary faces, other loops inside it intersect the zero-rapidity line in a nested fashion, in which the loops deepest in the nesting cut the zero-rapidity line in adjacent faces. We can resolve this recursively, starting with the loops deepest in the nest, and then the loops enclosing them, and so on. We conclude that in a system of size L , in which one of

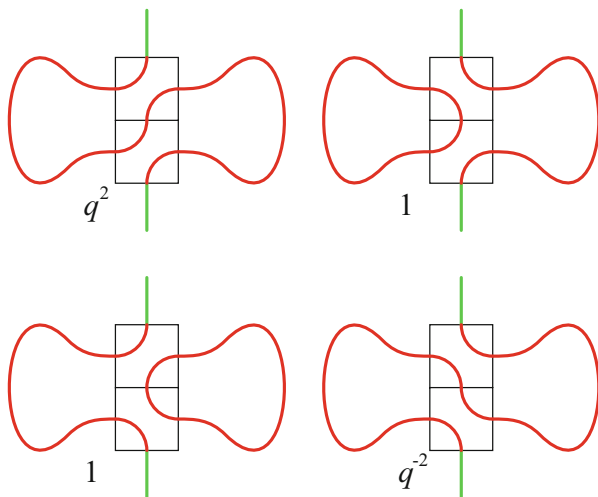


Fig. 3 A closed, contractible loop that cuts the rapidity-zero line in consecutive faces. The path as resolved in the size- L system is drawn red, and the continuation of the zero-rapidity line is shown in green. The total weight of the two faces is given for each configuration

the variables is zero, the relative weight of configurations with any given number of loops surrounding an operator insertion is completely the same as in the system of size $(L - 1)$ with only the $(L - 1)$ non-zero rapidities. This suggests that also the braid recursion relation is the same for Φ_n as for Φ_d . However, this is not the case. Since the braid recursion relation relates even and odd sized systems, a defect line may appear or disappear or turn into a loop.

In Eq. (1) the determinantal expression is divided by $(a + 1/a)$ for odd L . So in the determinant itself, while the loops that surround the operator insertion have weight $(a^2 + a^{-2})$, the defect line has weight $(a + 1/a)$. It is convenient to take this operator to replace the original LWCO: in the numerator the surrounding loops have weight $(a^2 + a^{-2})$, and the defect line has co-varying weight $(a + 1/a)$. In analogy, in the denominator all loops have weight 1, but the defect line has weight $\sqrt{3}$. So, from now on, we multiply the LWCO in odd- L systems by the simple factor $(a + 1/a)/\sqrt{3}$, and rename the resulting one-point function $\overline{\Phi}$ and similarly its numerator and denominator $\overline{\Phi}_n$ and $\overline{\Phi}_d$. The results for the fusion recursion relations are not altered, as the factors appear on the RHS and LHS of the recursion equally.

First we will consider a system of odd size L , and send one of its rapidities to zero. We have already seen that for a single insertion of the LWCO, all the loops in the corresponding system of size $(L - 1)$ that surround it, translate into a surrounding loop in the system of size L , with precisely the same weight. What changes is that in all configurations in the larger system a defect line appears, of which the weight is always $(a + 1/a)$. Thus we find for odd L

$$\overline{\Phi}_n(a^2 + a^{-2}, \mathbf{z} | z_i \rightarrow 0) = (a + a^{-1}) \overline{\Phi}_n(a^2 + a^{-2}, \mathbf{z} \setminus i) F_i(\mathbf{z}), \tag{20}$$

where the function $F_i(\mathbf{z})$ is a symmetric function of $(\mathbf{z} \setminus i)$, which we do not need, as it does not depend on a .

Second we consider a system of even size L , and send one of its rapidities to zero. Now the larger system has a defect line, and it has weight $(a + a^{-1})$ in all configurations. We have seen above that the zero-rapidity line effectively creates a new path. The defect path in the system of size $(L - 1)$ joins up with this new path to form a loop. While it does this the puncture for either end of the cylinder is placed to the right or left of the juncture with weight b or c , respectively, as illustrated in Fig. 4. The defect line and the zero-rapidity line form a loop. If this loop separates the two punctures, it must wind the cylinder. This happens with weight $c^2 + b^2 = -q - q^{-1} = 1$. The loop surrounds the operator insertion if it separates this from both punctures. Irrespective of where the insertion is, it is surrounded with weight 1 and not surrounded with weight 1. Thus, in total the insertion is surrounded with weight 1, and not surrounded with weight 2. Thus we see that the weight $(a + a^{-1})$ for the defect line in the size- $(L - 1)$ system is replaced by $(a^2 + a^{-2})$ with weight 1, and by 1 with weight 2. In total the weight $(a + a^{-1})$ in the small system is replaced by $(a^2 + a^{-2} + 2) = (a + a^{-1})^2$ in the big system. Thus also for even L we find

$$\overline{\Phi}_n(a^2 + a^{-2}, \mathbf{z} | z_i \rightarrow 0) = (a + a^{-1}) \overline{\Phi}_n(a^2 + a^{-2}, \mathbf{z} \setminus i) F_i(\mathbf{z}) \tag{21}$$

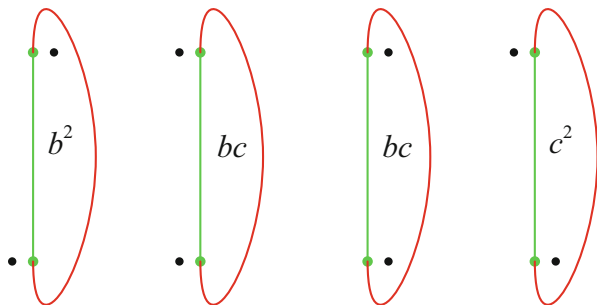


Fig. 4 The possible configurations of the puncture relative to the closed path in the width- $(L + 1)$ system made up of the unmatched path in the width- L system and the zero-rapidity line, with respective weights (top) and current factor (bottom)

just as for odd L . With $\overline{\Phi}_d(\mathbf{z}) = \overline{\Phi}_n(1, \mathbf{z})$ and $\overline{\Phi}(w, \mathbf{z}) = \overline{\Phi}_n(w, \mathbf{z})/\overline{\Phi}_d(\mathbf{z})$ this completes the braid recursion relation for the LWCO.

5 The Inhomogeneous Expression for LWCO

Now that we have found recursion relations that should be satisfied by the LWCO of the inhomogeneous TL model, we have the means to prove an expression if we have it. However, until now only the homogeneous limit was known. In the Matrix workshop on Statistical Mechanics, Combinatorics and Conformal Field Theory in 2017, Christian Hagendorf pointed us to his publication [8] in which an expression appears equivalent to (1). And it also gives an more general expression of which (1) is the homogeneous limit. Since we had calculated explicit expressions for the inhomogeneous LWCO for small systems ($L < 9$), it was not difficult to verify that indeed his expression is what we need. To make this explicit, we introduce the shorthand $[x] \equiv (x - x^{-1})$, and we propose that

$$\overline{\Phi}_n(a^2 + a^{-2}, \mathbf{z}) = \prod_{1 \leq i < j \leq L} \frac{(qz_i - q^{-1}z_j)(qz_j - q^{-1}z_i)}{\left[\begin{smallmatrix} 1/2 & -1/2 \\ z_i & z_j \end{smallmatrix} \right] \left[\begin{smallmatrix} 1/2 & -1/2 \\ z_j & z_i \end{smallmatrix} \right]} \times \prod_{i,j=1}^L \left[qz_i^{1/2} z_j^{-1/2} \right] \det_{i,j=1}^L \left(\frac{a^{-1}}{\left[\begin{smallmatrix} 1/2 & -1/2 \\ qz_i & z_j \end{smallmatrix} \right]} + \frac{a}{\left[\begin{smallmatrix} 1/2 & -1/2 \\ qz_j & z_i \end{smallmatrix} \right]} \right) \tag{22}$$

That this expression does indeed satisfy the recursion relations we have derived, can be shown by explicit row and column manipulations of the matrix after specification of the last variable $z_L \rightarrow 0$ or $z_L \rightarrow qz_j$ or $z_L \rightarrow q^{-1}z_j$ for any $j \neq L$ respectively. As mentioned before this gives $2L - 1$ values for a polynomial that has degree $2L - 2$

in z_L and thus proves the inhomogeneous expression for the LWCO. The fact that (1) is the homogeneous limit is shown in [8], which proves the expression conjectures in [7].

Acknowledgements We thank Christian Hagendorf for his important contribution to this result, and the MATRix organization for the excellent opportunity for scientific exchange.

References

1. Baxter, R.J., Kelland, S.B., Wu, F.Y.: Equivalence of Potts model or Whitney polynomial with an ice-type model. *J. Phys. A* **9** 397 (1976)
2. Razumov, A.V., Stroganov, Y.G.: Spin chains and combinatorics. *J. Phys. A* **34**, 3185 (2001). arXiv:cond-mat/0012141
3. Kuperberg, G.: Symmetry classes of alternating-sign matrices under one roof. *Ann. Math.* **156**, 835 (2002). arXiv:math/0008184
4. Batchelor, M.T., de Gier, J., Nienhuis, B.: The quantum symmetric XXZ chain at $\Delta = -1/2$, alternating sign matrices and plane partitions. *J. Phys. A* **34**, L265 (2001). arXiv:cond-mat/0101385
5. Razumov, A.V., Stroganov, Y.G.: Combinatorial nature of ground state vector of $O(1)$ loop model. *Theor. Math. Phys.* **138**, 333 (2004); *Teor.Mat.Fiz.* **138** (2004) 395; arXiv:math/0104216
6. Cantini, L., Sportiello, A.: Proof of the Razumov-Stroganov conjecture. *J. Combin. Theor. A* **118**, 1549 (2011). arXiv:1003.337
7. Mitra, S., Nienhuis, B.: Exact conjectured expressions for correlations in the dense $O(1)$ loop model on cylinders. *J. Stat. Mech.* P10006 (2004). arXiv:cond-mat/0407578
8. Hagendorf, C., Morin-Duchesne, A.: Symmetry classes of alternating sign matrices in the nineteen-vertex model. *J. Stat. Mech.* 053111 (2016). arXiv:1601.01859
9. Di Francesco, P., Zinn-Justin, P.: Around the Razumov-Stroganov conjecture: proof of a multi-parameter sum rule. *Electron. J. Combin.* **12**, R6 (2005). arXiv:math-ph/0410061
10. Knizhnik, V.G., Zamolodchikov, A.B.: Current-algebra and Wess-Zumino model in 2 dimensions. *Nucl. Phys. B* **247**, 83 (1984)
11. Di Francesco, P., Zinn-Justin, P., Zuber, J.-B.: Sum rules for the ground states of the $O(1)$ loop model on a cylinder and the XXZ spin chain. *J. Stat. Mech.* P08011 (2006). arXiv:math-ph/0603009

A Note on Optimal Double Spending Attacks



Juri Hinz and Peter Taylor

Abstract In the present note we address the important problem of stability of blockchain systems. The so-called “double-spending attacks” (attempts to spend digital funds more than once) have been analyzed by several authors. We re-state these questions under more realistic assumptions than previously discussed and show that they can be formulated as an optimal stopping problem.

1 Introduction

In recent years, novel concepts originating from the blockchain idea have gained popularity. Their rapidly emerging software realizations are based on a mixture of traditional techniques (peer-to-peer networking, data encryption) and more modern concepts (consensus protocols). Digital currencies represent assets of these systems, their transactions are written and kept in an electronic ledger as part of the operation of the blockchain system. The main difference from a traditional financial system is that crypto-currencies are not issued and supervised by a central authority, but are maintained by joint efforts of a network consisting of independent computers (all running the same/similar software). Such a network searches for consensus which yields a common version of the ledger shared by all participants. The consensus is reached by means of a process, which is called *mining* and is usually backed by economic incentives. Blockchain systems are believed to achieve the same level of certainty and security as those governed by a central authority but do that at significantly lower costs. Furthermore, because of the distributed, decentralized, and

J. Hinz (✉)

School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway, NSW, Australia

e-mail: Juri.Hinz@uts.edu.au

P. Taylor

School of Mathematics and Statistics, University of Melbourne, Melbourne, VIC, Australia

e-mail: taylorpg@unimelb.edu.au

homogeneous architecture of the network, a blockchain system can reach a very high level of stability due to data redundancy and hard/software replicability.

Following the mining process, all network participants append, validate, and mutually agree on a common version of the data history, which is usually referred to as the *blockchain ledger*. Although the invention of mining is considered to be a real break-through which solves the long-standing consensus problem in computer science, there is criticism of this approach. The problem is that to reach consensus, real physical resources/efforts must be spent or at least allocated. For instance, the traditional Bitcoin protocol requires participants to solve cryptographic puzzles with real consumption of computing power and energy. This process is referred to as the so-called *proof of work*. Other blockchain systems avoid resource consumption and require temporary allocation of diverse resources, for instance the ownership of the underlying digital assets (proof of stake) or their spending (proof of burn). Furthermore, commitment of storage capacity (proof of storage) or a diverse combination of resource allocation/consumption can also be used.

Let us briefly elaborate on the proof of work; more details can be found in the excellent book by “Mastering Bitcoin” [1] by Andreas Antonopoulos. We focus on the Bitcoin protocol which was been initiated by Nakamoto [4], with a refinement on the double-spending problem in [5] and later in [3] with further considerations addressing propagation delay in [2]. In this framework, the ledger consists of a chain of blocks and each block contains valid transactions. The nodes compete to add a new block to the chain, and while doing so, each node attempts to collect transactions and solve a cryptographic puzzle. Once this puzzle is solved, it is made public to other nodes. This protocol also prescribes that if a peer node reports a completed block, then it must be verified, and if this block is valid, it must be attached to the chain, all uncompleted blocks shall be abandoned and a new block continuing the chain must be started. However, even following these rules, the chain forks regularly, which results in different nodes working on different branches. In order to reach a consensus in such cases, the protocol prescribes that a branch with shorter length must be abandoned as soon as a longer branch becomes known.

2 The Double-Spending Problem

Now we return to the resilience of the protocol to attacks. Note that within a blockchain system, the nodes are running publicly available open-source software (for mining) which can easily be modified by any private user to control the computer nodes in order to undermine the system. In principle, there are many ways of doing this. One of the most obvious among malicious strategies would be an attempt to spend the electronic money more than once. The analysis of such a strategy is referred to as the double-spending problem.

In the classical [4, 5] formulation of this problem a merchant waits for $n \in \mathbb{N}_0$ confirming blocks after a payment by a buyer, before providing a product or service. While the network is mining these n blocks, the attacker tries to build his/her own

secret branch containing a version of the history in which this payment is not made. The idea is to not include the paying transaction in the private secret branch whose length will overtake the official branch to be then published. If this strategy succeeds, then the private secret branch becomes official and the payment disappears in the ledger after the product/service is taken by the attacker. Nakamoto [4] provides and Rosenfeld [5] refines an estimate of the attacker’s success probability depending on his/her computational power and the number n of confirming blocks.

Let us briefly discuss their result before we elaborate on further details. In the framework of double-spending problem, it is assumed that a continuous-time Markov chain taking values in \mathbb{Z} describes the difference in blocks between the official and secret branches. As in [5], we consider this process at time points at which a new block in one of the branches is completed, which yields a discrete-time Markov chain $(Z_t)_{t=0}^\infty$. Having started secret mining after the block including the attacker’s payment (at block time $t = 0, Z_0 = 0$) the attacker considers the following situation: At each time $t = 1, 2, 3 \dots$, a new block in one of the branches (official or secret) is found, the block difference changes by ± 1 with probabilities (see [2, 5])

$$\begin{aligned} \mathbb{P}(Z_t = z + 1 | Z_{t-1} = z) &= 1 - q \\ \mathbb{P}(Z_t = z - 1 | Z_{t-1} = z) &= q. \end{aligned}$$

where $q \in]0, 1[$ is the ratio of the computational power controlled by the attacker to the total mining capacity. Consider a realistic case where the attacker controls a smaller part of the mining power $0 < q < 1/2$ than that controlled by honest miners. In this case, if at any block time $t = 0, 1, 2, \dots$ the block difference is $z \in \mathbb{Z}$, then the probability $a_\infty(z, \infty)$ that the secret branch overtakes the official branch within unlimited time after t is given by

$$a_\infty(z, q) = \mathbb{P}(\min_{u=0}^\infty Z_u < 0 | Z_0 = z) = \begin{cases} 1 & \text{if } z < 0 \\ (\frac{q}{1-q})^{z+1} & \text{otherwise.} \end{cases} \tag{2.1}$$

Furthermore, at the time when the n -th block in the official branch is mined, the probability that the attacker has mined $m = 0, 1, 2, \dots$ blocks follows the *negative binomial distribution* whose distribution function is given by

$$F_{q,n}(k) = \sum_{m=0}^k \binom{n+m-1}{m} (1-q)^n q^m, \quad k = 0, 1, 2, \dots \tag{2.2}$$

Both results (2.1) and (2.2) are combined in [5] to obtain the success probability of the double-spending as follows: Consider the situation where at the time the n -th block in the official chain is completed, the attacker has mined $m > n$ blocks which can be published immediately. The probability of this event is given by

$$\sum_{m=n+1}^\infty \binom{n+m-1}{m} (1-q)^n q^m = 1 - \sum_{m=0}^n \binom{n+m-1}{m} (1-q)^n q^m = 1 - F_{q,n}(n).$$

Next, consider the opposite event, assuming that when the n -th official block is completed, the attacker has not overtaken the official chain in which case $m \leq n$. In this case, the probability of winning the race is given by

$$\begin{aligned} \sum_{m=0}^n \binom{n+m-1}{m} (1-q)^n q^m a_{\infty}(n-m, q) &= \frac{q}{1-q} \sum_{m=0}^n \binom{n+m-1}{m} (1-q)^m q^n \\ &= \frac{q}{1-q} F_{1-q, n}(n). \end{aligned}$$

Clearly, the success probability is given by

$$1 - F_{q, n}(n) + \frac{q}{1-q} F_{1-q, n}(n). \tag{2.3}$$

For instance, if the merchant waits for six confirming blocks the attack succeeds with probabilities

$$0.00037\% \quad \text{for } q = 6\%, \quad \text{and with } 0.0025\% \quad \text{for } q = 8\%.$$

As a result, waiting for six blocks after the payment has been considered as secure in the sense that with realistic efforts it is practically impossible to succeed with double spending.

Remark Note that in the original work [5] it was assumed that the attacker can start the race having pre-mined one block. This leads to a different success probability

$$1 - F_{q, n}(n-1) + F_{1-q, n}(n-1). \tag{2.4}$$

While the difference between (2.4) and (2.3) can be significant (see Fig. 1), it is not clear how to achieve an advantage of being able to start the race with one block ahead of the official chain. The present note is devoted to this interesting question.

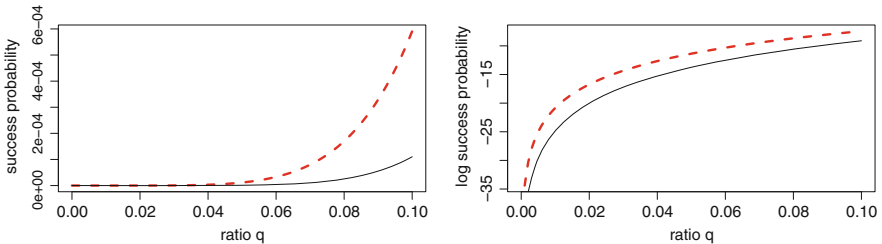


Fig. 1 The success probability (and its logarithm) of the double spending attack for $n = 6$ confirming blocks depending on the mining ratio $q \in [0, \frac{1}{10}]$ calculated by (2.3) (solid line) versus (2.4) (dashed line)

The above analysis [5] calculates the probability of the alternative blockchain getting ahead of the official one. It doesn't consider revenues and losses from a successful/failed attack. Furthermore, the possibility of canceling the secret mining (if the block difference becomes too high) is not considered. Most important, however, is the question why the paying transaction must be placed right after fork-off. Note that this assumption is justifiable only if the merchant requires immediate payment after the purchase is agreed upon, otherwise canceling the deal. However, in reality, the attacker may be able to freely choose the time of payment, in particular when buying goods from web portals. That is, an attempt to overtake the official chain before launching an attack can give an advantage in the spirit of the above remark.

3 A Refinement of the Double-Spending Problem

Let us consider an alternative situation. Assume that the attacker can freely choose the time of payment. Doing so, he/she can start working on a private secret branch long before the payment is placed. For such situations, the analysis of the double-spending is different and requires solving (multiple) stopping problems.

Consider a finite time horizon where $t \in \{0, \dots, T\}$ represents the number of blocks mined in the official chain since the branch has forked off. That is, we suppose that our secret mining starts at the block time $t = 0$. We interpret $T \in \mathbb{N}_0$ as the maximal length of the official branch, which can be abandoned if a longer branch has been discovered. To the best of our knowledge, the current Bitcoin protocol does not have such a restriction, meaning that the shorter branch must always be discarded, independently of its length. However, other blockchain systems discuss "checkpoints" and "gates" with a similar functionality. A finite time horizon yields conceptual advantages and presents a negligible deviation from the reality since T can be sufficiently large.

Recall the process $(Z_t)_{t=0}^\infty$, describing the branch length difference at times the next block in one of the branches has been mined. Now, consider another process $(X_t)_{t=0}^\infty$, where X_t stands for the branch length difference between the official and secret branches at the times $t = 0, 1 \dots$, when one new block in the *official* branch is completed. It turns out that $(X_t)_{t=0}^\infty$ follows a Markov chain, whose transition from state $X_t = x$ to $X_{t+1} = x + 1 - y$ describes the event that while one official block has been mined, $y = 0, 1, 2, \dots$ secret blocks have been obtained. The transition probabilities of $(X_t)_{t=0}^T$ are given for all $t = 0, 1 \dots$, by

$$\mathbb{P}(X_{t+1} = x + 1 - k | X_t = x) = \begin{cases} G(k), & k \in \mathbb{N}_0, \\ 0, & k \in \mathbb{Z} \setminus \mathbb{N}_0, \end{cases} \tag{3.1}$$

in terms of the geometric distribution

$$G(k) = (1 - q)q^k \quad \text{for } k \in \mathbb{N}_0 = \{0, 1, 2, \dots\}. \tag{3.2}$$

Suppose that the secret branch contains the invalidation of the paying transaction. Recall this can be reached by a simple non-inclusion of the attacker’s paying transaction. If the attack is launched at a block time $\tau = \{0, \dots, T\}$, then the payment will be included into block $\tau + 1$ of the official branch. In this context, the crucial question is whether to attack or not and how to choose the time $\tau = 0, \dots, T$ optimally. It turns out that under specific assumptions, this question can be treated as an optimal stopping problem, which we formulate next.

According to our modeling with a finite time horizon, we agree that for $\tau > T - n$ a successful attack is not possible. Namely, since the payment is placed into block $\tau + 1$ and n confirming blocks are expected, the last confirmation block $\tau + n > T$ would be beyond the maximal branch length which can be abandoned. That is, we can assume that the time τ must be chosen within the finite horizon $\tau = 0, \dots, \tilde{T}$ with the last time point $\tilde{T} = T - n$. The decision whether to attack must be based on the current block time $t = 0, \dots, \tilde{T}$ and on the recent block difference X_t . In order to optimize the time $\tau = 0, 1, \dots, \tilde{T}$, we define the success event $S(\tau)$ for the attack launched at τ as

$$S(\tau) = \left\{ \min_{i=\tau+n+1}^{T+1} X_i \leq 0 \right\} \quad \tau = 0, \dots, \tilde{T}.$$

Hence the expected reward of the attack is

$$\begin{aligned} R_\tau(x) &= \mathbb{E}(C 1_{S(\tau)} - c 1_{S(\tau)^c} | X_\tau = x) \\ &= (C + c)\mathbb{P}(S(\tau) | X_\tau = x) - c, \quad \tau = 0, \dots, \tilde{T}, \quad x \in \mathbb{Z} \end{aligned} \quad (3.3)$$

where the numbers $C > 0$ and $c > 0$ represent the revenue and loss resulting from the success or failure of the attack. Let us agree that $\tau = +\infty$ stands for the attacker’s option to not attack, which can be optimal if the chance of overtaking the official branch is too low. In order to model such an opportunity, we extend the reward function (3.3) for the time argument $t = \infty$ as

$$R_\infty(x) = 0, \quad x \in \mathbb{Z}. \quad (3.4)$$

In this context, the choice of the optimal payment time τ^* yields an optimal stopping problem of the following type:

$$\begin{aligned} &\text{determine a maximizer } \tau^* \text{ to } \mathcal{T} \rightarrow \mathbb{R}, \quad \tau \mapsto \mathbb{E}(R_\tau(X_\tau)) \text{ where} \\ &\mathcal{T} \text{ denotes all } \{0, \dots, \tilde{T}\} \cup \{+\infty\}\text{-valued stopping times.} \end{aligned} \quad (3.5)$$

4 Conclusion

Having assumed that the payment moment can be chosen by the attacker, the solution to the double spending problem consists of secret mining, followed by a later payment. The optimal payment time is determined by the current numbers of blocks in both chains since their fork off. The success probability of such an attack is dependent on the required confirmation block number n and the revenue/loss $C > 0$, $c > 0$ caused by the success/failure of the attack. This contribution shows that the optimization of the attack under these assumptions requires solving an optimal stopping problem. The authors will address these problems in future research.

Acknowledgement The first author expresses his warmest gratitudes to Kostya Borovkov for useful remarks.

References

1. Antonopoulos, A.M.: *Mastering Bitcoin: Programming the Open Blockchain*, 2nd edn. O'Reilly Media, Inc., Sebastopol (2017). ISBN 1491954388, 9781491954386
2. Goebel, J., Keeler, H.P., Krzesinski, A.E., Taylor, P.G.: Bitcoin blockchain dynamics: the selfish-mine strategy in the presence of propagation delay. *Perform. Eval.* **104**(Suppl. C), 23–41 (2016). ISSN 0166–5316. <http://www.sciencedirect.com/science/article/pii/S016653161630089X>
3. Gruenspan, C., Perez-Marco, R.: Double spend races. Working paper (2017)
4. Nakamoto, S.: A peer-to-peer electronic cash system. Working paper (2008)
5. Rosenfeld, M.: Analysis of hashrate-based double spending. Working paper (2014)

Stochastic Maximum Principle on a Continuous-Time Behavioral Portfolio Model



Qizhu Liang and Jie Xiong

Abstract In this short note, we consider the optimization problem with probability distortion when the objective functional involves a running term which is given by an S -shaped function. A stochastic maximum principle is presented.

1 Introduction

There are several epoch-making achievements in the history of finance theory over the past 70 years. The first is the expected utility maximization proposed by von Neumann and Morgenstern [17]. It is premised on the tenets that decision makers are rational and consistently risk averse under uncertainty. Later on, a Nobel-prize-winning work, Markowitz's mean-variance model [12] came out. Along with these theories in continuous portfolio selection problems, many approaches, such as dynamic programming, stochastic maximum principle, martingale and convex duality have been developed, see Merton [13], Peng [14], Duffie and Epstein [4], Yong and Zhou [19], Karatzas et al. [9].

On the account of substantial phenomena violating the basic tenets of conventional financial theory, for instance, Allais paradox [1], Tversky and Kahneman [16] put forward cumulative prospect theory (CPT) and Benartzi and Thaler [3] proposed behavioral economics. Both of them integrate psychology with finance and economics. To study the continuous-time portfolio choice problem, we concentrate on CPT in this paper. Its key elements are: (1) benchmark (evaluated at terminal time T) serves as a base point to distinguish gains from losses (Without loss of generality, it is assumed to be 0 in this paper); (2) Utility functions are concave for gains and convex for losses, and steeper for losses than for gains; (3) Probability distortions

Q. Liang

Department of Mathematics, University of Macau, Macau, China

e-mail: yb47422@connect.umac.mo

J. Xiong (✉)

Department of Mathematics, Southern University of Science and Technology, Shenzhen, China

e-mail: xiongj@sustc.edu.cn

(or weighting) are nonlinear transformation of the probability measures, which overweight small probabilities and underweight moderate and high probabilities.

There have been burgeoning research merge CPT into portfolio investment. Most of them are limited to the discrete-time setting, see for example Benartzi and Thaler [2], Shefrin and Statman [15], Levy and Levy [10]. The pioneering analytical research on continuous-time asset allocation featuring behavioral criteria is done by Jin and Zhou [7]. Since then, a few extensive works have been published, see He and Zhou [5, 6], Xu and Zhou [18], Jin and Zhou [8] and so on. Jin and Zhou developed a new theory to work out the optimal terminal value in a continuous-time CPT model. Nonetheless, their theory aims at a particular portfolio choice problem in a self-financing market.

This article is to deal with probability distortion for model with running utilities. In order to come closer to reality, bankruptcy is not allowed in our problem. The remainder is organized as follows. Next section will formulate a general continuous-time portfolio selection model under the CPT, featuring S-shaped utility functions and probability distortions. The stochastic maximum principle as well as a solvable example are finally presented.

2 Problem Formulation

Let $T > 0$ be a fixed time horizon and $(\Omega, \mathcal{F}, \mathbb{P}, \{\mathcal{F}_t\}_{t \geq 0})$ a filtered complete probability space on which is defined a standard \mathcal{F}_t -adapted m -dimensional Brownian motion $W_t \equiv (W_t^1, \dots, W_t^m)^\top$ with $W_0 = 0$. It is assumed that $\mathcal{F}_t = \sigma\{W_s : 0 \leq s \leq t\}$, augmented by all the null sets. Throughout this paper A^\top denotes the transpose of a matrix A ; a^\pm denote the positive and negative parts of the real number a .

We define a positive state process

$$\begin{cases} dX_t = b(t, u_t, X_t)dt + \sigma(t, u_t, X_t)dW_t \\ X_0 = x_0 > 0, \end{cases} \tag{2.1}$$

and the agent’s prospective functional

$$\begin{aligned} J(u.) = & \mathbb{E} \int_0^T (\zeta_+(u_t^+) \varpi'_+(1 - F_{u_t^+}(u_t^+)) - \zeta_-(u_t^-) \varpi'_-(1 - F_{u_t^-}(u_t^-))) dt \\ & + \mathbb{E} (l(X_T) w'(1 - F_{X_T}(X_T))), \end{aligned} \tag{2.2}$$

where $u.$ is a control process taking values in a convex set $U \subseteq \mathbb{R}$. According to CPT, the following assumptions will be in force throughout this paper, where x denotes the state variable, u denotes the control variable.

We make the following assumptions throughout this article.

(H.1) $b(\cdot, \cdot, \cdot) : [0, T] \times U \times \mathbb{R}^+ \rightarrow \mathbb{R}$, $\sigma(\cdot, \cdot, \cdot) : [0, T] \times U \times \mathbb{R}^+ \rightarrow \mathbb{R}$, are continuously differentiable with respect to (u, x) with Lipschitz continuous first derivatives. We further assume $b(t, u, 0) = \sigma(t, u, 0) = 0$.

(H.2) $\zeta_{\pm}(\cdot), l(\cdot) : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ are differentiable, strictly increasing, strictly concave, with $\zeta_{\pm}(0) = l(0) = 0$ and $\zeta'_{\pm}(0+) = l'(0+) = \infty$.

(H.3) $\varpi_{\pm}(\cdot), w(\cdot) : [0, 1] \rightarrow [0, 1]$, are differentiable and strictly increasing, with $\varpi_{\pm}(0) = w(0) = 0, \varpi_{\pm}(1) = w(1) = 1$. Moreover, the first derivatives of $\varpi_{\pm}(\cdot), w(\cdot)$ are all bounded.

Let

$$\mathcal{U} = \left\{ u : [0, T] \times \Omega \rightarrow U \mid u_t \text{ is } \mathcal{F}_t\text{-adapted and } \mathbb{E} \int_0^T |u_t|^4 dt < \infty \right\}.$$

Definition 1 A control process $u. \in \mathcal{U}$ is said to be admissible, and $(u., X.)$ is called an admissible pair, if

1. $X.$ is the unique solution of Eq. (2.1) under $u.$;
2. For any $t \in [0, T]$, the distribution functions of u_t^{\pm} are continuous except at 0;
3. $\mathbb{E} \int_0^T |\zeta_{\pm}(u_t^{\pm})\varpi'_{\pm}(1 - F_{u_t^{\pm}}(u_t^{\pm}))|^8 dt < \infty$.
4. $\mathbb{E} \int_0^T \left(\left| \frac{d}{du} \ln \zeta_{\pm}(u_t^{\pm}) \right|^8 + |\zeta''_{\pm}(u_t^{\pm})|^4 \right) dt < \infty$.

The set of all admissible controls is denoted by \mathcal{U}_{ad} .

Meanwhile, the following technical assumption for the terminal state are in force throughout this paper.

Assumption 1 The terminal state X_T corresponding to the control process $u. \in \mathcal{U}_{ad}$ is supposed to has continuous distribution function. Besides,

$$\mathbb{E} |l(X_T)w'(1 - F_{X_T}(X_T))|^8 + \mathbb{E} \left| \frac{d}{dx} \ln l(X_T) \right|^8 + \mathbb{E} |l''(X_T)|^4 < \infty. \tag{2.3}$$

Problem Our optimal control problem is to find $\bar{u}. \in \mathcal{U}_{ad}$ such that

$$J(\bar{u}.) = \max_{u. \in \mathcal{U}_{ad}} J(u.). \tag{2.4}$$

3 A Necessary Condition for Optimality

The current section presents our main result of the article. Let $(\bar{u}., \bar{X}.)$ be an optimal pair of the problem (2.4). We proceed to presenting the condition it must satisfy. To this end, we formulate the adjoint equation

$$\begin{cases} dp_t = -(b_x(t, \bar{u}_t, \bar{X}_t)p_t + \sigma_x(t, \bar{u}_t, \bar{X}_t)q_t)dt + q_t dW_t, \\ p_T = l'(\bar{X}_T)w'(1 - F_{\bar{X}_T}(\bar{X}_T)). \end{cases} \tag{3.1}$$

Here is the necessary condition we obtained for the optimality of the control.

Theorem 1 *If \bar{u} . is the optimal control with the state trajectory \bar{X} ., then there exists a pair (p, q) of adapted processes which satisfies (3.1) such that a.e. $t \in [0, T]$,*

$$p_t b_u(t, \bar{u}_t, \bar{X}_t) + \sigma_u(t, \bar{u}_t, \bar{X}_t) q_t = \begin{cases} -\zeta'_+(\bar{u}_t^+) \varpi'_+(1 - F_{\bar{u}_t^+}(\bar{u}_t^+)) & \text{if } \bar{u}_t > 0, \\ -\zeta'_-(\bar{u}_t^-) \varpi'_-(1 - F_{\bar{u}_t^-}(\bar{u}_t^-)) & \text{if } \bar{u}_t < 0, \end{cases} \quad a.s. \tag{3.2}$$

Recall the state equation (2.1) and the adjoint equation (3.1). Given an optimal control \bar{u} ., there exists a unique solution $\bar{X}(\bar{u}.)$ to the state equation. As p_T is known, the unique solution $(p(\bar{u}.), q(\bar{u}.)$ for the backward SDE (3.1) is obtained. Plugging $\bar{X}(\bar{u}.)$ and $(p(\bar{u}.), q(\bar{u}.)$ into (3.2), the optimal control \bar{u} . is narrowed to one of the solution of so obtained algebraic equation.

In what follows, we present a solvable example and compare the result with the one without probability distortions. The process u_t^\pm in the objective functional are replaced by $u_t^\pm X_t$, signifying the proportion of wealth process. We study a case with compounded cost function.

Example 1 Let $u_t, X_t > 0$, $b(t, u, x) = -ux$, and $\sigma(t, u, x) = x$. We take utility function $\zeta_+(x) = \frac{x^\alpha}{\alpha}$ ($0 < \alpha < 1$), and distortion function (see, Lopes [11])

$$\varpi_+(p) = \nu p^{\gamma+1} + (1 - \nu)[1 - (1 - p)^{\beta+1}], \quad \gamma, \beta \geq 0, 0 \leq \nu \leq 1.$$

Then,

$$dX_t = -u_t X_t dt + X_t dW_t, \quad X_0 = x_0,$$

and

$$J(u.) = \mathbb{E} \int_0^T \left(\frac{1}{\alpha} (u_t X_t)^\alpha \varpi'_+(1 - F_{u_t X_t}(u_t X_t)) + X_t \right) dt.$$

By Theorem 1, its optimal solution $(\bar{u}., \bar{X}.)$ should satisfy

$$p_t = (\bar{u}_t \bar{X}_t)^{\alpha-1} \varpi'_+(1 - F_{\bar{u}_t \bar{X}_t}(\bar{u}_t \bar{X}_t)), \quad a.e. t \in [0, T], a.s., \tag{3.3}$$

and

$$dp_t = (\bar{u}_t p_t - q_t - (\bar{u}_t \bar{X}_t)^{\alpha-1} \varpi'_+(1 - F_{\bar{u}_t \bar{X}_t}(\bar{u}_t \bar{X}_t)) \bar{u}_t - 1) dt + q_t dW_t, \quad p_T = 0.$$

It yields $p_t = T - t$, $q_t = 0$, $\forall t \in [0, T]$. Plugging back to equality (3.3), we obtain that $\bar{u}_t \bar{X}_t = \left(\frac{T-t}{(1-\nu)(\beta+1)} \right)^{1/(\alpha-1)}$, a.e. $t \in [0, T]$, a.s. Solving the state equation, we arrive at

$$\bar{u}_t = (T-t)^{1/(\alpha-1)} / V_t (x_0((1-\nu)(\beta+1)))^{1/(\alpha-1)} + \int_0^t \frac{(T-s)^{1/(\alpha-1)}}{V_s} ds, \quad a.s.,$$

where $V_t = \exp\{B_t - \frac{t}{2}\}$.

Acknowledgements This research is supported by Macao Science and Technology Development Fund FDCT 025/2016/A1 and Southern University of Science and Technology Start up fund Y01286220.

References

- Allais, M.: Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école américaine. *Econom. J. Econom. Soc.* **21**, 503–546 (1953). <https://doi.org/10.2307/1907921>
- Benartzi, S., Thaler, R.H.: Myopic loss aversion and the equity premium puzzle. *Q. J. Econ.* **110**, 73–92 (1995). <https://doi.org/10.3386/w4369>
- Benartzi, S., Thaler, R.H.: Behavioral economics and the retirement savings crisis. *Science* **339**, 1152–1153 (2013). <https://doi.org/10.1126/science.1231320>
- Duffie, D., Epstein, L.G.: Stochastic differential utility. *Econom. J. Econom. Soc.* **60**, 353–394 (1992). <https://doi.org/10.2307/2951600>
- He, X.D., Zhou, X.Y.: Portfolio choice under cumulative prospect theory: an analytical treatment. *Manag. Sci.* **57**, 315–331 (2011). <https://doi.org/10.1287/mnsc.1100.1269>
- He, X.D., Zhou, X.Y.: Portfolio choice via quantiles. *Math. Financ.* **21**, 203–231 (2011). <https://doi.org/10.1111/j.1467-9965.2010.00432.x>
- Jin, H.Q., Zhou, X.Y.: Behavioral portfolio selection in continuous time. *Math. Financ.* **18**, 385–426 (2008). <https://doi.org/10.1111/j.1467-9965.2008.00339.x>
- Jin, H.Q., Zhou, X.Y.: Greed, leverage, and potential losses: a prospect theory perspective. *Math. Financ.* **23**, 122–142 (2013). <https://doi.org/10.1111/j.1467-9965.2011.00490.x>
- Karatzas, I., Lehoczky, J.P., Shreve, S.E., Xu, G.L.: Martingale and duality methods for utility maximization in an incomplete market. *SIAM J. Control. Optim.* **29**, 702–730 (1991). <https://doi.org/10.1137/0329039>
- Levy, H., Levy, M.: Prospect theory and mean-variance analysis. *Rev. Financ. Stud.* **17**, 1015–1041 (2003). <https://doi.org/10.1093/rfs/hhg062>
- Lopes, L.L.: Between hope and fear: the psychology of risk. *Adv. Exp. Soc. Psychol.* **20**, 255–295 (1987). [https://doi.org/10.1016/S0065-2601\(08\)60416-5](https://doi.org/10.1016/S0065-2601(08)60416-5)
- Markowitz, H.: Portfolio selection. *J. Financ.* **7**, 77–91 (1952). <https://doi.org/10.1111/j.1540-6261.1952.tb01525.x>
- Merton, R.C.: Lifetime portfolio selection under uncertainty: the continuous-time case. *Rev. Econ. Stat.* **51**, 247–257 (1969). <https://doi.org/10.2307/1926560>
- Peng, S.G.: A general stochastic maximum principle for optimal control problems. *SIAM J. Control. Optim.* **28**, 966–979 (1990). <https://doi.org/10.1137/0328054>
- Shefrin, H., Statman, M.: Behavioral portfolio theory. *J. Financ. Quant. Anal.* **35**, 127–151 (2000). <https://doi.org/10.2307/2676187>

16. Tversky, A., Kahneman, D.: Advances in prospect theory: cumulative representation of uncertainty. *J. Risk Uncertain.* **5**, 297–323 (1992). <https://doi.org/10.1007/bf00122574>
17. Von Neumann, J., Morgenstern, O.: *Theory of Games and Economic Behavior*. Princeton University Press, Princeton (2007)
18. Xu, Z.Q., Zhou, X.Y.: Optimal stopping under probability distortion. *Ann. Appl. Probab.* **23**, 251–282 (2013). <https://doi.org/10.1214/11-AAP838>
19. Yong, J.M., Zhou, X.Y.: *Stochastic Controls: Hamiltonian Systems and HJB Equations*. Springer, New York (1999)

Number of Claims and Ruin Time for a Refracted Risk Process



Yanhong Li, Zbigniew Palmowski, Chunming Zhao, and Chunsheng Zhang

Abstract In this paper, we consider a classical risk model refracted at given level. We give an explicit expression for the joint density of the ruin time and the cumulative number of claims counted up to ruin time. The proof is based on solving some integro-differential equations and employing the Lagrange's Expansion Theorem.

1 Introduction

Between 20.11.2017 and 8.12.2017 an international research institute MATRIX in Creswick, Australia, run research program *Mathematics of Risk* during which four 5-h workshops were given. In particular, Z. Palmowski presented a workshop entitled *Ruin probabilities: exact and asymptotic results*. This paper is closely related with the topics introduced during his lectures.

The joint density of the ruin time and the numbers of claims counted until ruin time has been already studied for a classical risk process over last years. Dickson [3] derived special expression for it using probabilistic arguments. Landriault et al. [11] analyzed this object for the Sparre Andersen risk model with the exponential claims. Later Frostig et al. [6] generalized it to the case of a renewal risk model with the

Y. Li
Sichuan University, Chengdu, China
e-mail: yanhonglink@qq.com

Z. Palmowski
Faculty of Pure and Applied Mathematics, Wrocław University of Science and Technology,
Wrocław, Poland

C. Zhao (✉)
Department of Statistics, School of Mathematics, Southwest Jiaotong University, Chengdu,
Sichuan, China
e-mail: cmzhao@swjtu.cn

C. Zhang
School of Mathematical Sciences and LPMC Nankai University, Tianjin, China
e-mail: zhangcs@nankai.edu.cn

phase-type claims and inter-arrival times. The main tool used there was the duality between the risk model and a workload of a single server queueing model. Zhao and Zhang [20] considered a delayed renewal risk model, where the claim size is Erlang(n) distributed and the inter-arrival time is assumed to be infinitely divisible.

Our goal is to derive expression for the joint density of the ruin time and the numbers of claims counted until ruin time for a refracted classical risk process (see Kyrianiou and Loeffen [9] for a formal definition). It is also called a compound Poisson risk model under a threshold strategy. The latter process is a classical risk process whose dynamic is changed by subtracting off a fixed linear drift whenever the cumulative risk process is above a pre-specified level b . This subtracting of the linear drift corresponds to the dividend payments and the considered strategy is also known as a threshold strategy. Dividend strategies for insurance risk models were first proposed by De Finetti [2] to reflect more realistically the surplus cash flows in an insurance portfolio. More recently, many kind of risk related quantities under threshold dividend strategies have been studied by Lin and Pavlova [16], Zhu and Yang [22], Lu and Li [14, 15, 18], Badescu et al. [1], Gao and Yin [7] (see references therein). The case when the drift of the refracted process is disappearing (everything above threshold b is paid as dividends) is called barrier strategy, see Lin et al. [17], Li and Garrido [12], Zhou [21] and in the references therein.

The paper is organized as follows. In Sect. 2 we define the model we deal with in this paper. In Sect. 3 we recall properties of the translation operator and the root of the Lundberg fundamental equation. In particular, we introduce the Lagrange's expansion theorem and some notation. In Sect. 4 we construct two integro-differential equations identifying the joint Laplace transform of joint density of the numbers of claims counted up to ruin time and the ruin time. Analytical solutions of these two integro-differential equations are given in Sect. 5. Applying the Lagrange's expansion theorem in Sect. 6 we give the expression for above mentioned density.

2 Model

The classical risk process is given by

$$U(t) = u + c_1t - S(t), \quad (1)$$

where $U(0) = u$ denotes initial capital, c is the premium rate and $S(t) = \sum_{i=1}^{N_t} X_i$ represents the total amount of claims appeared up to time $t \geq 0$. That is, $\{X_i\}_{i \in \mathbb{N}}$ are non-negative i.i.d. random variables with pdf $f(x)$ and cdf $F(x)$ and $\{N_t\}_{t \geq 0}$ is an independent Poisson process with a parameter λ . To take into account dividend payments paid when regulated process (after deduction of dividends) is above fixed

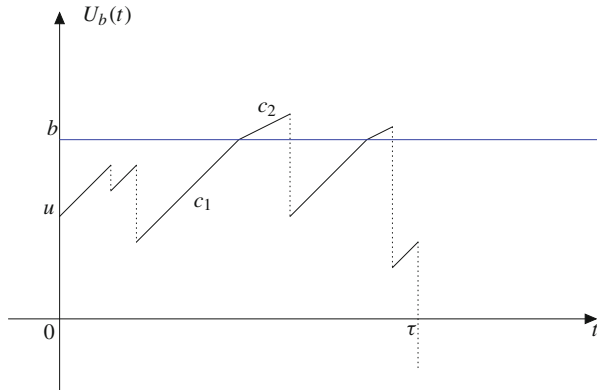


Fig. 1 Graphical representation of the surplus process $U_b(t)$

threshold level $b > 0$, we consider so-called refracted process given formally for $c_2 < c_1$ by:

$$dU_b(t) = \begin{cases} c_1 dt - dS(t), & 0 \leq U_b(t) \leq b \\ c_2 dt - dS(t), & U_b(t) > b \end{cases} \tag{2}$$

and $U_b(0) = u$. In this case $c_1 - c_2$ denotes intensity of dividend payments, see Fig. 1.

Throughout this paper, we will assume that $c_2 > \lambda EX_1$, which means refracted process $U_b(t)$ tends to infinity almost surely. We can then consider the ruin time:

$$\tau = \inf\{t > 0, U_b(t) < 0\},$$

($\tau = \infty$ if ruin does not occur). Note that N_τ represents the number of claims counted until the ruin time. The main goal of this paper is identification of the density of (τ, N_τ) . We start from analyzing its Laplace transform:

$$\phi(u) = E[r^{N_\tau} e^{-\delta\tau} \mathbb{I}(\tau < \infty) | U_b(0) = u] \tag{3}$$

$$= \sum_{n=1}^{\infty} r^n \int_0^{\infty} e^{-\delta t} w(u, n, t) dt, \tag{4}$$

where

$$w(u, n, t) = P(N_\tau = n, \tau \in dt | U_b(0) = u) / dt$$

is the joint density of (τ, N_τ) when $U_b(0) = u$. In above definition we have $\delta > 0$ and $r \in (0, 1]$. Later we will use the following notation

$$w_1(u, n, t) = w(u, n, t) \quad \text{for } u \leq b$$

and

$$w_2(u, n, t) = w(u, n, t) \quad \text{for } u > b.$$

3 Preliminaries

In this section we introduce few facts used further in this paper. We start from recalling the translation operator T_s ; see Dickson and Hipp [4]. For any integrable real-valued function f it is defined as

$$T_s f(x) = \int_x^\infty e^{-s(y-x)} f(y) dy, \quad x \geq 0.$$

The operator T_s satisfies the following properties:

1. $T_s f(0) = \int_0^\infty e^{-sx} f(x) dx = \hat{f}(s)$ which is the Laplace transform of f ;
2. The operator T_s is commutative, i.e. $T_s T_r = T_r T_s$. Moreover, for $s \neq r$ and $x \geq 0$

$$T_s T_r f(x) = T_r T_s f(x) = \frac{T_s f(x) - T_r f(x)}{r - s}. \tag{5}$$

More properties of the translation operator T_s can be found in Li and Garrido [13] and Gerber and Shiu [8].

For any function g we will denote by $\hat{g}(s)$ its Laplace Transform, that is $\hat{g}(s) = \int_0^\infty e^{-sx} g(x) dx$. Next, for $i = 1, 2$ let ρ_i be the positive root of the Lundberg fundamental equation

$$c_i s - (\lambda + \delta) + \lambda r \hat{f}(s) = 0. \tag{6}$$

The positive roots always exists for $\delta > 0$; see Fig. 2.

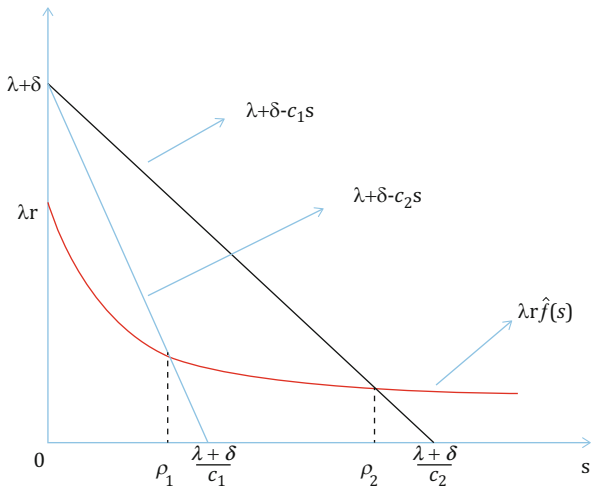
Lagrange’s Expansion Theorem In this paper we will also use the Lagrange’s Expansion Theorem; see pages 251–326 of Lagrange [10]. Given two functions $\alpha(z)$ and $\beta(z)$ which are both analytic on and inside a contour D surrounding a point a , if r satisfies the inequality

$$|r\beta(z)| < |z - a|, \tag{7}$$

for every z on the perimeter of D , then $z - a - r\varphi(z)$, as a function of z , has exactly one zero η in the interior of D , and we further have:

$$\alpha(\eta) = \alpha(a) + \sum_{k=1}^\infty \frac{r^k}{k!} \frac{d^{k-1}}{dx^{k-1}} (\alpha'(x)\beta^k(x)) \Big|_{x=a}. \tag{8}$$

Fig. 2 Roots for Lundberg’s fundamental equation



Finally, we define also the impulse function

$$\delta_x(t) = \begin{cases} 0, & t \neq x \\ \infty, & t = x \end{cases}$$

with $\int_0^\infty \delta_x(t)dt = 1$. We denote $g^{k*}, k \geq 0$, with $g^{1*} = g$ and $g^{0*}(t) = \delta_0(t)$ the k -fold convolution of g with itself, where

$$(g * h)(t) = \int_0^t g(x)h(t - x)dx, \quad t \geq 0$$

for two functions g and h supported on $(0, \infty)$.

4 Integro-Differential Equations for the Joint Laplace Transform

In this section, we derive two integro-differential equations identifying $\phi(u)$ defined in (3). We will follow the idea given in Lin and Pavlova [16]. Denote

$$\phi(u) = \begin{cases} \phi_1(u), & u \leq b, \\ \phi_2(u), & u > b. \end{cases} \tag{9}$$

Theorem 1 *The joint Laplace transform ϕ satisfies the following integro-differential equations:*

$$\begin{cases} \phi'_1(u) = \frac{\lambda+\delta}{c_1}\phi_1(u) - \frac{\lambda r}{c_1} \int_0^u \phi_1(u-x)f(x)dx - \frac{\lambda r}{c_1}\bar{F}(u), & 0 \leq u \leq b \\ \phi'_2(u) = \frac{\lambda+\delta}{c_2}\phi_2(u) - \frac{\lambda r}{c_2} \left(\int_0^{u-b} \phi_2(u-x)f(x)dx + \int_{u-b}^u \phi_1(u-x)f(x)dx \right) - \frac{\lambda r}{c_2}\bar{F}(u), & u > b \end{cases} \tag{10}$$

with the boundary condition

$$\phi_1(b) = \phi_2(b) := \lim_{u \rightarrow b^+} \phi_2(u). \tag{11}$$

Remark 1 Note that from the integro-differential equations (10) follows that the joint Laplace transform with initial surplus above the barrier depends on the respective function with initial surplus below the barrier, but the reverse relationship does not hold true.

Proof Let first $0 \leq u \leq b$. Then conditioning on the occurrence of the first claim we will have two cases: the first claim occurs before the surplus has reached the barrier level b or it occurs after reaching this barrier. There are also two other cases at the moment of the arrival of the first claim: either the risk process starts all over again with new initial surplus or the first claim leads already to ruin. Hence:

$$\begin{aligned} \phi(u) &= \phi_1(u) \\ &= \int_0^{\frac{b-u}{c_1}} \lambda r e^{-\lambda t} e^{-\delta t} \left(\int_0^{u+c_1 t} \phi(u+c_1 t-x)f(x)dx + \bar{F}(u+c_1 t) \right) dt \\ &\quad + \int_{\frac{b-u}{c_1}}^\infty \lambda r e^{-\lambda t} e^{-\delta t} \left(\int_0^{b+c_2(t-\frac{b-u}{c_1})} \phi(b+c_2(t-\frac{b-u}{c_1})-x)f(x)dx + \bar{F}(b+c_2(t-\frac{b-u}{c_1})) \right) dt \\ &= \lambda r \int_0^{\frac{b-u}{c_1}} e^{-(\lambda+\delta)t} \gamma(u+c_1 t) dt + \lambda r \int_{\frac{b-u}{c_1}}^\infty e^{-(\lambda+\delta)t} \gamma(b+c_2(t-\frac{b-u}{c_1})) dt, \end{aligned} \tag{12}$$

where $\gamma(t) = \int_0^t \phi(t-x)f(x)dx + \bar{F}(t)$.

Changing variables in (12) and rearranging leads to the following equation for $0 \leq u \leq b$:

$$\begin{aligned} \phi_1(u) &= \frac{\lambda r}{c_1} e^{(\lambda+\delta)u/c_1} \int_u^b e^{-(\lambda+\delta)t/c_1} \gamma(t) dt + \frac{\lambda r}{c_2} e^{(\lambda+\delta)u/c_1} \\ &\quad \times \int_b^\infty e^{-(\lambda+\delta)[t-(c_1-c_2)b/c_1]/c_2} \gamma(t) dt. \end{aligned} \tag{13}$$

Differentiating both sides of (13) with respect to u yields first equation.

Similarly, for $u > b$ we have:

$$\begin{aligned}
 \phi(u) &= \phi_2(u) \\
 &= \int_0^\infty \lambda r e^{-\lambda t} e^{-\delta t} \left(\int_0^{u+c_2t} \phi(u+c_2t-x) f(x) dx + \bar{F}(u+c_2t) \right) dt \\
 &= \lambda r \int_0^\infty e^{-(\lambda+\delta)t} \gamma(u+c_2t) dt \\
 &= \frac{\lambda r}{c_2} e^{(\lambda+\delta)u/c_2} \int_u^\infty e^{-(\lambda+\delta)t/c_2} \gamma(t) dt.
 \end{aligned} \tag{14}$$

Differentiating both sides of (14) with respect to u produces the second equation.

Note also that from Eqs. (13) and (14) it follows that $\phi(u)$ is continuous at $u = b$ and hence (11) holds. This completes the proof.

5 The Analytical Expression for $\phi(u)$

In this section, we derive the analytical expression for $\phi_i(u)$ ($i = 1, 2$) using the translation operator introduced in Sect. 3.

Theorem 2 *The function $\phi_2(u)$ can be expressed analytically as follows:*

$$\phi_2(u) = \sum_{n=0}^\infty \left(\frac{\lambda r}{c_2} \right)^{n+1} (T_{\rho_2} f)^{n*} * h(u-b), \quad u > b, \tag{15}$$

where

$$h(u) := \int_u^{u+b} \phi_1(u+b-x) T_{\rho_2} f(x) dx + T_{\rho_2} \bar{F}(u+b). \tag{16}$$

Proof We adopt the approach of Willmot and Dickson [19]. Consider the second equation in (10) for $u > b$. For a fixed $s > 0$, we multiply both sides of this equation by $e^{-s(u-b)}$ and integrate it with respect to u from b to ∞ :

$$\begin{aligned}
 &c_2 \int_b^\infty e^{-s(u-b)} \phi_2'(u) du \\
 &= (\lambda + \delta) T_s \phi_2(b) - \lambda r \int_b^\infty e^{-s(u-b)} \int_0^{u-b} \phi_2(u-x) f(x) dx du \\
 &\quad - \lambda r \int_b^\infty e^{-s(u-b)} \int_0^b \phi_1(y) f(u-y) dy du - \lambda r T_s \bar{F}(b)
 \end{aligned}$$

$$\begin{aligned}
&= (\lambda + \delta)T_s\phi_2(b) - \lambda r \int_0^\infty e^{-sx} f(x) \int_{x+b}^\infty e^{-s(u-x-b)} \phi_2(u-x) du dx \\
&\quad - \lambda r \int_0^b \phi_1(y) \int_b^\infty e^{-s(u-b)} f(u-y) du dy - \lambda r T_s \bar{F}(b) \\
&= (\lambda + \delta)T_s\phi_2(b) - \lambda r \hat{f}(s) T_s \phi_2(b) - \lambda r \int_0^b \phi_1(y) T_s f(b-y) dy - \lambda r T_s \bar{F}(b).
\end{aligned}$$

Integrating by parts gives:

$$c_2 \int_b^\infty e^{-s(u-b)} \phi_2'(u) du = c_2 s T_s \phi_2(b) - c_2 \phi_2(b).$$

Hence

$$\begin{aligned}
&c_2 s T_s \phi_2(b) - c_2 \phi_2(b) \\
&= (\lambda + \delta)T_s\phi_2(b) - \lambda r \hat{f}(s) T_s \phi_2(b) - \lambda r \int_0^b \phi_1(y) T_s f(b-y) dy - \lambda r T_s \bar{F}(b)
\end{aligned}$$

and simple rearranging leads to:

$$(c_2 s - (\lambda + \delta) + \lambda r \hat{f}(s)) T_s \phi_2(b) = c_2 \phi_2(b) - \lambda r \int_0^b \phi_1(y) T_s f(b-y) dy - \lambda r T_s \bar{F}(b). \quad (17)$$

Taking $s = \rho_2$ for the solution ρ_2 of the Lundberg Fundamental Equation (6) gives

$$c_2 \phi_2(b) = \lambda r \int_0^b \phi_1(y) T_{\rho_2} f(b-y) dy + \lambda r T_{\rho_2} \bar{F}(b).$$

Then Eq. (17) is equivalent to:

$$\begin{aligned}
&[c_2(s - \rho_2) + \lambda r \hat{f}(s) - \lambda r \hat{f}(\rho_2)] T_s \phi_2(b) \\
&= \lambda r \int_0^b \phi_1(y) [T_{\rho_2} f(b-y) - T_s f(b-y)] dy + \lambda r [T_{\rho_2} \bar{F}(b) - T_s \bar{F}(b)].
\end{aligned}$$

Now dividing above equation by $s - \rho_2$ and using property 2 of the translation operator introduced in Sect. 2 produces:

$$c_2 T_s \phi_2(b) = \lambda r T_s T_{\rho_2} f(0) T_s \phi_2(b) + \lambda r \int_0^b \phi_1(y) T_s T_{\rho_2} f(b-y) dy + \lambda r T_s T_{\rho_2} \bar{F}(b). \quad (18)$$

Inverting the translation operators of (18) yields the following renewal equation for $\phi_2(u)$:

$$\phi_2(u) = \frac{\lambda r}{c_2} \left[\int_0^{u-b} \phi_2(u-x) T_{\rho_2} f(x) dx + \int_{u-b}^u \phi_1(u-x) T_{\rho_2} f(x) dx + T_{\rho_2} \bar{F}(u) \right]. \tag{19}$$

Taking $y = u - b$ and $g(y) = \phi_2(y + b)$ we can rewrite (19) as follows:

$$g(y) = \frac{\lambda r}{c_2} \int_0^y g(y-x) T_{\rho_2} f(x) dx + \frac{\lambda r}{c_2} h(y), \quad y > 0,$$

where

$$h(y) = h(u - b) = \int_{u-b}^u \phi_1(u-x) T_{\rho_2} f(x) dx + T_{\rho_2} \bar{F}(u), \quad u > b.$$

Hence

$$\begin{aligned} \phi_2(u) &= g(y) \\ &= \frac{\lambda r}{c_2} \int_0^y g(y-x) T_{\rho_2} f(x) dx + \frac{\lambda r}{c_2} h(y) \\ &= \sum_{n=0}^{\infty} \left(\frac{\lambda r}{c_2} \right)^{n+1} (T_{\rho_2} f)^{n*} * h(y) \\ &= \sum_{n=0}^{\infty} \left(\frac{\lambda r}{c_2} \right)^{n+1} (T_{\rho_2} f)^{n*} * h(u - b) \end{aligned}$$

which completes the proof.

The expression for $\phi_1(u)$ could be also derived in terms of the translation operator.

Theorem 3 *The function $\phi_1(u)$ can be expressed analytically in the following form:*

$$\phi_1(u) = \phi_{\infty}(u) + \frac{\frac{\lambda r}{c_2} [\phi_{\infty} * T_{\rho_2} f(b) + T_{\rho_2} \bar{F}(b)] - \phi_{\infty}(b)}{v(b) - \frac{\lambda r}{c_2} v * T_{\rho_2} f(b)} v(u), \tag{20}$$

where

$$\phi_{\infty}(u) := \sum_{n=0}^{\infty} \left(\frac{\lambda r}{c_1} \right)^{n+1} (T_{\rho_1} f)^{n*} * T_{\rho_1} \bar{F}(u) \tag{21}$$

and

$$v(x) := \sum_{n=0}^{\infty} \left(\frac{\lambda r}{c_1}\right)^n (T_{\rho_1} f)^{n*} * p(x) \tag{22}$$

with $p(x) = e^{\rho_1 x}$.

Proof We will follow Landriault et al. [11]. Note that the first equation in (10) does not involve the barrier level b :

$$\phi_1'(u) = \frac{\lambda + \delta}{c_1} \phi_1(u) - \frac{\lambda r}{c_1} \int_0^u \phi_1(u-x) f(x) dx - \frac{\lambda r}{c_1} \bar{F}(u). \tag{23}$$

The information about the barrier b is included in the boundary condition:

$$\phi_1(b) = \phi_2(b) := \lim_{u \rightarrow b^+} \phi_2(u).$$

Lin et al. [16] showed that the general solution of (23) is of the form

$$\phi_1(u) = \phi_{\infty}(u) + kv(u), \tag{24}$$

where $\phi_{\infty}(u)$ is the joint Laplace transform of density of the ruin time and number of claims counted up to ruin time for the classical risk process (1) without any barrier applied. That is,

$$\phi_{\infty}(u) := \sum_{n=1}^{\infty} r^n \int_0^{\infty} e^{-\delta t} w_{\infty}(u, n, t) dt \tag{25}$$

for

$$w_{\infty}(u, n, t) := P(N_{\tau} = n, \tau \in dt | U(0) = u) / dt. \tag{26}$$

In above Eq.(24) the quantity k is a constant which we can specify by implementing (24) and (19):

$$k = \frac{\frac{\lambda r}{c_2} \left[\int_0^b \phi_{\infty}(b-x) T_{\rho_2} f(x) dx + T_{\rho_2} \bar{F}(b) \right] - \phi_{\infty}(b)}{v(b) - \frac{\lambda r}{c_2} \int_0^b v(b-x) T_{\rho_2} f(x) dx}. \tag{27}$$

We express now the function ϕ_{∞} in terms of a compound geometric distribution. Indeed, since ϕ_{∞} also satisfies Eq. (23), taking Laplace transforms of its both sides for sufficiently large s gives:

$$(c_1 s - (\lambda + \delta) + \lambda r \hat{f}(s)) \hat{\phi}_{\infty}(s) = c_1 \phi_{\infty}(0) - \lambda r \hat{F}(s), \quad s \geq 0. \tag{28}$$

To determine the constant term $c_1\phi_\infty(0)$ in (28), we substitute the solution ρ_1 of the Lundberg Fundamental Equation (6) for s :

$$c_1\phi_\infty(0) = \lambda r \hat{F}(\rho_1) = \lambda r T_{\rho_1} \hat{F}(0). \tag{29}$$

Consequently, the Eq. (28) reduces to

$$[c_1(s - \rho_1) + \lambda r \hat{f}(s) - \lambda r \hat{f}(\rho_1)]\hat{\phi}_\infty(s) = \lambda r \hat{F}(\rho_1) - \lambda r \hat{F}(s).$$

Dividing above equation by $s - \rho_1$ and simple rearranging along with implementation of the formula (5) produces:

$$c_1 \hat{\phi}_\infty(s) = \lambda r \hat{\phi}_\infty(s) T_s T_{\rho_1} f(0) + \lambda r T_s T_{\rho_1} \bar{F}(0).$$

Inverting this Laplace transforms gives classical renewal equation:

$$\phi_\infty(u) = \frac{\lambda r}{c_1} \phi_\infty * T_{\rho_1} f(u) + \frac{\lambda r}{c_1} T_{\rho_1} \bar{F}(u) \tag{30}$$

having the solution given as an Neumann infinite series (21).

To prove the last statement (22) note that the function $v(u)$ satisfies the following integro-differential equation:

$$c_1 v'(u) - (\lambda + \delta)v(u) + \lambda r \int_0^u v(u-x)f(x)dx = 0, \quad u \geq 0, \tag{31}$$

with the initial condition $v(0) = 1$. To get the analytical expression of $v(u)$ we take the Laplace transforms of both sides of (31) for sufficiently large s ($s > \rho_1$). This yields:

$$c_1 s \hat{v}(s) - c_1 v(0) = (\lambda + \delta)\hat{v}(s) - \lambda r \hat{f}(s)\hat{v}(s).$$

Since $v(0) = 1$,

$$(s + \frac{\lambda r}{c_1} \hat{f}(s) - \frac{\lambda + \delta}{c_1})\hat{v}(s) = 1. \tag{32}$$

Recalling that ρ_1 is the root of (6), we can rewrite (32) as

$$(s - \rho_1 + \frac{\lambda r}{c_1} [\hat{f}(s) - \hat{f}(\rho_1)])\hat{v}(s) = 1,$$

which, by dividing by $s - \rho_1$ and implementing (5), produces:

$$\hat{v}(s) = \frac{\lambda r}{c_1} \hat{v}(s) T_s T_{\rho_1} f(0) + \frac{1}{s - \rho_1}. \tag{33}$$

Inverting the Laplace transforms in (33) leads to the Eq. (22). Including all above identities in (24) completes the proof.

6 The Joint Density of (τ, N_τ)

In this section we give the joint density of the number of claims counted until ruin time and the ruin time using the Lagrange’s Expansion theorem. We start with few facts that will be useful in the proof of the main result.

Recall that by $w_\infty(u, n, t)$ we denote the joint density of (τ, N_τ) for the classical risk process (1) (with infinite barrier $b = +\infty$); see (26). For $i = 1, 2$ we denote

$$g_i(x, 0, t) := \delta_{x/c_i}(t) e^{-\lambda x/c_i},$$

$$g_i(x, n, t) := x t^{n-1} e^{-\lambda t} \lambda^n f^{n*}(c_i t - x) / n!.$$

Following Dickson [3] we can state the following lemma.

Lemma 1 *We have*

$$w_\infty(u, 1, t) = \lambda e^{-\lambda t} \bar{F}(u + c_1 t).$$

For $n = 1, 2, 3, \dots$ the following holds:

$$w_\infty(u, n + 1, t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t} \int_0^{u+c_1 t} f^{n*}(u + c_1 t - x) \lambda \bar{F}(x) dx$$

$$- c_1 \sum_{j=1}^n \int_0^t \frac{(\lambda s)^j}{j!} e^{-\lambda s} f^{j*}(u + c_1 s) w_\infty(0, n + 1 - j, t - s) ds,$$
(34)

where

$$w_\infty(0, n, t) = \frac{\lambda}{c_1} \int_0^{c_1 t} \bar{F}(x) g_1(x, n - 1, t) dx, \quad n = 1, 2, \dots \tag{35}$$

Proof Using Lagrange’s Expansion Theorem presented in Sect. 2 with $\alpha(z) = e^{-zx}$, $\beta(z) = -\frac{\lambda}{c_i} \hat{f}(s)$, $a = (\lambda + \delta)/c_i$ and $D = \{z \mid |z - a| \leq a\}$ ($i = 1, 2$) and the Lundberg fundamental equation (6) we can conclude the following identity:

$$\begin{aligned} e^{-\rho_i x} &= e^{-(\lambda+\delta)x/c_i} + \sum_{n=1}^{\infty} \frac{r^n}{n!} \frac{d^{n-1}}{ds^{n-1}} \left(-x e^{-sx} \left(-\frac{\lambda}{c_i} \hat{f}(s) \right)^n \right) \Big|_{s=(\lambda+\delta)/c_i} \\ &= e^{-(\lambda+\delta)x/c_i} + \sum_{n=1}^{\infty} \frac{r^n}{n!} \frac{d^{n-1}}{ds^{n-1}} \left((-1)^{n+1} \lambda^n x / c_i^n \int_0^{\infty} e^{-s(x+y)} f^{n*}(y) dy \right) \Big|_{s=(\lambda+\delta)/c_i} \\ &= e^{-(\lambda+\delta)x/c_i} + \sum_{n=1}^{\infty} \frac{\lambda^n r^n}{n! c_i^n} \int_0^{\infty} x(x+y)^{n-1} e^{-(\lambda+\delta)(x+y)/c_i} f^{n*}(y) dy. \end{aligned}$$

Substituting $t := (x + y)/c_i$ and rearranging leads to:

$$\begin{aligned} e^{-\rho_i x} &= e^{-(\lambda+\delta)x/c_i} + \sum_{n=1}^{\infty} r^n \frac{\lambda^n}{n!} \int_{x/c_i}^{\infty} x t^{n-1} e^{-\lambda t} e^{-\delta t} f^{n*}(c_i t - x) dt \\ &= \sum_{n=0}^{\infty} r^n \int_{x/c_i}^{\infty} e^{-\delta t} g_i(x, n, t) dt. \end{aligned} \tag{36}$$

Therefore,

$$\begin{aligned} T_{\rho_i} f(x) &= \int_x^{\infty} e^{-\rho_i(u-x)} f(u) du \\ &= \int_x^{\infty} \sum_{n=0}^{\infty} r^n \int_{(u-x)/c_i}^{\infty} e^{-\delta t} g_i(u-x, n, t) dt f(u) du \\ &= \sum_{n=0}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \int_x^{c_i t+x} f(u) g_i(u-x, n, t) du dt. \end{aligned} \tag{37}$$

Since $\phi_{\infty}(u)$ defined in (25) is the joint Laplace transform under the classical compound Poisson risk model without a barrier we can use Dickson [3] to complete the proof.

Moreover, the following result holds true.

Lemma 2 *The function $v(u)$ given in (22) equals*

$$v(u) = \sum_{n=0}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \varpi(u, n, t) dt, \tag{38}$$

where

$$\varpi(u, 0, t) := g_1(-u, 0, t),$$

$$\varpi(u, n, t) := \sum_{m=1}^n \left(\frac{\lambda}{c_1}\right)^m \int_0^{c_1 t} \int_0^u g_{c_1}(y, n - m, t) b_m(u - x, y + x) dx dy + g_{c_1}(-u, n, t), \quad n \geq 1$$

$$b_n(u, y) := \sum_{j=0}^{n-1} \binom{n}{j} \frac{(-1)^j}{\Gamma(n)} \int_0^u (u - x)^{n-1} f^{(n-j)*}(y + u - x) f^{j*}(x) dx.$$

Proof Our goal is to express $v(u)$ as the Laplace transform:

$$v(u) = \int_0^\infty e^{-\rho_1 t} \xi(u, t) dt. \tag{39}$$

We start from definition (22):

$$\begin{aligned} v(u) &= \sum_{n=0}^\infty \left(\frac{\lambda r}{c_1}\right)^n (T_{\rho_1} f)^{n*} * p(u) \\ &= \sum_{n=1}^\infty \left(\frac{\lambda r}{c_1}\right)^n \int_0^u (T_{\rho_1} f)^{n*}(u - x) e^{\rho_1 x} dx + e^{\rho_1 u}. \end{aligned} \tag{40}$$

Using Dickson and Willmot [5] we can obtain the following representation:

$$(T_{\rho_i} f)^{n*}(u) = \int_0^\infty e^{-\rho_i y} b_n(u, y) dy \tag{41}$$

for

$$b_n(u, y) := \sum_{j=0}^{n-1} \binom{n}{j} \frac{(-1)^j}{\Gamma(n)} \int_0^u (u - x)^{n-1} f^{(n-j)*}(y + u - x) f^{j*}(x) dx.$$

By (40)

$$\begin{aligned} v(u) &= \sum_{n=1}^\infty \left(\frac{\lambda r}{c_1}\right)^n \int_0^u \int_0^\infty e^{-\rho_1 y} b_n(u - x, y) dy e^{\rho_1 x} dx + e^{\rho_1 u} \\ &= \sum_{n=1}^\infty \left(\frac{\lambda r}{c_1}\right)^n \int_0^\infty e^{-\rho_1 t} \int_0^u b_n(u - x, t + x) dx dt \\ &\quad + \sum_{n=1}^\infty \left(\frac{\lambda r}{c_1}\right)^n \int_{-u}^0 e^{-\rho_1 t} \int_{-t}^u b_n(u - x, t + x) dx dt + \int_0^\infty e^{-\rho_1 t} \delta_{-u}(t) dt. \end{aligned}$$

Comparing the coefficients of $e^{-\rho_1 t}$ in (39) gives:

$$\xi(u, t) = \sum_{n=1}^{\infty} \left(\frac{\lambda r}{c_1}\right)^n \int_0^u b_n(u-x, t+x) dx + \delta_{-u}(t); \tag{42}$$

see also [11]. Using (36) and (42) in (39) we end up with:

$$\begin{aligned} v(u) &= \int_0^{\infty} e^{-\rho_1 y} \xi(u, y) dy + e^{\rho_1 u} \\ &= \int_0^{\infty} \sum_{n=0}^{\infty} r^n \int_{y/c_1}^{\infty} e^{-\delta t} g_{c_1}(y, n, t) dt \xi(u, y) dy \\ &= \sum_{n=0}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \int_0^{c_1 t} g_{c_1}(y, n, t) \xi(u, y) dy dt \\ &= \sum_{n=1}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \left(\sum_{m=1}^n \left(\frac{\lambda}{c_1}\right)^m \int_0^{c_1 t} \int_0^u g_{c_1}(y, n-m, t) b_m(u-x, y+x) dx dy + g_{c_1}(-u, n, t) \right) dt \\ &\quad + \int_0^{\infty} e^{-\delta t} g_{c_1}(-u, 0, t) dt \end{aligned}$$

which completes the proof.

Using above lemmas we will prove the main result of this paper.

Theorem 4 For $0 \leq u \leq b$ and $m > 1$ the joint density of the number of claims until ruin N_{τ} and the time to ruin τ is given by

$$\begin{aligned} w_1(u, 1, t) &= \frac{\lambda}{c_2} e^{-\lambda t} \bar{F}(c_2 t + b + \frac{c_2}{c_1}(u-b)) \\ w_1(u, m, t) &= e^{-\frac{\lambda b}{c_1}} \left[\sum_{n=1}^m \vartheta(u, m, n, t - \frac{b}{c_1}) - \sum_{n=1}^{m-1} \int_0^{t-\frac{b}{c_1}} \varsigma(b, m-n, t - \frac{b}{c_1} - z) w_1(u, n, z) dz \right], \end{aligned} \tag{43}$$

where for $n \geq 1$

$$\begin{aligned} \varsigma(b, 0, t) &:= \varpi(b, 0, t) = g_{c_1}(-b, 0, t), \\ \varsigma(b, n, t) &:= \varpi(b, n, t) - \sum_{m=0}^{n-1} \frac{\lambda}{c_2} \int_0^b \int_0^t \varpi(b-x, n-1-m, t-z) \int_x^{c_2 z+x} f(y) g_2(y-x, m, z) dy dz dx, \\ \gamma(b, 1, t) &:= \frac{\lambda}{c_2} \int_b^{c_2 t+b} \bar{F}(y) g_2(y-b, 0, t) dy, \\ \gamma(b, n, t) &:= \sum_{m=0}^{n-2} \frac{\lambda}{c_2} \int_0^b \int_0^t w_{\infty}(b-x, n-m-1, t-z) \int_x^{c_2 z+x} f(y) g_2(y-x, m, z) dy dz dx, \\ \vartheta(u, m, n, t) &:= \int_0^t \varsigma(b, m-n, t-z) w_{\infty}(u, n, z) + (\gamma(b, n, t-z) - w_{\infty}(b, n, t-z)) \varpi(u, m-n, z) dz. \end{aligned}$$

Proof In order to get the joint density $w(u, n, t)$, we have to take inverse Laplace transform with respect to δ rather than ρ_1 and ρ_2 . To do this we must find firstly the relationship between transforms with respect to ρ_1, ρ_2 and δ by applying the Lagrange’s Expansion theorem. For convenience, we will denote:

$$\chi(b) := v(b) - \frac{\lambda r}{c_2} v * T_{\rho_2} f(b). \tag{44}$$

Then we can rewrite (20) as follows:

$$\chi(b)\phi_1(u) = \chi(b)\phi_\infty(u) + \frac{\lambda r}{c_2} [\phi_\infty * T_{\rho_2} f(b) + T_{\rho_2} \bar{F}(b)] v(u) - \phi_\infty(b)v(u). \tag{45}$$

Putting (37) and (38) into (44) we will derive:

$$\begin{aligned} \chi(b) &= \sum_{n=0}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \varpi(b, n, t) dt - \frac{\lambda r}{c_2} \int_0^b v(b-x) T_{\rho_2} f(x) dx \\ &= \sum_{n=0}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \varpi(b, n, t) dt - \frac{\lambda r}{c_2} \sum_{n=0}^{\infty} r^n \sum_{m=0}^n \int_0^b \int_0^{\infty} e^{-\delta t} \varpi(b-x, n-m, t) dt \int_0^{\infty} e^{-\delta z} \\ &\quad \int_x^{c_2 z+x} f(y) g_2(y-x, m, z) dy dz dx \\ &= \sum_{n=0}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \varpi(b, n, t) dt - \sum_{n=1}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \left\{ \sum_{m=0}^{n-1} \frac{\lambda}{c_2} \int_0^b \int_0^t \varpi(b-x, n-1-m, t-z) \right. \\ &\quad \left. \int_x^{c_2 z+x} f(y) g_2(y-x, m, z) dy dz dx \right\} dt \\ &= \sum_{n=1}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \left\{ \varpi(b, n, t) - \sum_{m=0}^{n-1} \frac{\lambda}{c_2} \int_0^b \int_0^t \varpi(b-x, n-1-m, t-z) \int_x^{c_2 z+x} f(y) \right. \\ &\quad \left. g_2(y-x, m, z) dy dz dx \right\} dt + \int_0^{\infty} e^{-\delta t} \varpi(b, 0, t) dt \\ &= \sum_{n=0}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \zeta(b, n, t) dt. \end{aligned} \tag{46}$$

Similarly, using Lemma 1, we can check that:

$$\frac{\lambda r}{c_2} [\phi_\infty * T_{\rho_2} f(b) + T_{\rho_2} \bar{F}(b)] = \sum_{n=1}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \gamma(b, n, t) dt. \tag{47}$$

Using (38), (46) and (47) in (45) we obtain:

$$\begin{aligned} & \sum_{m=1}^{\infty} r^m \int_0^{\infty} e^{-\delta t} \sum_{n=1}^m \int_0^t \zeta(b, m-n, t-z) (w_1(u, n, z) - w_{\infty}(u, n, z)) dz dt \\ &= \sum_{m=1}^{\infty} r^m \int_0^{\infty} e^{-\delta t} \sum_{n=1}^m \int_0^t (\gamma(b, n, t-z) - w_{\infty}(b, n, t-z)) \varpi(u, m-n, z) dz dt \end{aligned}$$

or equivalently that

$$\begin{aligned} & \sum_{n=1}^m \int_0^t \zeta(b, m-n, t-z) w_1(u, n, z) dz \\ &= \sum_{n=1}^m \int_0^t \zeta(b, m-n, t-z) w_{\infty}(u, n, z) \\ & \quad + (\gamma(b, n, t-z) - w_{\infty}(b, n, t-z)) \varpi(u, m-n, z) dz. \end{aligned}$$

Now, if $m = 1$ then

$$\int_0^t \zeta(b, 0, t-z) w_1(u, 1, z) dz = \vartheta(u, 1, 1, t).$$

In this case

$$\begin{aligned} \int_0^t \delta_{-b/c_1}(t-z) e^{\frac{\lambda b}{c_1}} w_1(u, 1, z) dz &= e^{\frac{\lambda b}{c_1}} w_1(u, 1, t + \frac{b}{c_1}) \\ &= \frac{\lambda}{c_2} e^{-\lambda t} \bar{F}(c_2 t + b + \frac{c_2}{c_1} u) \end{aligned}$$

and

$$w_1(u, 1, t) = \frac{\lambda}{c_2} e^{-\lambda(t+\frac{b}{c_1})} \bar{F}(c_2 t + b + \frac{c_2}{c_1}(u - b)).$$

Similarly, if $m = 2$ then

$$\int_0^t \delta_{-b/c_1}(t-z) e^{\frac{\lambda b}{c_1}} w_1(u, 2, z) dz = \sum_{n=1}^2 \vartheta(u, 2, n, t) - \int_0^t \zeta(b, 1, t-z) w_1(u, 1, z) dz$$

and

$$w_1(u, 2, t) = e^{-\frac{\lambda b}{c_1}} \left[\sum_{n=1}^2 \vartheta(u, 2, n, t - \frac{b}{c_1}) - \int_0^{t-\frac{b}{c_1}} \zeta(b, 1, t - \frac{b}{c_1} - z) w_1(u, 1, z) dz \right].$$

Similarly we can prove the assertion for any $m > 1$.

Theorem 5 For $u > b$ and $m > 1$ the joint density of the number of claims until ruin N_τ and the time to ruin τ is given by

$$w_2(u, m, t) = \left(\frac{\lambda}{c_2}\right)^m \sum_{k=0}^{m-1} \sum_{n=0}^{m-k-1} \int_0^{u-b} \int_0^t \int_0^{c_2z} g_2(y, k, z) b_{m-k-n-1}(u-b-x, y) \varepsilon(x, n, t-z) dy dz dx, \tag{48}$$

where

$$\begin{aligned} \varepsilon(u, 0, t) &:= \int_{u+b}^{c_2t+u+b} \bar{F}(y) g_2(y-u-b, 0, t) dy, \\ \varepsilon(u, m, t) &:= \sum_{n=1}^m \int_u^{u+b} \int_0^t w_1(u+b-x, n, t-z) \int_x^{c_2z+x} f(y) g_2(y-x, m-n, z) dy dz dx \\ &\quad + \int_{u+b}^{c_2t+u+b} \bar{F}(y) g_2(y-u-b, m, z) dy, \quad n \geq 1. \end{aligned}$$

Proof To obtain an expression for $w_2(u, m, t)$ we first consider $h(x)$ defined in (16). Using (37) we can derive:

$$\begin{aligned} h(u) &= \int_u^{u+b} \sum_{m=1}^\infty r^m \int_0^\infty e^{-\delta t} w_1(u+b-x, m, t) dt \\ &\quad \times \sum_{n=0}^\infty r^n \int_0^\infty e^{-\delta z} \int_x^{c_2z+x} f(y) g_2(y-x, n, z) dy dz dx \\ &\quad + \sum_{n=0}^\infty r^n \int_0^\infty e^{-\delta t} \int_{u+b}^{c_2t+u+b} \bar{F}(y) g_2(y-u-b, n, t) dy dt \\ &= \sum_{n=1}^\infty r^n \int_0^\infty e^{-\delta t} \left[\sum_{m=1}^n \int_u^{u+b} \int_0^t w_1(u+b-x, m, t-z) \right. \\ &\quad \times \int_x^{c_2z+x} f(y) g_2(y-x, n-m, z) dy dz dx \\ &\quad \left. + \int_{u+b}^{c_2t+u+b} \bar{F}(y) g_2(y-u-b, n, z) dy \right] dt \tag{49} \end{aligned}$$

$$\begin{aligned} &+ \int_0^\infty e^{-\delta t} \int_{u+b}^{c_2t+u+b} \bar{F}(y) g_2(y-u-b, 0, t) dy dt \\ &= \sum_{n=0}^\infty r^n \int_0^\infty e^{-\delta t} \varepsilon(u, n, t) dt. \tag{50} \end{aligned}$$

Moreover, substituting (41), (49) and (36) into (15) gives:

$$\begin{aligned}
 \phi_2(u) &= \sum_{m=0}^{\infty} \left(\frac{\lambda r}{c_2}\right)^{m+1} \int_0^{u-b} (T_{\rho_2} f)^{m*}(u-b-x)h(x)dx \\
 &= \sum_{m=0}^{\infty} \left(\frac{\lambda r}{c_2}\right)^{m+1} \int_0^{u-b} \int_0^{\infty} e^{-\rho_2 y} b_m(u-b-x, y)dy \sum_{n=0}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \varepsilon(x, n, t)dt dx \\
 &= \sum_{m=0}^{\infty} \left(\frac{\lambda r}{c_2}\right)^{m+1} \int_0^{u-b} \int_0^{\infty} \sum_{k=0}^{\infty} r^k \int_{y/c_2}^{\infty} e^{-\delta z} g_2(y, k, z)dz b_m(u-b-x, y)dy \\
 &\quad \times \sum_{n=0}^{\infty} r^n \int_0^{\infty} e^{-\delta t} \varepsilon(x, n, t)dt dx \\
 &= \sum_{m=1}^{\infty} r^m \int_0^{\infty} e^{-\delta t} \left\{ \left(\frac{\lambda}{c_2}\right)^m \sum_{k=0}^{m-1} \sum_{n=0}^{m-k-1} \int_0^{u-b} \int_0^t \int_0^{c_2 z} g_2(y, k, z) b_{m-k-n-1}(u-b-x, y) \right. \\
 &\quad \left. \times \varepsilon(x, n, t-z) dy dz dx \right\} dt. \tag{51}
 \end{aligned}$$

Comparing Eqs. (51) and (4) completes the proof.

Acknowledgements Chunming Zhao is supported by the Fundamental Research Funds for the Central Universities (Grant No. 2682017CX065) and by the FP7 Grant PIRSES-GA-2012-318984. Zbigniew Palmowski is supported by the National Science Centre under the grant 2013/09/B/HS4/01496. Zbigniew Palmowski thanks the organizers of the wonderful program Mathematics of Risk for all work done to make this event happened. In particular, he is grateful to Kostya Borovkov (University of Melbourne) and Kais Hamza (Monash University) for all the help and nice discussions.

References

1. Badescu, A., Drekcic, S., Landriault, D.: Analysis of a threshold dividend strategy for a MAP risk model. *Scand. Actuar. J.* **4**, 227–247 (2007)
2. De Finetti, B.: Su un'impostazione alternativa della teoria collettiva del rischio. In: *Transactions of the XVth International Congress of Actuaries*, pp. 433–443 (1957)
3. Dickson, D.C.M.: The joint distribution of the time to ruin and the number of claims until ruin in the classical risk model. *Insur. Math. Econ.* **50**(3), 334–337 (2012)
4. Dickson, D.C.M., Hipp, C.: On the time to ruin for Erlang (2) risk processes. *Insur. Math. Econ.* **29**(3), 333–344 (2001)
5. Dickson, D.C.M., Willmot, G.E.: The density of the time to ruin in the classical Poisson risk model. *Astin Bull.* **35**(1), 45–60 (2005)
6. Frostig, E., Pitts, S.M., Politis, K.: The time to ruin and the number of claims until ruin for phase-type claims. *Insur. Math. Econ.* **51**(1), 19–25 (2012)
7. Gao, H., Yin, C.: The perturbed Sparre Andersen model with a threshold dividend strategy. *J. Comput. Appl. Math.* **220**(1–2), 394–408 (2008)
8. Gerber, H.U., Shiu, E.S.W.: The time value of ruin in a Sparre Andersen model. *North Am. Actuar. J.* **9**(2), 49–69 (2005)

9. Kyprianou, A.E., Loeffen, R.L.: Refracted Lévy processes. *Ann. Inst. H. Poincaré Probab. Stat.* **46**(1), 24–44 (2010)
10. Lagrange, J.L.: Nouvelle méthode pour résoudre les équations littérales par le moyen des séries. Chez Haude et Spener, Libraires de la Cour & de l'Académie royale (1770)
11. Landriault, D., Shi, T., Willmot, G.E.: Joint densities involving the time to ruin in the Sparre Andersen risk model under exponential assumptions. *Insur. Math. Econ.* **49**(3), 371–379 (2011)
12. Li, S., Garrido, J.: On a class of renewal risk models with a constant dividend barrier. *Insur. Math. Econ.* **35**(3), 691–701 (2004)
13. Li, S., Garrido, J.: On ruin for the Erlang (n) risk process. *Insur. Math. Econ.* **34**(3), 391–408 (2004)
14. Li, S., Lu, Y.: The distribution of total dividend payments in a Sparre Andersen model. *Statist. Probab. Lett.* **79**(9), 1246–1251 (2009)
15. Li, S., Lu, Y.: On the time and the number of claims when the surplus drops below a certain level. *Scand. Actuar. J.* **5**, 420–445 (2016)
16. Lin, X.S., Pavlova, K.P.: The compound Poisson risk model with a threshold dividend strategy. *Insur. Math. Econ.* **38**(1), 57–80 (2006)
17. Lin, X.S., Willmot, G.E., Drekić, S.: The classical risk model with a constant dividend barrier: analysis of the Gerber-Shiu discounted penalty function. *Insur. Math. Econ.* **33**(3), 551–566 (2003)
18. Lu, Y., Li, S.: The Markovian regime-switching risk model with a threshold dividend strategy. *Insur. Math. Econ.* **44**(2), 296–303 (2009)
19. Willmot, G.E., Dickson, D.C.M.: The Gerber-Shiu discounted penalty function in the stationary renewal risk model. *Insur. Math. Econ.* **32**(3), 403–411 (2003)
20. Zhao, C., Zhang, C.: Joint density of the number of claims until ruin and the time to ruin in the delayed renewal risk model with Erlang (n) claims. *J. Comput. Appl. Math.* **244**, 102–114 (2013)
21. Zhou, X.: On a classical risk model with a constant dividend barrier. *North Am. Actuar. J.* **9**(4), 95–108 (2005)
22. Zhu, J., Yang, H.: Ruin theory for a Markov regime-switching model under a threshold dividend strategy. *Insur. Math. Econ.* **42**(1), 311–318 (2008)

Numerical Approximations to Distributions of Weighted Kolmogorov-Smirnov Statistics via Integral Equations



Dan Wu, Lin Yee Hin, Nino Kordzakhia, and Alexander Novikov

Abstract We show that the distribution of two-sided weighted Kolmogorov-Smirnov (wK-S) statistics can be obtained via the solution of the system of two Volterra type integral equations for corresponding boundary crossing probabilities for a diffusion process. Based on this result we propose a numerical approximation method for evaluating the distribution of wK-S statistics. We provide the numerical solutions to the system of the integral equations which were also verified via Monte Carlo simulations.

1 Introduction

The applications of one-sided and two-sided weighted Kolmogorov-Smirnov (wK-S) statistical tests are ubiquitous in diverse areas of applications, including physics, finance, computational biology and Gene Set Enrichment Analysis (GSEA), see e.g. [8, 14, 21]. In some cases there exist modifications of the wK-S

D. Wu · L. Y. Hin

School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway, NSW, Australia

e-mail: dan.wu@uts.edu.au; LinYee.Hin@uts.edu.au

N. Kordzakhia

Department of Statistics, Macquarie University, North Ryde, NSW, Australia

e-mail: nino.kordzakhia@mq.edu.au

A. Novikov (✉)

School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway, NSW, Australia

Steklov Institute of Mathematics, Moscow, Russia

e-mail: Alex.Novikov@uts.edu.au

whose limit distributions (for large sample sizes) can be represented as the following random variable

$$D_{g,f} := \sup_{t \in T} \frac{|B_t - g(t)\xi|}{f(t)}, \tag{1.1}$$

where $g(t)$ and $f(t)$ are some deterministic functions of t , $B = \{B_t, t \in [0, 1]\}$ is a standard Brownian bridge, the random variable ξ is independent of B and has the standard normal distribution, $\xi \sim N(0, 1)$, $T \subseteq [0, 1]$. Note that analytical expressions for the distribution function $P\{D_{g,f} < x\}$, $x > 0$, are not available in closed form besides the classical case when $g(t) = 0$, $f(t) = 1$ (see Kolmogorov [15]).

Recent applications of wK-S in GSEA (see e.g. [6]) require the development of fast and accurate numerical approximations for the cumulative distribution functions (cdf) of $D_{g,f}$ for specific functions f and g .

This paper addresses the issues of approximating cdf of the $D_{g,f}$ under the following two important settings:

1. $f(t) = 1, g = \{g(t) = t^\alpha - t, 1/2 < \alpha < 1, t \in T = [0, 1]\}$,
2. $f(t) = \sqrt{t(1-t)}, t \in T = [a, b], g(t) = 0, 0 < a \leq b < 1$.

Our goal is to find accurate numerical approximations for the following corresponding tail distributions

$$P_1(x) := P\{\sup_{t \in T} |B_t - g(t)\xi| > x\}; P_2(x) := P\left\{ \sup_{t \in [a,b]} \frac{|B_t|}{\sqrt{t(1-t)}} > x \right\}.$$

Setting 1 was recently discussed in the context of GSEA, see e.g. [6, 16, 17]. The family of functions g is of special relevance there. In particular, the case $\alpha = 2/3$ in Setting 1 corresponds to GSEA analysis where the weights of the genes in question are replaced by their respective ranks obtained based on their expressions in typical experiments. Examples of such gene expression profiles are accessible from the Gene Expression Omnibus repository [12]. Note that $g = 0$ corresponds to the classical Kolmogorov-Smirnov test statistic where the closed-form expression for $P\{D_{0,1} < x\}$ is well known [15]. See also [10], [11] and [20] for a historical account.

Setting 2 corresponds to the wK-S test suggested by Anderson and Darling [1]. It is designed with the purpose to increase the sensitivity for the tails of empirical distributions compared to the classical K-S test. Some asymptotic result for the tail probabilities under this setting have been derived recently in [7].

In general, finding the distribution of $D_{g,f}$ is a computationally intensive numerical problem, a subject of pursuit for many different approaches, each with its own merits and shortcomings, that are devised specifically to address this problem from different perspectives. In [9] and [2], among others, the authors reduce the problem of approximating the extrema of modified Brownian bridges to finding boundary crossing probabilities (BCP) with respect to Brownian motion. In line

with this approach, piecewise linear boundaries were used to replace nonlinear boundaries and approximate the desired distribution by an n -dimensional integrals in a similar way as used in [3, 18], and [22]. The convenient feature of this approach, as demonstrated in [17] and [16] for the case of the one-sided version of wK-S, is a possibility to obtain analytical upper and lower bounds for the tail distributions of statistics which lead to fast and reasonably accurate approximations. However, for the case of the two-sided wK-S this approach requires substantial computational cost in situations when highly accurate approximations are needed.

Historically it was Kolmogorov [15] who first found the distribution of $D_{0,1}$ as a solution of a partial differential equation (PDE). Subsequently, Anderson and Darling in [1], had applied this approach for solving related problems in the construction of goodness-of-fit tests. Under Settings 1 and 2 it is possible to use the finite-difference schemes or finite element method to obtain numerical approximations. However, the PDE approach seems to be not computationally efficient due to the fact that not only function evaluations are required at larger number of discretised points, but substantially higher computational burden is incurred due to the element assembly process [4].

The technique discussed in our paper is inspired by the work of Peskir (see Theorem 2.2 in [19]) who derived an integral equation of Volterra type for BCP with one-sided boundaries. In this paper we are expanding this technique for BCP with two-sided boundaries deriving a system of two integral equations of Volterra type. Note that this system of integral equations was derived in a different way by Buonocore et al. [5] using a different approach. As a matter of fact our technique is applicable to all regular diffusion processes where transition probabilities are available in a closed form. The advantage of this approach is that a system of integral equations are rather straight forward to obtain for all one-dimensional diffusion processes and efficient numerical techniques can be easily developed to solve the equations. The complete results including the case of general diffusion processes will be presented in another publication.

The paper is organised as follows. In Sect. 2 we formulate the results which provide a system of two Volterra integral equations for $P\{D_{g,f} < x\}$. In Sect. 3 we describe the numerical algorithm and results as well as comparisons with Monte Carlo simulation.

2 Construction of a System of Integral Equations of Volterra Type

In this section, we formulate a general result on BCP for two-sided boundaries which will be used for deriving a system of integral equations to evaluate the distribution of $D_{g,f}$.

Note that in Setting 1 we need to find BCP of the following form

$$P_1(x, y) := P\{\sup_{t \in T} |B_t - g(t)y| < x\} = P\{L(t) \leq B_t \leq U(t), t \in [0, 1]\}$$

where

$$L(t) = -x + g(t)y, \quad U(t) = x + g(t)y, \quad x > 0, y \in R.$$

Having the function $P_1(x, y)$ the distribution of two-sided wK-S statistics can be evaluated via integration with respect the density function $\phi(y, \sigma) := \frac{1}{\sigma\sqrt{2\pi}}e^{-(y/\sigma)^2/2}$:

$$P_1(x) = 1 - P\{D_{g,1} < x\} = \int_{-\infty}^{\infty} \phi(y, 1)(1 - P_1(x, y))dy. \tag{2.1}$$

In Setting 2 a similar integral representation holds due to the following relations:

$$\begin{aligned} P_2(x) &= P\left(\sup_{t \in [a,b]} \frac{|B_t|}{\sqrt{t(1-t)}} > x\right) = P(|B_a| > x\sqrt{a(1-a)}) + \\ &P\left(\left\{|B_a| \leq x\sqrt{a(1-a)}\right\} \cap \left\{\sup_{t \in [a,b]} \frac{|B_t|}{\sqrt{t(1-t)}} > x\right\}\right) \\ &= 1 - \operatorname{erf}(x/\sqrt{2}) + \int_{-x\sqrt{t_a}}^{x\sqrt{t_a}} \phi(y, \sqrt{t(1-t)})(1 - P_2(x, y))dy \end{aligned} \tag{2.2}$$

with

$$\begin{aligned} P_2(x, y) &= P\{L(t) \leq B_t \leq U(t), t \in [a, b] | B_a = y\}, y \in R, \\ L(t) &= -x\sqrt{t(1-t)}, \quad U(t) = x\sqrt{t(1-t)}, \quad x > 0. \end{aligned}$$

To derive equations for BCP under the general setting for a general diffusion process $X = \{X_t, t \geq 0\}$ (defined on a suitable probability space) we set

$$\tau_L = \inf_{t \geq t_0} \{t : X_t \leq L(t); B_s < U(s), \forall s \in (t_0, t) | X_{t_0} = x_0\}, \tag{2.3}$$

$$\tau_U = \inf_{t \geq t_0} \{t : X_t \geq U(t); X_s > L(s), \forall s \in (t_0, t) | X_{t_0} = x_0\}, \tag{2.4}$$

$$\tau = \inf_{t \geq t_0} \{t : X_t \notin (L(t), U(t)) | X_{t_0} = x_0\} = \inf\{\tau^-, \tau^+\}. \tag{2.5}$$

Let $f_L(t|x_0, t_0)$, $f_U(t|x_0, t_0)$ and $f(t|x_0, t_0)$ be the densities of τ_L , τ_U and τ respectively (subject their existence) and $X_{t_0} = x_0$.

Now we state a modified version of Theorem 2.2 in [19] for the two-sided BCP.

Theorem 1 *Let X be a one-dimensional diffusion process with boundaries U and L being continuously differentiable functions satisfying inequalities $L(s) < x_0 < U(s)$ and $L(t) < U(t)$ for all $t > s$. Let F_U and F_L be the cumulative distribution functions of τ_U and τ_L respectively. The following system of integral equations*

$$P(G_1, t|x_0, s) = \int_s^t P(G_1, t|U(s), s)F_U(ds|x_0, s) + \int_s^t P(G_1, t|L(s), s)F_L(ds|x_0, s), \tag{2.6}$$

$$P(G_2, t|x_0, s) = \int_s^t P(G_2, t|U(s), s)F_U(ds|x_0, s) + \int_s^t P(G_2, t|L(s), s)F_L(ds|x_0, s), \tag{2.7}$$

hold for any measurable sets $G_1 \subseteq [U(t), \infty)$ and $G_2 \subseteq (-\infty, L(t)]$.

The proof of this result will be presented in a full version of this paper; we just mention that it is based on the use of the Chapman-Kolmogorov equation as a starting point.

In both Settings 1 and 2 we need to derive equations for the case when $X = B$ is a standard Brownian bridge. Since B is a Gauss-Markov process we have

$$P(y, t|x, s) \sim N\left(\frac{R(s, t)}{R(s, s)}x, R(t, t) - \frac{R^2(s, t)}{R(s, s)}\right) \tag{2.8}$$

where R is the covariance function of B . Using this representation, upon the substitution of the initial condition $B_{t_0} = x_0$ into the Eqs. (2.6) and (2.7). Letting

$$\Psi(y|x, s) = \Psi\left(\frac{y - \frac{1-t}{1-s}x}{\sqrt{\frac{(t-s)(1-t)}{(1-s)}}}\right), \quad \Phi(y|x, s) = \Phi\left(\frac{y - \frac{1-t}{1-s}x}{\sqrt{\frac{(t-s)(1-t)}{(1-s)}}}\right),$$

we have

$$\Psi(U(t)|x_0, t_0) = \int_{t_0}^t \Psi(U(t)|U(s), s)f_U(s|x_0, t_0) ds + \int_{t_0}^t \Psi(U(t)|L(s), s)f_L(s|x_0, t_0) ds, \tag{2.9}$$

$$\Phi(L(t)|x_0, t_0) = \int_{t_0}^t \Phi(L(t)|U(s), s)f_U(s|x_0, t_0) ds + \int_{t_0}^t \Phi(L(t)|L(s), s)f_L(s|x_0, t_0) ds \tag{2.10}$$

respectively, where $\Phi(x) = \int_{-\infty}^x \phi(z, 1) dz$ and $\Psi(x) = 1 - \Phi(x)$.

Since by assumptions $U(t)$ and $L(t)$ are differentiable, we have

$$\lim_{t \rightarrow s} \Psi(U(t), t|U(s), s) = \lim_{t \rightarrow s} \Phi(L(t), t|L(s), s) = \Psi(0) = \frac{1}{2}, \tag{2.11}$$

$$\lim_{t \rightarrow s} \Psi(U(t), t|L(s), s) = \lim_{t \rightarrow s} \Phi(L(t), t|U(s), s) = \Psi(-\infty) = 0. \tag{2.12}$$

Hence, the kernels $\Psi(\cdot)$ and $\Phi(\cdot)$ are non-singular and are differentiable with respect to t for the case $X = B$.

The system of integral equations (2.6) and (2.7) (and the corresponding Eqs. (2.9) and (2.10) for the case of Brownian bridge) are Volterra equations of the first kind; they can be reduced to Volterra integral equations of the second kind which are numerically more suitable.

Theorem 2 *Let $f_U(t|x_0, t_0)$ and $f_L(t|x_0, t_0)$ be the probability density functions of τ_U and τ_L respectively and let $p(y, t|x, s) = \frac{\partial}{\partial t} P(y, t|x, s)$, then*

$$\begin{aligned}
 f_U(t|x_0, t_0) &= 2p(y_1, t|x_0, t_0) - 2 \int_{t_0}^t p(y_1, t|U(s), s) f_U(s|x_0, t_0) ds \\
 &\quad - 2 \int_{t_0}^t p(y_1, t|L(s), s) f_L(s|x_0, t_0) ds, \tag{2.13}
 \end{aligned}$$

$$\begin{aligned}
 f_L(t|x_0, t_0) &= 2p(y_2, t|x_0, t_0) - 2 \int_{t_0}^t p(y_2, t|U(s), s) f_U(s|x_0, t_0) ds \\
 &\quad - 2 \int_{t_0}^t p(y_2, t|L(s), s) f_L(s|x_0, t_0) ds \tag{2.14}
 \end{aligned}$$

hold for any $y_1 \subseteq [U(t), \infty)$ and $y_2 \subseteq (-\infty, L(t)]$.

Equations (2.13) and (2.14) are obtained by differentiating (2.6) and (2.7) with respect to t and then using the relation

$$\lim_{s \rightarrow t} P(U(t), t|U(s), s) = \lim_{s \rightarrow t} P(L(t), t|L(s), s) = \frac{1}{2}.$$

Note that a similar approach for the one-sided case was used by Fortet [13].

It follows that for a standard Brownian bridge with the initial condition $B_{t_0} = x_0$, we have

$$\begin{aligned}
 f_U(t|x_0, t_0) &= 2 \frac{\partial \Psi(U(t)|x_0, t_0)}{\partial t} - 2 \int_{t_0}^t \frac{\partial \Psi(U(t)|U(s), s)}{\partial t} f_U(s|x_0, t_0) ds \\
 &\quad - 2 \int_{t_0}^t \frac{\partial \Psi(U(t)|L(s), s)}{\partial t} f_L(s|x_0, t_0) ds, \tag{2.15}
 \end{aligned}$$

and

$$\begin{aligned}
 f_L(t|x_0, t_0) = & 2 \frac{\partial \Phi(L(t)|x_0, t_0)}{\partial t} - 2 \int_{t_0}^t \frac{\partial \Phi(L(t)|U(s), s)}{\partial t} f_U(s|x_0, t_0) ds \\
 & - 2 \int_{t_0}^t \frac{\partial \Phi(L(t)|L(s), s)}{\partial t} f_L(s|x_0, t_0) ds.
 \end{aligned}
 \tag{2.16}$$

Due to properties (2.11) and (2.12), singularities in the denominator of the kernels can be removed.

3 Numerical Integration Procedure for Approximation BCP

For Settings 1 and 2 we use (2.1) and (2.2)

$$P_1(x) = \int_{-\infty}^{\infty} \int_0^1 \phi(y, 1)(f_U(t|0, 0) + f_L(t|0, 0))dy$$

and

$$P_2(x) = \operatorname{erfc}(x/\sqrt{2}) + \int_{-x\sqrt{t_a}}^{x\sqrt{t_a}} \int_a^b \phi(y, \sqrt{t(1-t)})(f_U(t|y, a) + f_L(t|y, a))dy$$

with boundaries L and U shown in Sect. 1 respectively. To calculate f_U and f_L we use (2.15) and (2.16).

Let $t_i = t_0 + ih, i = 1, \dots, m$, where h is the time step size of uniform discretisations. We use the Euler approximation to obtain $f_L(t_i) = f_L(t_i|x, s)$ and $f_U(t_i) = f_U(t_i|x, s)$ at an increasing sequence of knots t_i and the appropriate Gaussian quadrature for numerical integrations.

For each of the aforementioned cases, we preform N simulations and use m equally-spaced discretization in the time interval T , and we estimate the tail probabilities

$$p_{f,g;m}(x) := P \left[D_{f,g;m}^{(i)} \geq x \right] \approx \frac{1}{N} \sum_{i=1}^N I(D_{f,g;m}^{(i)} \geq x),$$

where $I(\cdot)$ is an indicator function.

Table 1 Setting 1: estimated tail probabilities $\hat{P}_1(x)$ using integral equations, compared to simulations $p_{f,g;m}(x)$

x	$\hat{P}_1(x)$	$p_{f,g;m}(x)$	$Var[p_{f,g;m}(x)] \times 10^7$	$ \hat{P}_1(x) - p_{f,g;m}(x) $
0.4	0.997467	0.997395	0.025982	0.000072
1.2	0.128036	0.127972	1.115952	0.000064
2.0	0.001056	0.001040	0.010389	0.000016
2.2	0.000219	0.000233	0.002329	0.000014

For Setting 1 we calculate $\hat{P}_1(x)$, the tail probabilities approximated using the aforementioned approximation for $p_{f,g;m}(x)$ with $m = 2^{10}$ discretised time steps and the usage of $n = 20$ Gauss-Hermite nodes in the numerical integrations. Table 1 contains the comparisons of $\hat{P}_1(x)$ and $p_{f,g;m}$.

For Setting 2 we calculate $\hat{P}_2(x)$, the tail probabilities approximated using the aforementioned approximation for $p_{f,g;m}(x)$ in this setting, with $m = 2^{10}$ discretised time steps and the usage of $n = 20$ Gauss-Legendre nodes in the numerical integrations. In Table 2 we compare $\hat{P}_2(x)$ to simulation results, and asymptotic estimators $\tilde{P}_2(x)$ generated from [7].

The numerical approximations using integral equations are performed on a Macintosh laptop computer running OS X with 8 GB RAM and 1600 MHz CPU, and computer programs are implemented in C++98. Each point of $\hat{p}_g(x)$ and $\hat{p}_t(x)$ in the tables takes 1.92 and 1.87 s respectively. The Monte Carlo simulations for all cases are carried out on a cluster computer with 28 parallel CPUs using $N = 10^7$ simulation runs and $m = 2^{19}$ equally-spaced discretization time intervals.

In both settings, tail probabilities estimated using integral equations and simulation approach differ only at the third decimal place and beyond, suggesting that the numerical integration approach delivers a level of accuracy comparable to those of the simulation approach despite the use of a comparatively coarser grain discretization time step, i.e., $m = 2^{10}$ in the former as opposed to $m = 2^{19}$ in the latter.

The major advantage of our method is the relative simplicity and fast calculations compared to other techniques. For approximations using integral equations, since discretization time steps of the order $m = 2^{10}$ is sufficient to deliver tail probability estimates comparable to those of simulation at $N = 10^7$ and $m = 2^{19}$, they can be evaluated in a modest computational framework to reduce costs.

Table 2 Setting 2: estimated tail probabilities $\hat{P}_2(x)$ using integral equations, compared to simulations $p_{f,g;m}(x)$; $\tilde{P}_2(x) = 1 - \tilde{A}(x) (1/\underline{u})^{-\theta_0(x)}$ is the asymptotic tail probability estimator in equations (9), (10) and (13) of Chicheportiche and Bouchaud [7] where $\tilde{A}(x) = (\text{erf}(x/\sqrt{2}))^2$, $\theta_0(x) = \sqrt{2/\pi} x e^{-x^2/2}$, and $1/\underline{u}$ is the sample size which is set to be 100 in this numerical example

x	$\hat{P}_2(x)$	$p_{f,g;m}(x)$	$\text{Var}[p_{f,g;m}(x)] \times 10^7$	$\tilde{P}_2(x)$	$ \hat{P}_2(x) - p_{f,g;m}(x) $	$ \tilde{P}_2(x) - p_{f,g;m}(x) $
1.2	0.997822	0.997820	0.021752	0.930700	0.000002	0.067120
2.0	0.695515	0.698015	2.107901	0.663005	0.002500	0.035010
2.8	0.175139	0.177285	1.458550	0.192957	0.002146	0.015672
3.6	0.018395	0.018750	0.183984	0.020708	0.000355	0.001958
4.4	0.000925	0.000969	0.009681	0.001032	0.000044	0.000063

Acknowledgements The authors are grateful to the International Mathematical Research Institute MATRIX for hosting and supporting the Mathematics of Risk program during which they worked on the results presented in this note.

This research was partially funded by the Australian Government through the Australian Research Council (project DP150102758).

References

1. Anderson, T.W., Darling, A.D.: Asymptotic theory of certain “goodness of fit” criteria based on stochastic processes. *Ann. Math. Stat.* **23**, 193–212 (1952)
2. Bischoff, W., Hashorva, E., Hüslér, J., Müller, F.: Exact asymptotics for boundary crossings of the brownian bridge with trend with application to the Kolmogorov test. *Ann. Inst. Stat. Math.* **55**, 849–864 (2003)
3. Borovkov, K., Novikov, A.: Explicit bounds for approximation rates of boundary crossing probabilities for the Wiener process. *J. Appl. Prob.* **42**, 82–92 (2005)
4. Brebbia, C.A., Dominguez, J.: Boundary element methods for potential problems. *Appl. Math. Model.* **1**, 372–378 (1977)
5. Buonocore, A., Giorno, V., Nobile, A., Ricciardi, L.: On the two-boundary first-crossing-time problem for diffusion processes. *J. Appl. Probab.* **27**(1), 102–114 (1990)
6. Champi, K., Ycart, B.: Weighted Kolmogorov-Smirnov testing: an alternative for gene set enrichment analysis. *Stat. Appl. Genet. Mol. Biol.* **14**, 279–293 (2015)
7. Chicheportiche, R., Bouchaud, J.-P.: Weighted Kolmogorov-Smirnov test: accounting for the tails. *Phys. Rev.* **86**, 041115 (2012)
8. Delande, D., Gay, J.C.: Quantum chaos and statistical properties of energy levels: numerical study of the hydrogen atom in a magnetic field. *Phys. Rev. Lett.* **57** (2006). Published 20 October 1986; Erratum *Phys. Rev. Lett.* **57**, 2877 (1986)
9. Durbin, J.: Boundary-crossing probabilities for the Brownian motion and Poisson processes and techniques for computing the power of the Kolmogorov-Smirnov test. *J. Appl. Probab.* **8**, 431–453 (1971)
10. Durbin, J.: *Distribution Theory for Tests Based on the Sample Distribution Theory*. SIAM, Philadelphia (1973)
11. Durbin, J.: The first-passage density of a continuous gaussian process to a general boundary. *J. Appl. Probab.* **22**, 99–122 (1985)
12. Edgar, R., Domrachev, M., Lash, A.E.: Gene expression omnibus: ncbi gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002)
13. Fortet, R.: Les fonctions aléatoires du type de markoff associées à certaines équations linéaires aux dérivées partielles du type parabolique. *J. Math. Pures. Appl.* **22**, 177–243 (1943)
14. Klein, J.P., Logan, B., Harhoff, M., Andersen, P.K.: Analyzing survival curves at a fixed point in time. *Stat. Med.* **26**, 4505–4519 (2007)
15. Kolmogorov, A.N.: Sulla Determinazione Empirica di una Legge di Distribuzione. *Giorn. Ist. Ital. degli Attuari.* **4**, 83–91 (1933)
16. Kordzakhia, N., Novikov, A.: Bounds and approximations for distributions of weighted Kolmogorov-Smirnov tests. In: *From Statistics to Mathematical Finance*, pp. 235–250. Springer, Cham (2017)
17. Kordzakhia, N., Novikov, A., Ycart, B.: Approximations for weighted Kolmogorov-Smirnov distributions via boundary crossing probabilities. *Stat. Comput.* **27**, 1–11 (2016)
18. Novikov, A., Frishling, V., Kordzakhia, N.: Approximations of boundary crossing probabilities for a brownian motion. *J. Appl. Probab.* **36**, 1019–1030 (1999)
19. Peskir, G.: On integral equations arising in the first-passage problem for Brownian motion. *J. Integral Equ. Appl.* **14**, 397–423 (2002)

20. Stephens, M.A.: Introduction to Kolmogorov (1933) on the empirical determination of a distribution. In: *Breakthroughs in Statistics*, pp. 93–105. Springer, New York (1992)
21. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005)
22. Wang, L., Pötzelberger, K.: Boundary crossing probability for Brownian motion and general boundaries. *J. Appl. Probab.* **34**, 54–65 (1997)

Introduction to Extreme Value Theory: Applications to Risk Analysis and Management



Marie Kratz

Abstract We present an overview of Univariate Extreme Value Theory (EVT) providing standard and new tools to model the tails of distributions. One of the main issues in the statistical literature of extremes concerns the tail index estimation, which governs the probability of extreme occurrences. This estimation relies heavily on the determination of a threshold above which a Generalized Pareto Distribution (GPD) can be fitted. Approaches to this estimation may be classified into two classes, one qualified as ‘supervised’, using standard Peak Over Threshold (POT) methods, in which the threshold to estimate the tail is chosen graphically according to the problem, the other class collects unsupervised methods, where the threshold is algorithmically determined.

We introduce here a new and practically relevant method belonging to this second class. It is a self-calibrating method for modeling heavy tailed data, which we developed with N. Debbabi and M. Mboup. Effectiveness of the method is addressed on simulated data, followed by applications in neuro-science and finance. Results are compared with those obtained by more standard EVT approaches.

Then we turn to the notion of dependence and the various ways to measure it, in particular in the tails. Through examples, we show that dependence is also a crucial topic in risk analysis and management. Underestimating the dependence among extreme risks can lead to serious consequences, as for instance those we experienced during the last financial crisis. We introduce the notion of copula, which splits the dependence structure from the marginal distribution, and show how to use it in practice. Taking into account the dependence between random variables (risks) allows us to extend univariate EVT to multivariate EVT. We only give the first steps of the latter, to motivate the reader to follow or to participate in the increasing research development on this topic.

M. Kratz (✉)
ESSEC Business School, CREAR, Paris, France
e-mail: kratz@essec.edu

1 Introduction

Quantitative risk analysis used to rely, until recently, on classical probabilistic modeling where fluctuations around the average were taken into account. The standard deviation was the usual way to measure risk, like, for instance, in Markowitz portfolio theory [26], or in the Sharpe Ratio [34]. The evaluation of “normal” risks is more comfortable because it can be well modelled and predicted by the Gaussian model and so is easily insurable. The series of catastrophes that hit the World at the beginning of this century (see Fig. 1), natural (earthquakes, volcano eruption, tsunamis, ...) or financial (subprime crisis, sovereign crisis) or political (Arab Spring, ISIS, Ukraine, ...), made it clear that it is crucial nowadays to take also extreme occurrences into account; indeed, although it concerns events that rarely occur (i.e. with a very small probability), their magnitude is such that their consequences are dramatic when they hit unprepared societies.

Including extreme risks in probabilistic models is recognized nowadays as a necessary condition for good risk management in any institution, and not restricted anymore to reinsurance companies, who are the providers of covers for natural catastrophes. For instance in finance, minimizing the impact of extreme risks, or even ignoring them because of a small probability of occurrence, has been considered by many professionals and supervisory authorities, as a factor of

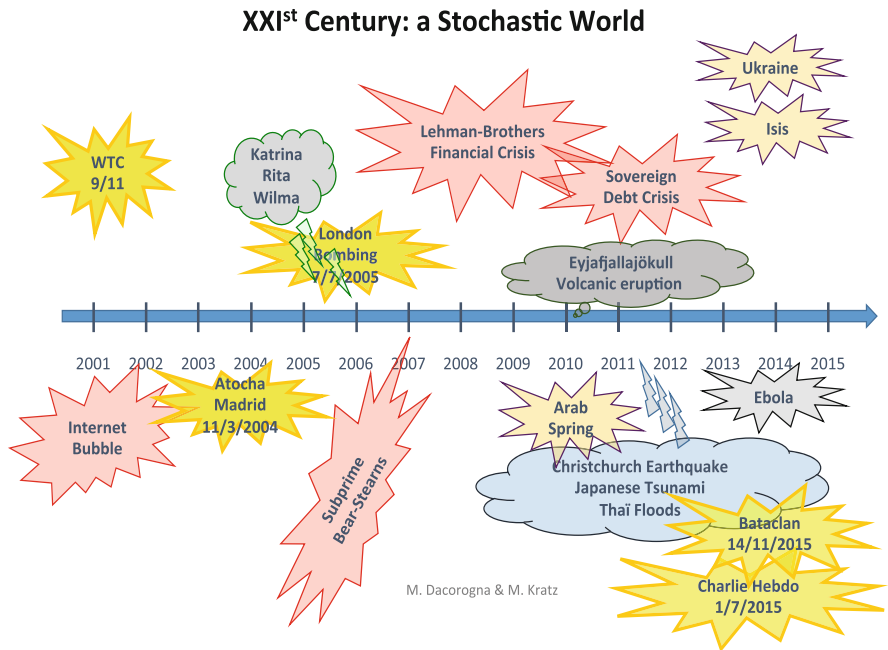


Fig. 1 Some of the extreme events (covered by reinsurances) that hit the World between 2001 and 2015

aggravation of the financial crisis of 2008–2009. The American Senate and the Basel Committee on Banking Supervision confirm this statement in their reports. Therefore, including and evaluating correctly extreme risks has become very topical and crucial for building up the resilience of our societies.

The literature on extremes is very broad; we present here an overview of some standard and new methods in univariate Extreme Value Theory (EVT) and refer the reader to books on the topic [1, 8, 12, 24, 31–33] and also on EVT with applications in finance or integrated in quantitative risk management [25, 27, 29]. Then we develop the concept of dependence to extend univariate EVT to multivariate EVT. All along applications in various fields including finance, insurance and quantitative risk management, illustrate the various concepts or tools.

1.1 What Is Risk?

Risk is a word widely used by many people and not only by professional risk managers. It is therefore useful to spend a bit of time analysing this concept. We start by looking at its definition in common dictionaries. There, we find that it is mainly identified to the notion of danger of loss:

The Oxford English Dictionary: Hazard, a chance of bad consequences, loss or exposure to mischance.

For financial risks: “Any event or action that may adversely affect an organization’s ability to achieve its objectives and execute its strategies” or, alternatively, “the quantifiable likelihood of loss or less-than-expected returns”.

Webster’s College Dictionary (insurance): “The chance of loss” or “The degree of probability of loss” or “The amount of possible loss to the insuring company” or “A person or thing with reference to the risk involved in providing insurance” or “The type of loss that a policy covers, as fire, storm, etc.”

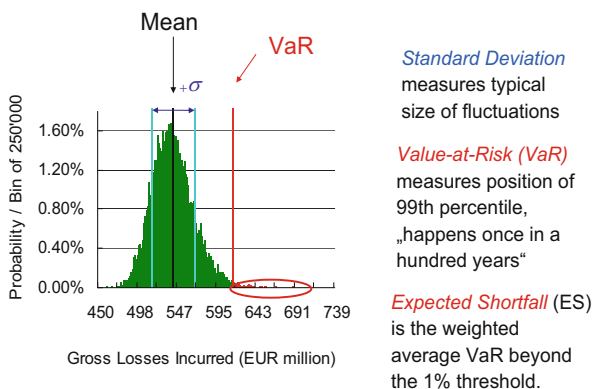
However, strictly speaking, risk is not simply associated to a danger. In its modern acceptance, it can also be seen as an opportunity for a profit. It is the main reason why people would accept to be exposed to risk. In fact, this view started to develop already in the eighteenth century. For instance, the French philosopher, Etienne de Condillac (1714–1780), defined risk as “The chance of incurring a bad outcome, coupled, with the hope, if we escape it, to achieve a good one”.

Another concept born from the management of risk is the insurance industry. In the seventeenth century, the first insurance for buildings is created after the big London fire (1666). During the eighteenth century appears the notion that social institutions should protect people against risk. This contributed to the development of life insurance that was not really acceptable by religion at the time. In this context, the definition of ‘Insurance’ as the transfer of risk from an individual to a group (company) takes its full meaning. The nineteenth century sees the continuous development of private insurance companies.

Independently of any context, risk relates strongly to the notion of randomness and the uncertainty of future outcomes. The distinction between “uncertainty” and “risk” was first introduced by the American economist Frank H. Knight (1885–1972), although they are related concepts. According to Knight, risk can be defined as randomness with knowable probabilities, contrary to uncertainty, which is randomness with unknowable probabilities. We could then define risk as a measurable uncertainty, the ‘known-unknowns’ according to Donald Rumsfeld’s terminology, whereas uncertainty is unmeasurable, the ‘unknown-unknowns’ (D. Rumsfeld). Of course, research and improved knowledge help to transform some uncertainty into risk with knowable probabilities.

In what follows, we focus on the notion of risk, in particular extreme risk, and its quantification. We choose a possible definition of risk, used within probabilistic framework, namely the variation from the expected outcome over time.

There are numerous ways to measure risk and many risk measures have been developed in the literature. Most modern measures of the risk in a portfolio are statistical quantities describing the conditional or unconditional loss distribution of the portfolio over some predetermined horizon. Here we present only popular ones, used in particular in regulation. We recall their definition and refer to Emmer et al. [13] and references therein for their mathematical properties.



Standard Deviation σ : Portfolio theory (Markowitz)

Value-at-Risk (VaR_α) = quantile $q_\alpha(L) = \inf\{q \in \mathbb{R} : \mathbb{P}(L > q) \leq 1 - \alpha\}$

Expected Shortfall (ES) (or TVaR)

$$\begin{aligned}
 ES_\alpha(L) &= \frac{1}{1 - \alpha} \int_\alpha^1 q_\beta(L) d\beta \\
 &= \underset{L \text{ cont}}{E}[L \mid L \geq q_\alpha(L)]
 \end{aligned}$$

We will focus on the analysis of extreme risks, related to unexpected, abnormal or extreme outcomes.

Many questions arise as: How to model extreme risks? How to study the behavior in the tails of the distribution of the model? How to capture dependency and measure it in risk models? which methods can be used? What about aggregation of risks?

First we present the main concepts of univariate EVT. Then we introduce the issue of dependence among random variables (risks).

1.2 Impact of Extreme Risks

When considering financial assets, because of the existence of a finite variance, a normal approximation is often chosen in practice for the unknown distribution of the yearly log returns, justified by the use of the Central Limit Theorem (CLT), when assuming independent and identically distributed (iid) observations. Such a choice of modeling, in particular using light tail distributions, has shown itself grossly inadequate during the last financial crisis when dealing with risk measures because it leads to underestimating the risk.

On Fig. 2, the QQ-plot of the S&P 500 daily returns from 1987 to 2007, helps to detect a heavy tail. When aggregating the daily returns into monthly returns, the QQ-plot looks more as a normal one, and the very few observations appearing above the threshold of $\text{VaR}_{99\%}$, among which the financial crises of 1998 and 1987, could almost be considered as outliers, as it is well known that financial returns are almost symmetrically distributed. Now, look at Fig. 3. When adding data from 2008 to 2013, the QQ plot looks pretty the same, i.e. normal, except that another “outlier” appears . . . with the date of October 2008! Instead of looking again on daily data for the same years, let us consider a larger sample of monthly data from 1791 to 2013 (as compiled by *Global Finance Data*). With a larger sample size, the heavy

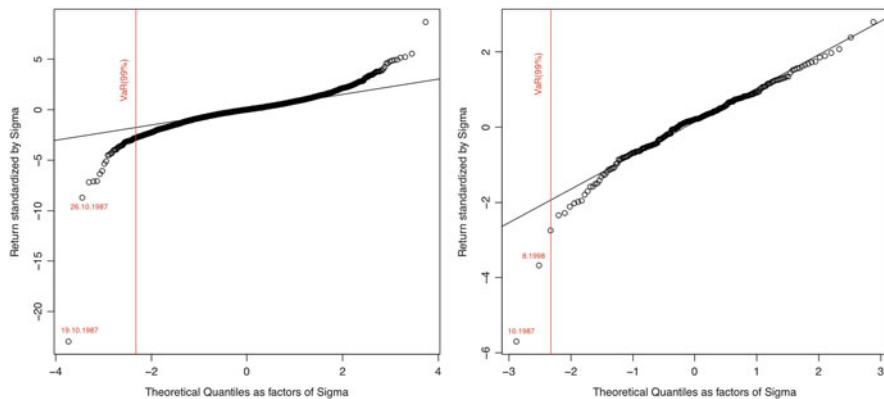


Fig. 2 QQ-plots of the S&P500 daily (left) and monthly (right) log-returns from 1987 to 2007

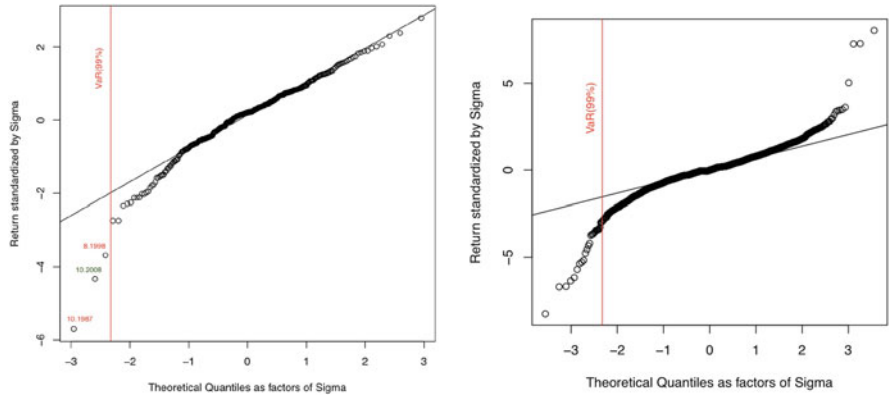


Fig. 3 QQ-plots of the S&P500 monthly log-returns from 1987 (left) or 1791 (right) to 2013

tail becomes again visible. And now we see that the financial crisis of 2008 does belong to the heavy tail of the distribution and cannot be considered anymore as an outlier. Although it is known, by Feller theorem, that the tail index of the underlying distribution remains constant under aggregation, we clearly see the importance of the sample size to make the tail visible. The figures on the S&P 500 returns illustrate very clearly this issue.

2 Univariate EVT

Let $(X_i)_{i=1, \dots, n}$ be iid random variables (rv) with parent rv X and continuous cumulative distribution function (cdf) F (that is unknown). The associated order statistics are denoted by $\min_{1 \leq i \leq n} (X_i) = X_{1,n} \leq X_{2,n} \leq \dots \leq X_{n-1,n} \leq X_{n,n} = \max_{1 \leq i \leq n} (X_i)$.

2.1 CLT Versus EVT

- *Mean behavior.* Assuming the existence of the variance σ^2 of X , the Central Limit Theorem (CLT) tells us that the empirical mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, when normalized (since $var(\bar{X}_n) = \frac{1}{n} var(X) \xrightarrow{n \rightarrow \infty} 0$), has an asymptotic standard Gaussian distribution (whatever is F):

$$\frac{\bar{X}_n - \mathbb{E}(\bar{X}_n)}{\sqrt{var(\bar{X}_n)}} = \sqrt{\frac{n}{var(X)}} (\bar{X}_n - \mathbb{E}(X)) = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sqrt{\sigma^2}} \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, 1)$$

i.e. $\lim_{n \rightarrow \infty} \mathbb{P}[(\bar{X}_n - b_n)/a_n \leq x] = F_{\mathcal{N}(0,1)}(x)$ with $b_n = \mathbb{E}(X)$, $a_n = \sqrt{\frac{\text{var}(X)}{n}}$.

- *Extreme behavior.* Instead of looking at the mean behavior, consider now the extreme behavior, with for instance the maximum. Noticing that

$$\mathbb{P}[\max_{1 \leq i \leq n} X_i \leq x] = \prod_{i=1}^n \mathbb{P}[X_i \leq x] = F^n(x) \xrightarrow{n \rightarrow \infty} \begin{cases} 0 & \text{if } F(x) < 1 \\ 1 & \text{if } F(x) = 1, \end{cases}$$

could we find, as for the CLT, a linear transformation to avoid such degeneracy, and say that there exist sequences (a_n) , (b_n) and a rv Z with cdf H such that $\lim_{n \rightarrow \infty} \mathbb{P}[(\max X_i - b_n)/a_n \leq x] = H(x)$? It comes back to look for (a_n) and (b_n) , and a non-degenerated cdf H s.t.

$$\mathbb{P}\left[\frac{\max X_i - b_n}{a_n} \leq x\right] = \mathbb{P}[\max_{1 \leq i \leq n} X_i \leq a_n x + b_n] = F^n(a_n x + b_n) \underset{n \rightarrow \infty}{\simeq} H(x).$$

It can be proved that there is not a unique limit distribution as for the CLT, but three possible asymptotic distributions (whatever is F), namely:

Theorem 1 (The ‘Three-Types Theorem’; Fréchet-Fisher-Tippett Theorem, 1927–1928; [15]) *The rescaled sample extreme (max renormalized) has a limiting distribution H that can only be of three types:*

$$\begin{aligned} H_{1,a}(x) &:= \exp\{-x^{-a}\} \mathbb{1}_{(x>0)} && (a > 0) : \mathbf{Fréchet} \\ H_{2,a}(x) &:= \mathbb{1}_{(x \geq 0)} + \exp\{-(-x)^a\} \mathbb{1}_{(x < 0)} && (a > 0) : \mathbf{Weibull} \\ H_{3,0}(x) &:= \exp\{-e^{-x}\}, && \forall x \in \mathbb{R} : \mathbf{Gumbel} \end{aligned}$$

(A similar result holds for the minimum).

We can then classify the distributions according to the three possible limiting distributions of the (rescaled) maximum, introducing the notion of *Maximum Domain of Attraction* (MDA):

$$F \in \text{MDA}(H) \iff \exists (a_n) > 0, (b_n) : \forall x \in \mathbb{R}, \lim_{n \rightarrow \infty} F^n(a_n x + b_n) = H(x).$$

For instance, for Fréchet, $a_n = F^{-1}(1 - 1/n)$ and $b_n = 0$. (Note that most of the cdf F we use, usually belong to a MDA.)

To mimick the CLT, the three types of extreme value distribution have been combined into a single three-parameter family [18, 19, 36] known as **Generalized Extreme Value Distribution** (GEV).

Theorem 2 (The EV Theorem) *If $F \in MDA(G)$ then, necessarily, G is of the same type as the GEV cdf H_ξ (i.e. $G(x) = H_\xi(ax + b)$, $a > 0$), defined by*

$$H_\xi(x) = \begin{cases} \exp\left[-(1 + \xi x)_+^{-\frac{1}{\xi}}\right] & \text{if } \xi \neq 0 \\ \exp(-e^{-x}) & \text{if } \xi = 0 \end{cases}$$

where $y_+ = \max(0, y)$. The parameter ξ , named the tail (or extreme-value) index, determines the nature of the tail distribution: if $\xi > 0$ then H_ξ is Fréchet, if $\xi = 0$ then H_ξ is Gumbel, and if $\xi < 0$ then H_ξ is Weibull.

We can write $G(x) = G_{\mu,\sigma,\xi}(x) = \exp\left[-\left(1 + \xi \frac{x - \mu}{\sigma}\right)^{-\frac{1}{\xi}}\right]$,

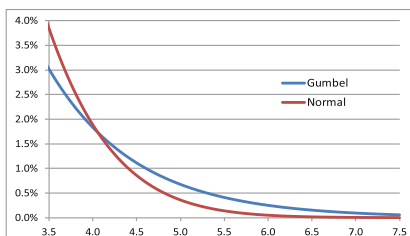
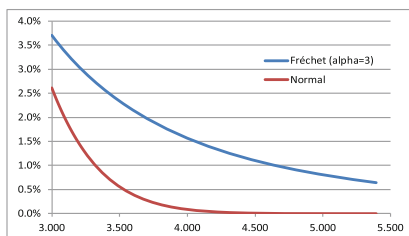
for $1 + \xi \frac{x - \mu}{\sigma} > 0$.

Moments of the GEV: the k th moment exists if $\xi < 1/k$ (in particular if $\mathbb{E}(X) < \infty$ if $\xi < 1$ and $var(X) < \infty$ if $\xi < 1/2$).

Example:

Sample cdf	MDA
Uniform	Weibull
Exponential(1) ($F(x) = 1 - e^{-x}$, $x > 0$)	Gumbel
Gaussian	Gumbel
Log-normal	Gumbel
Gamma (λ, r)	Gumbel
Cauchy ($F(x) = \frac{1}{2} + \frac{1}{\pi} \arctan x$)	Fréchet
Student	Fréchet
Pareto (β) ($F(x) = 1 - x^{-\beta}$, $x \geq 1$, $\beta > 0$)	Fréchet

Example of tails of distributions: Fréchet and Gumbel versus Gaussian (normal).



We observe that the tail can vary substantially according to the type of distributions. Here the tail of the Fréchet distribution is moderately heavy ($\alpha = \xi = 3$) although it looks much heavier than the Gaussian distribution

Exercise Considering the S&P500 daily log returns from 1987 to 2016, compute the chances to find a value smaller than the second minimum, i.e. $\mathbb{P}[X < x_{(n-1)}] = \Phi\left(\frac{x_{(n-1)} - \mu}{\sigma}\right)$ (with Φ the standard normal cdf), assuming the data are normally distributed. We obtain the following statistics on S&P500 daily log returns from 1987 to 2016:

Expected Value (μ)	0.029%	Maximum Value ($x_{(1)}$)	11.0%
Standard Deviation (σ)	1.192%	Minimum Value ($x_{(n)}$)	-22.9%
Probability of finding a value smaller than $x_{(n-1)}$ in a Gaussian Model	2.96E-16 (1 over 13.5 billion years)	Second Minimum ($x_{(n-1)}$)	-9.47%

Characteristic Property of the GEV. A distribution is a GEV if and only if it is *max-stable*, i.e. that it satisfies $\max_{1 \leq i \leq n} X_i \stackrel{d}{=} \alpha_n X + \beta_n$, with $\alpha_n > 0$.

For the three types of the GEV, we have:

Fréchet: $\max_{1 \leq i \leq n} X_i \stackrel{d}{=} n^{1/\xi} X$; Weibull: $\max_{1 \leq i \leq n} X_i \stackrel{d}{=} n^{-1/\xi} X$; Gumbel: $\max_{1 \leq i \leq n} X_i \stackrel{d}{=} X + \log n$.

2.2 A Limit Theorem for Extremes: The Pickands Theorem

Extracting more information in the tail of the distribution than just that given by the maximum should help for the evaluation of the tail. So considering the k th ($k \geq 1$) largest order statistics, we introduce the notion of ‘threshold exceedances’ where all data are extreme in the sense that they exceed a high threshold.

Picking up a high threshold $u < x_F^+$ (upper-end point of F), we study all exceedances above u .

Theorem 3 (Pickands Theorem, 1975) *If F does belong to one of the maximum domains of attraction (i.e. the limit distribution of $\max X_i$ is a GEV), then for a sufficiently high threshold u , $\exists \beta(u) > 0$ and $\xi \in \mathbb{R}$ such that the Generalized Pareto Distribution (GPD) $G_{\xi, \beta(u)}$, defined by $\overline{G}_{\xi, \beta(u)}(y) := 1 - G_{\xi, \beta(u)}(y) = \left(1 + \xi \frac{y}{\beta(u)}\right)^{-1/\xi} \mathbb{1}_{(\xi \neq 0)} + e^{-y/\beta(u)} \mathbb{1}_{(\xi = 0)}$, is a very good approximation to the excess cdf $F_u(\cdot) := \mathbb{P}[X - u \leq \cdot | X > u]$:*

$$\lim_{u \uparrow x_F^+} \sup_{0 \leq y \leq x_F^+ - u} |F_u(y) - G_{\xi, \beta(u)}(y)| = 0,$$

x_F^+ denoting the upper endpoint of F

As for the GEV, we have three cases for the GPD, depending on the sign of the tail index ξ :

- $\xi > 0$: $\overline{G}_{\xi,\beta}(y) \sim cy^{-1/\xi}$, $c > 0$ (“Pareto” tail): heavy-tail (note that $\mathbb{E}(X^k) = \infty$ for $k \geq 1/\xi$).
- $\xi < 0$: $x_G^+ = \beta/|\xi|$ (upper endpoint of G), similar to the Weibull type of the GEV (short-tailed, Pareto type II distribution)
- $\xi = 0$: $\overline{G}_{\xi,\beta}(y) = e^{-y/\beta}$: light-tail (exponential distribution with mean β)

The mean of the GPD is defined for $\xi < 1$ by $\mathbb{E}(X) = \frac{\beta}{1 - \xi}$.

2.3 Supervised Methods in EVT: Standard Thresholds Methods

Univariate Extreme Value Theory (EVT) focuses on the tail distribution evaluation, more precisely on the estimation of the tail index. That is why the first and main question is how to determine the threshold above which observations are considered as extremes. Various methods have been developed to answer this question. We give here their main ideas and refer the reader e.g. to [12] for more details (see also the references therein).

2.3.1 Peak Over Threshold (POT) Method

This method developed for the GPD by Davison and Smith [7] helps to decide on an appropriate threshold for exceedance-based methods, when looking at the empirical Mean Excess Plot (MEP). This graphical method can be qualified as supervised.

The mean excess (ME) function defined by $e(u) = \mathbb{E}[X - u | X > u]$ can be computed for any rv X (whenever its expectation exists). For instance, if X is exponentially distributed $G_{\xi,\sigma}$, then its ME function is a constant. If X is GPD $G_{\xi,\sigma}$ distributed, with $\sigma > 0$ and $\xi < 1$, then its ME function is given by

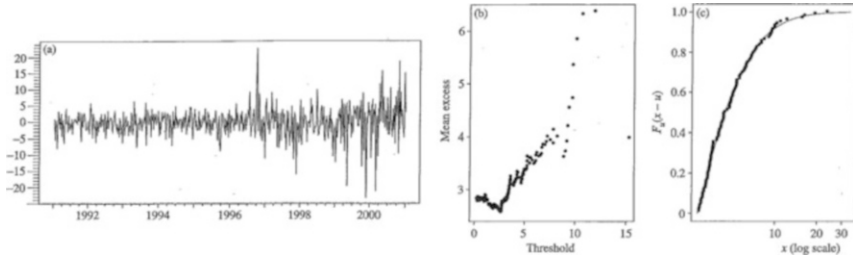
$$e(u) = \frac{\sigma + u\xi}{1 - \xi} 1_{(\sigma + u\xi > 0)}.$$

Hence, via the Pickands theorem, the MEP of X with unknown cdf F should stay reasonably close to a linear function from the threshold u at which the GPD provides a valid approximation to the excess distribution of X : $\mathbb{E}[X - u | X > u] \underset{u \rightarrow \infty}{\simeq} \sigma(u)/(1 - \xi)$. It will be the way to select u , when considering the empirical MEP

$\left(v, \frac{1}{n_v} \sum_{i=1}^{n_v} (x_{(i)} - v) : v < x_{n,n} \right)$, where the $x_{(i)}$ correspond to the n_v observations that exceed v .

Then, u being chosen, we can use ML or Moments estimators to evaluate the tail index ξ (and the scaling parameter β).

Illustration: Example from Embrechts et al.’s book [12]



Data set: time series plot (a) of AT&T weekly percentage loss data for the 521 complete weeks in the period 1991–2000

(b) Sample MEP. Selection of the threshold at a loss value of 2.75% (102 exceedances)

(c) Empirical distribution of excesses and fitted GPD, with ML estimators $\hat{\xi} = 0.22$ and $\hat{\beta} = 2.1$ (with Standard Error 0.13 and 0.34, respectively)

2.3.2 Tail Index Estimators for MDA(Fréchet) Distributions

To determine the tail index, other graphical methods than MEP may be used. Various estimators of the tail index have been (and still are) built, starting with the *Hill estimator* [17], a moment estimator [11], the QQ-estimator [22], . . . , the Hill estimator for truncated data [2], . . .

For a sample of size n , the tail index estimators are generally built on the $k = k(n)$ upper order statistics, with $k(n) \rightarrow \infty$ such that $k(n)/n \rightarrow 0$, as $n \rightarrow \infty$.

Choosing k is usually the Achilles heel of all these (graphical) supervised procedures, including the MEP one, as already observed.

Nevertheless it is remarkable to notice that for these methods, no extra information is required on the observations before the threshold (the $n - k$ th order statistics).

Let us present two tail index estimators under regular variation framework: the *Hill estimator* [17], as it is most probably still the most popular, and the *QQ-estimator* [22], which is based on a simple and intuitive idea (hence this choice).

Assume $F \in \text{MDA}(\text{Fréchet})$ with tail index $\xi > 0$, i.e. \bar{F} is regularly varying $RV_{-\alpha}$, with $\xi = \alpha^{-1}$. (Recall that a function f belongs to the class RV_ρ of regularly varying functions with index $\rho \in \mathbb{R}$ if $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ satisfies $\lim_{t \rightarrow \infty} f(tx)/f(t) = x^\rho$, for $x > 0$ (see [3].) Consider the threshold $u = X_{n-k,n}$ with $k = k(n) \rightarrow \infty$ and $k/n \rightarrow 0$ as $n \rightarrow \infty$.

- The *Hill estimator* $H_{k,n}$ of the tail index $\xi = \alpha^{-1}$ is defined by, and satisfies [17]

$$H_{k,n} := \frac{1}{k} \sum_{i=0}^{k-1} \log \left(\frac{X_{n-i,n}}{X_{n-k,n}} \right) \xrightarrow[n \rightarrow \infty]{P} \xi$$

This estimator is asymptotically normal, with a rate of convergence of $1/\alpha^2$. The Hill estimator can exhibit outrageous bias and graphical aids are often very difficult to interpret accurately. So it is wise to consider alternative methods to supplement

information given by the Hill estimator and associated plots. Thus we turn to the QQ-plot.

- The *QQ-estimator* (Kratz and Resnick [22]), $Q_{k,n}$, of the tail index ξ :

The QQ-method is based on the following simple observation: if we suspect that the n -sample X comes from the continuous cdf F , then the plot of

$$\left\{ \left(\frac{i}{n+1}, F(X_{i,n}) \right), 1 \leq i \leq n \right\}$$

should be roughly linear, hence also the QQ-plot of $\{(F^{\leftarrow}(\frac{i}{n+1}), X_{i,n}), 1 \leq i \leq n\}$ (considering the theoretical quantile $F^{\leftarrow}(\frac{i}{n+1})$ and the corresponding quantile $X_{i,n}$ of the empirical distribution function).

If $F = F_{\mu,\sigma}(x) = F_{0,1}(\frac{x-\mu}{\sigma})$, since $F_{\mu,\sigma}^{\leftarrow}(y) = \sigma F_{0,1}^{\leftarrow}(y) + \mu$, the plot of

$$\left\{ \left(G_{0,1}^{\leftarrow} \left(\frac{i}{n+1} \right), X_{i,n} \right), 1 \leq i \leq n \right\}$$

should be approximately a line of slope σ and intercept μ .

Take the example of a n -sample Pareto(α) distributed ($\bar{F}(x) = x^{-\alpha}$); then, for $y > 0$, $F_{0,\alpha}(y) := \mathbb{P}[\log X_1 > y] = e^{-\alpha y}$ and the plot of

$$\left\{ \left(F_{0,1}^{\leftarrow} \left(\frac{i}{n+1} \right), \log X_{i,n} \right), 1 \leq i \leq n \right\} = \left\{ \left(-\log \left(1 - \frac{i}{n+1} \right), \log X_{i,n} \right), 1 \leq i \leq n \right\}$$

should be approximately a line with intercept 0 and slope α^{-1} .

Now, just use the least squares estimator for the slope (SL), namely

$$SL(\{(x_i, y_i), 1 \leq i \leq n\}) = \frac{\sum_{i=1}^n x_i y_i - \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - \bar{x}^2}$$

to conclude that, for the Pareto example, an estimator of $\alpha^{-1} (= \xi)$ is

$$\widehat{\alpha^{-1}} = \frac{\sum_{i=1}^n -\log(\frac{i}{n+1}) \{n \log X_{n-i+1,n} - \sum_{j=1}^n \log X_{n-j+1,n}\}}{n \sum_{i=1}^n (-\log(\frac{i}{n+1}))^2 - (\sum_{i=1}^n -\log(\frac{i}{n+1}))^2},$$

which we call the QQ-estimator.

This method can be extended from Pareto to the general case $\bar{F} \sim RV_{-\alpha}$; we can define the QQ-estimator $Q_{k,n}$ of the tail index $\xi = \alpha^{-1}$, based on the upper k order statistics, by

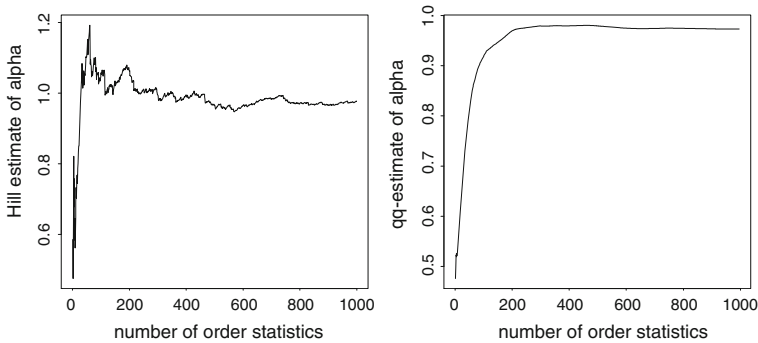
$$Q_{k,n} = SL(\{(-\log(1 - \frac{i}{k+1}), \log X_{n-k+i,n}), 1 \leq i \leq k\})$$

$$= \frac{\sum_{i=1}^k -\log\left(\frac{i}{k+1}\right) \left\{ k \log(X_{n-i+1,n}) - \sum_{j=1}^k \log(X_{n-j+1,n}) \right\}}{k \sum_{i=1}^k \left(-\log\left(\frac{i}{k+1}\right)\right)^2 - \left(\sum_{i=1}^k -\log\left(\frac{i}{k+1}\right)\right)^2}$$

and we can prove [22] that the QQ-estimator is weakly consistent ($Q_{k,n} \xrightarrow[n \rightarrow \infty]{P} \xi$) and asymptotically normal with a rate of convergence of $1/(2\alpha^2)$ (which is larger than for the Hill, but the Hill estimator exhibits considerable bias in certain circumstances). Whenever the threshold u is determined (corresponding to a k th order statistics), we can estimate the parameters, in particular the tail index.

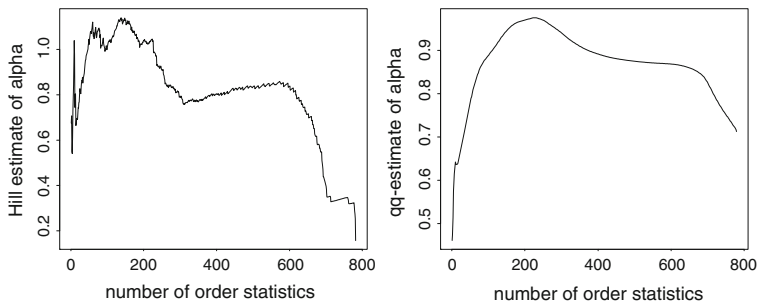
Illustration: Comparison of the Hill plot and the QQ-plot of estimates of α .

- On Pareto (1) simulated data (sample size $n = 1000$)



The QQ-plot shows $\widehat{\alpha^{-1}} \simeq 0.98$. It seems a bit less volatile than the Hill plot.

- On real data:



The Hill plot is somewhat inconclusive, whereas the QQ-plot indicates a value of about 0.97

The QQ-method in practice:

1. Make a QQ-plot of all the data (empirical vs theoretical quantile)
2. Choose k based on visual observation of the portion of the graph that looks linear
3. Compute the slope of the line through the chosen upper k order statistics and the corresponding exponential quantiles.

Alternatively, *for Hill and QQ methods:*

1. Plot $\{(k, \widehat{\alpha}^{-1}(k)), 1 \leq k \leq n\}$
2. Look for a stable region of the graph as representing the true value of α^{-1} .

Using those graphical (supervised) methods to determine u or, equivalently k , is an art as well as a science and the estimate of α is usually rather sensitive to the choice of k (but this is the price to pay for having a method to fit the tail without using any information before u).

Let us turn to another method, which answers this concern and provides an automatic (algorithmic) determination of the threshold u (but requiring, in this case, the data information before u).

2.4 A Self-Calibrated Method for Heavy-Tailed Data

It is based on a paper developed with N. Debbabi and M. Mboup [9].

We assume continuous (smooth transitions) and, with no loss of generality, right heavy tailed data (a similar treatment being possible on the left tail) belonging to the MDA(Fréchet).

Whereas one of the motivations for this new method is to be able to determine the threshold above which we fit the GPD in an unsupervised way, it will also provide a good fit for the entire distribution.

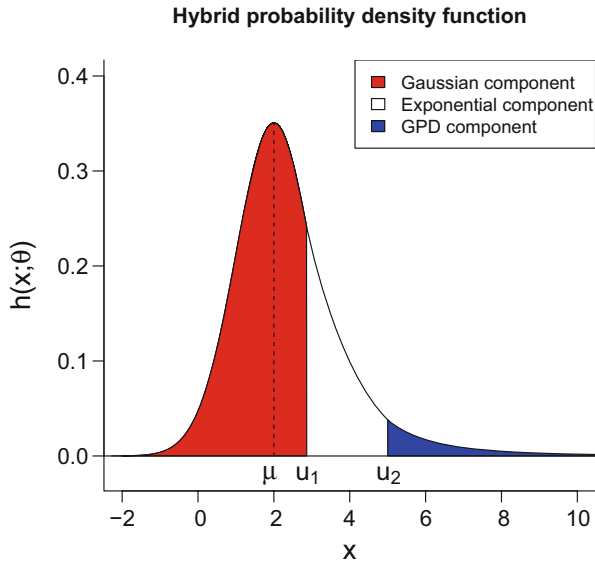
We introduce a hybrid model to fit the whole distribution underlying heavy tailed data. The idea is to consider both the mean and tail behaviors, and to use limit theorems for each one (as suggested and developed analytically in [21]), in order to make the model as general as possible. Therefore, we introduce a Gaussian distribution for the mean behavior, justified by the Central Limit Theorem (CLT), and a GPD for the tail (justified by the Pickands theorem). Then we bridge the gap between mean and asymptotic behaviors by inserting an exponential distribution used as a leverage to give full meaning of tail threshold to the junction point between the GPD and its exponential neighbour.

We assume that the hybrid model distribution (which belongs to the Fréchet MDA) has a density that is C^1 . It is the only assumption that is needed (no assumption on the dependence of the data). This model, denoted by G-E-GPD

(Gaussian-Exponential-Generalized Pareto Distribution), is characterized by its pdf h expressed as:

$$h(x; \theta) = \begin{cases} \gamma_1 f(x; \mu, \sigma), & \text{if } x \leq u_1, \\ \gamma_2 e(x; \lambda), & \text{if } u_1 \leq x \leq u_2, \\ \gamma_3 g(x - u_2; \xi, \beta), & \text{if } x \geq u_2, \end{cases}$$

where f is the Gaussian pdf (μ, σ^2), e is the exponential pdf with intensity λ , g is the GPD pdf with tail index ξ and scaling parameter β , and $\gamma_1, \gamma_2, \gamma_3$ are the weights (evaluated from the assumption).



Combining the facts that we are in the MDA(Fréchet) and that h is a C^1 pdf gives rise to six equations relating all model parameters:

$$\begin{cases} \beta = \xi u_2; & \lambda = \frac{1+\xi}{\beta}; & u_1 = \mu + \lambda \sigma^2; \\ \gamma_1 = \gamma_2 \frac{e(u_1; \lambda)}{f(u_1; \mu, \sigma)}; & \gamma_2 = \left[\xi e^{-\lambda u_2} + \left(1 + \lambda \frac{F(u_1; \mu, \sigma)}{f(u_1; \mu, \sigma)} \right) e^{-\lambda u_1} \right]^{-1}; & \gamma_3 = \beta \gamma_2 e(u_2; \lambda). \end{cases}$$

Consequently, the vector of the free parameters is reduced to $\theta = [\mu, \sigma, u_2, \xi]$.

Remark The main component in this hybrid model is the GPD one (for heavy tail), the mean behavior having to be adapted to the context. For instance, for *insurance claims*, we have replaced the Gaussian component with a *Lognormal* one (lognormal-E-GPD hybrid model).

2.4.1 Pseudo-code of the Algorithm for the G-E-GPD Parameters Estimation

Here we describe the iterative algorithm, which self-calibrates the G-E-GPD model, in particular the tail threshold above which a Fréchet distribution fits the extremes. We study its convergence, proving analytically the existence of a stationary point, then numerically that the stationary point is attractive and unique.

- 1: Initialization of $\tilde{p}^{(0)} = [\tilde{\mu}^{(0)}, \tilde{\sigma}^{(0)}, \tilde{u}_2^{(0)}]$, $\alpha, \epsilon > 0$, and k_{max} , then initialization of $\tilde{\xi}^{(0)}$ (recall that $\theta = [\mu, \sigma, u_2, \xi]$):

$$\tilde{\xi}^{(0)} \leftarrow \underset{\xi > 0}{\operatorname{argmin}} \left\| H(y; \theta \mid \tilde{p}^{(0)}) - H_n(y) \right\|_2^2,$$

where H_n is the empirical cdf of X and $\mathbf{y} = (y_j)_{1 \leq j \leq m}$ is a generated sequence of synthetic increasing data of size m (that may be different from n), with a logarithmic step, in order to increase the number of points above the tail threshold u_2 : $y_j = \min_{1 \leq i \leq n}(x_i) + (\max_{1 \leq i \leq n}(x_i) - \min_{1 \leq i \leq n}(x_i)) \log_{10} \left(1 + \frac{9(j-1)}{m-1} \right)$.

- 2: Iterative process:

- $k \leftarrow 1$

Step 1—Estimation of $\tilde{p}^{(k)}$: $\tilde{p}^{(k)} \leftarrow \underset{\substack{(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+^* \\ u_2 \in \mathbb{R}_+}}{\operatorname{argmin}} \left\| H(y; \theta \mid \tilde{\xi}^{(k-1)}) - H_n(y) \right\|_2^2$

Step 2—Estimation of $\tilde{\xi}^{(k)}$: $\tilde{\xi}^{(k)} \leftarrow \underset{\xi > 0}{\operatorname{argmin}} \left\| H(y; \theta \mid \tilde{p}^{(k)}) - H_n(y) \right\|_2^2$.

- $k \leftarrow k + 1$

until $(d(H(y; \theta^{(k)}), H_n(y)) < \epsilon$ and $d(H(y_{q_\alpha}; \theta^{(k)}), H_n(y_{q_\alpha})) < \epsilon)$ or $(k = k_{max})$

where ϵ is a positive real that is small enough, y_{q_α} represents the observations above a fixed high quantile q_α of arbitrary order $\alpha \geq 80\%$ associated with H and $d(a, b)$ denotes the distance between a and b , chosen in this study as the Mean Squared Error (MSE); it can be interpreted as the Cramér-von-Mises test of goodness of fit.

- 3: Return $\theta^{(k)} = [\tilde{\mu}^{(k)}, \tilde{\sigma}^{(k)}, \tilde{u}_2^{(k)}, \tilde{\xi}^{(k)}]$.

2.4.2 Performance of the Method (Algorithm) Tested via MC Simulations

To study the performance of the algorithm to self-calibrate the G-E-GPD model, we build on MC simulations. To do so, we proceed in four steps:

1. Consider
 - $\{X^q = (X_p^q)_{1 \leq p \leq n}\}_{1 \leq q \leq N}$: training sets of length n
 - $\{Y^q = (Y_p^q)_{1 \leq p \leq l}\}_{1 \leq q \leq N}$: test sets of length l
 - with a G-E-GPD parent distribution with a fixed parameters vector θ .
2. On each training set X^q , $1 \leq q \leq N$, evaluate $\tilde{\theta}^q = [\tilde{\mu}^q, \tilde{\sigma}^q, \tilde{u}_2^q, \tilde{\xi}^q]$ using our algorithm
3. Compute the empirical mean \tilde{a} and variance \tilde{S}_a of estimates of each parameter a over the N training sets. To evaluate the performance of the estimator \tilde{a} , we use two criteria:
 - (i) MSE expressed for any a as: $MSE_a = \frac{1}{N} \sum_{q=1}^N (\tilde{a}^q - a)^2$; a small value of MSE highlights the reliability of parameters estimation using the algorithm.
 - (ii) Test on the mean (with unknown variance): $\left| \begin{array}{l} H0 : \tilde{a} = a \\ H1 : \tilde{a} \neq a \end{array} \right.$
 (use for instance the normal test for a large sample)
4. Compare the hybrid pdf h (with the fixed θ) with the corresponding estimated one \tilde{h} , using $\tilde{\theta}^q$ on each test set Y^q . To do so, compute the average of the log-likelihood function \mathcal{D} , over N simulations, between $h(Y^q; \tilde{\theta}^q)$ and $h(Y^q; \theta)$: $\mathcal{D} = \frac{1}{Nl} \sum_{q=1}^N \sum_{p=1}^l \log(h(Y_p^q; \theta) / \tilde{h}(Y_p^q; \tilde{\theta}^q))$. The smallest the value of \mathcal{D} is, the most trustworthy is the algorithm.

Several MC simulations have been performed varying θ and n , to test the robustness of the algorithm (see [9, §4 and Appendix B]).

2.4.3 Application in Neuroscience: Neural Data

We consider the data corresponding to 20 s, equivalent to $n = 3.10^5$ observations, of real extracellular recording of neurons activities. The information to be extracted from these data (spikes or action potentials) lies on the extreme behaviors (left and right) of the data (Fig. 4).

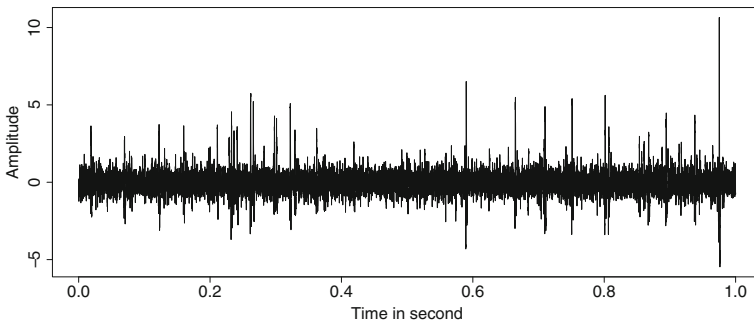


Fig. 4 One second of neural data, extracellularly recorded

Table 1 Comparison between the self-calibrating method and the three graphical methods: MEP, Hill and QQ ones

Model	Tail index (ξ)	Threshold (u_2)	N_{u_2}	Distance (tail distr.)	Distance (full distr.)
GPD	MEP (PWM): 0.3326	1.0855 = $q_{93.64\%}$	19,260	3.26×10^{-6}	
GPD	Hill-estimator: 0.599	1.0855 = $q_{93.64\%}$	19,260	2.07×10^{-6}	
GPD	QQ-estimator: 0.5104	1.0671 = $q_{93.47\%}$	19,871	1.26×10^{-5}	
G-E-GPD	Self-calibrating method: 0.5398	1.0301 = $q_{92.9\%}$	21,272	7.79×10^{-6}	9.31×10^{-5}

N_{u_2} represents the number of observations above u_2 . The distance gives the MSE between the empirical (tail or full respectively) distribution and the estimated one from a given model (GPD or hybrid G-E-GPD respectively). The neural data sample size is $n = 3 \times 10^5$

Since the neural data can be considered as symmetric, it is sufficient to evaluate the right side of the distribution with respect to its mode.

In Table 1, we present the results obtained with the self-calibrating method, the MEP, Hill and QQ methods. Since the three graphical approaches fit only the tail distribution, the comparison of the methods will focus on the goodness-of-fit of the GPD component. As observed in this table, the MSE between the estimated cdf and the empirical one, using only data above the selected threshold, is small enough for the four methods ensuring a reliable modeling of extremes. The GPD threshold and the estimated tail index are of the same order of magnitude for all methods; it confirms that our algorithm works in the right direction.

We can also notice the good performance of these methods through Fig. 5, where we plot the empirical quantile function and the estimated ones using the self-calibrating method and the various graphical ones. However, the advantage of our

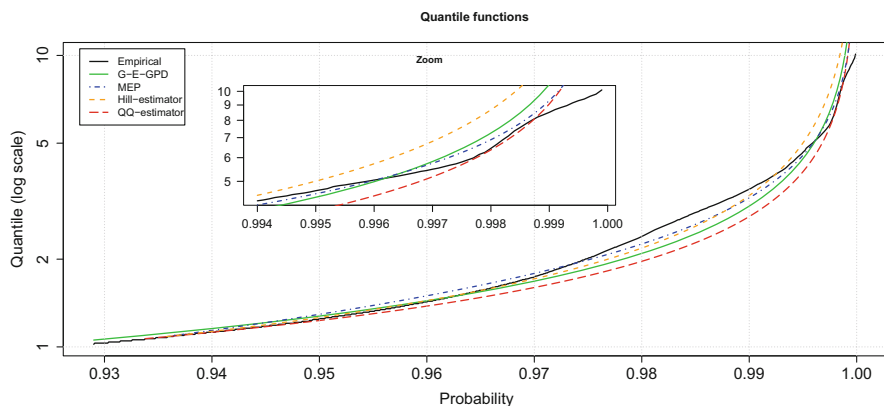


Fig. 5 Neural data: comparison between the empirical quantile function and the estimated ones via the self-calibrating method and the graphical methods

method is that it is unsupervised, i.e. it does not need the intervention of the user to select the threshold manually. Moreover it provides a good fit between the hybrid cdf estimated on the entire data sample (the right side for this data set) and the empirical cdf, with a MSE of order 10^{-5} .

2.4.4 Application in Finance: S&P 500 Data

Consider the S&P500 log-returns from January 2, 1987 to February 29, 2016, corresponding to $n = 7348$ observations, available in the `tseries` package of the R programming language. It is well known that log-returns of financial stock indices exhibit left and right heavy tails, with a slight different tail index from one to the other. It is important in such context to evaluate the nature of tail(s) in order to compute the capital needed by a financial institution to cover their risk, often expressed as a Value-at-Risk (i.e. a quantile) of high order.

The S&P500 log-returns being essentially symmetric around zero (representing the data mode), we kept the Gaussian component to model the mean behavior when applying the self-calibrating method. We modelled the negative log returns and the positive ones, respectively, then the full data set. When focusing on tails, we also compare our results with those obtained with MEP, Hill, and QQ methods. We present them in Tables 2 and 3. We observe that all methods offer a good fit of the tail distribution. However, the advantage of the self-calibrating method is that it does not need the intervention of the user to select the threshold manually, which is a considerable advantage in practice.

Now, to underline the good performance of the self-calibrating method even in the case when data are autocorrelated with a long memory, we apply it on the S&P500 absolute log-returns. Indeed, it is well known that the absolute value of financial returns are autocorrelated (see Fig. 6), but also that their extremes are not (for a thorough discussion of this point and empirical evidences, see (author?) [16]). In time of crisis, as e.g. in 2008–2009, we observe an increase of the dependence between various financial indices, in particular in the extremes. This is to be distinguished from a dependence of the extremes within a univariate financial index, which is not observed [16].

A comparison of the results obtained with our self-calibrating method and the graphical EVT ones is depicted in Table 4. In Fig. 7, we also give a comparison of the estimated quantile function using the G-E-GPD method and the graphical (MEP, Hill and QQ) ones. Through Table 4 and Fig. 7, we can highlight once again the good performance of the self-calibrating method to estimate the tail distribution as well as the entire distribution of autocorrelated data. Note that the estimate of the tail index is of the same order as those of the upper and lower tail indices evaluated in the previous section.

Table 2 Lower tail modeling of the S&P500 log-returns

Model	Tail index (ξ)	Threshold (u_2)	N_{u_2}	Distance (tail distr.)	Distance (positive distr.)
GPD	MEP: 0.3640	$0.0270 = q_{98.36\%}$	120	1.19×10^{-7}	
GPD	Hill-estimator: 0.3601	$0.0301 = q_{98.84\%}$	86	6.43×10^{-8}	
GPD	QQ-estimator: 0.3813	$0.0313 = q_{99.00\%}$	74	3.54×10^{-8}	
G-E-GPD	Self-calibrating method: 0.3545	$0.0289 = q_{98.63\%}$	100	2.64×10^{-7}	3.11×10^{-6}

Comparison between the self-calibrating method and the three graphical methods: MEP, Hill and QQ ones, applied on the right side of the S&P500 opposite log-returns ($-X$). N_{u_2} represents the number of observations above the tail threshold u_2 . The distance gives the MSE between the empirical tail (from u_2), or positive side (for $x \geq 0$) respectively, distribution and the estimated one from a given model (GPD, or hybrid G-E-GPD respectively)

Table 3 Upper tail modeling of the S&P500 log-returns

Model	Tail index (ξ)	Threshold (u_2)	N_{u_2}	Distance (tail distr.)	Distance (positive distr.)
GPD	MEP: 0.2715	$0.0209 = q_{96.89\%}$	229	4.91×10^{-7}	
GPD	Hill-estimator: 0.3225	$0.0288 = q_{98.84\%}$	86	4.42×10^{-7}	
GPD	QQ-estimator: 0.2859	$0.0321 = q_{99.03\%}$	71	5.06×10^{-8}	
G-E-GPD	Self-calibrating method: 0.3360	$0.0266 = q_{98.51\%}$	109	3.89×10^{-7}	2.49×10^{-6}

Comparison between the self-calibrating method and the graphical methods applied on the right side of the S&P500 log-returns

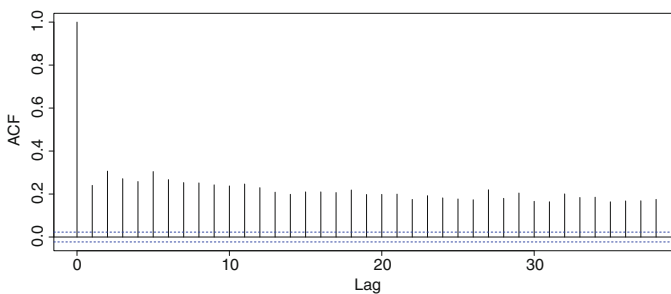


Fig. 6 AutoCorrelation function (ACF) of the S&P500 absolute log-returns

Table 4 Comparison between the self-calibrating method and the three graphical methods: MEP, Hill and QQ ones

Model	Tail index (ξ)	Threshold (u_2)	N_{u_2}	Distance (tail distr.)	Distance (full distr.)
GPD	MEP: 0.3025	0.0282 = $q_{97.21\%}$	206	1.78×10^{-7}	
GPD	Hill-estimator: 0.3094	0.0382 = $q_{98.85\%}$	85	4.49×10^{-8}	
GPD	QQ-estimator: 0.3288	0.0323 = $q_{98.14\%}$	137	6.01×10^{-8}	
G-E-GPD	Self-calibrating method: 0.3331	0.0290 = $q_{97.49\%}$	184	2.00×10^{-7}	1.05×10^{-5}

The S&P500 absolute log-returns data sample size is $n = 7348$

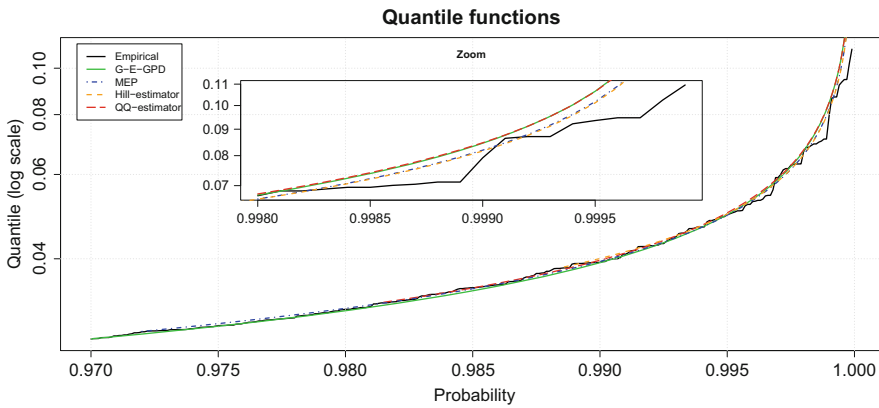


Fig. 7 S&P 500 absolute log-returns data: comparison between the empirical quantile function and the estimated ones via the self-calibrating method and the graphical methods

3 Dependence

3.1 Motivation

3.1.1 Impact of the Dependence on the Diversification Benefit

The diversification performance is at the heart of the strategy of a company. It can be measured, for a portfolio of n risks, via the diversification benefit $D_{n,\alpha}$ at a threshold α ($0 < \alpha < 1$) defined by $D_{n,\alpha} = 1 - \frac{\rho_\alpha(\sum_{i=1}^n L_i)}{\sum_{i=1}^n \rho_\alpha(L_i)}$, where ρ denotes a risk measure. This indicator, not universal as it depends on the number of the risks undertaken and on the chosen risk measure ρ , helps to determine the optimal portfolio of the company since the diversification reduces the risk and thus enhances the performance. This is key to both insurances and financial institutions.

Before developing an example in the insurance context (it would be the same for investment banks) to point out the impact of the dependence on the diversification benefit, let us recall some standard notions in insurance.

Insurance Framework

In insurance, the risk is priced based on the knowledge of the loss probability distribution. The occurrence of a loss L being random, we define it as a random variable (rv) on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ (note that, in insurance context, we often use *risk* and *loss* for one another).

The role of capital for an insurance company is to ensure that the company can pay its liability even in the worst cases, up to some threshold.

It means to define the capital to put behind the risk. That is why we introduce a risk measure (say ρ), defined on the loss distribution, in order to estimate the capital needed to ensure payment of the claim up to a certain confidence level.

Now let us define some useful quantities, as:

Risk-adjusted-capital. The risk can be defined as the deviation from the expectation, hence the notion of Risk-Adjusted-Capital (RAC), say K , which is a function of the risk measure ρ associated to the risk L , defined by $K = \rho(L) - \mathbb{E}[L]$.

Risk Loading. An insurance is a company in which shareholders can invest. They expect a return on investment. So the insurance firm has to make sure that the investors receive their dividends. It corresponds to the cost of capital, η , that the insurance company must charge on its premium. Consider a portfolio of N similar policies. The risk loading per policy, say R , is defined as $R = \eta \frac{K_N}{N} = \eta \left(\frac{\rho(L^{(N)})}{N} - \mathbb{E}[L_1] \right)$, where K_N is the capital assigned to the entire portfolio, $L^{(N)} = \sum_{i=1}^N L_i$ is the total loss of the portfolio, and $L_1 = L$ is the loss incurred by one policy (of the portfolio).

Technical risk premium. For one policy case, incurring a loss L , the technical premium, P , that needs to be paid can be defined by $\mathbb{P} = \mathbb{E}(L) + \eta K + e$, where η is the return expected by shareholders before tax, K is the RAC (i.e. the capital assigned to this risk), ηK corresponds to the Risk loading (per policy), and e are the expenses incurred by the insurer to handle this case.

Assuming that the expenses are a small portion of the expected loss, i.e. $e = a\mathbb{E}[L]$ with $0 < a \ll 1$, then the premium can be written as $P = (1 + a)\mathbb{E}[L] + R$.

Generalizing to a portfolio of N similar (iid) policies, the total loss is $L^{(N)} = \sum_{i=1}^N L_i$, hence the premium for one policy in the portfolio becomes:

$$P = \frac{(1+a)\mathbb{E}[L^{(N)}] + \eta K_N}{N} = (1+a)\mathbb{E}[L] + \eta \frac{K_N}{N},$$

where $\eta \frac{K_N}{N}$ is the risk loading per policy.

Let us then develop our toy model (see [4, 5]) to show the dependence impact on the diversification benefit.

Suppose an insurance company has underwritten N policies of a given risk. To price these policies, the company must know the underlying probability distribution of this risk. Assume that each policy is exposed n times to this risk, thus in a portfolio of N policies, the risk may occur $n \times N$ times.

Let us introduce a sequence $(X_i, i = 1, \dots, Nn)$ of rv's X_i to model the occurrence of the risk, with a given severity l (for simplicity, take it deterministic). Hence the total loss amount, say L , associated to this portfolio is given by $L = l \sum_{i=1}^{Nn} X_i := l S_{Nn}$.

We are going to consider three models for the occurrence of the risk, depending on the dependence structure.

(a) A First Simple Model, Under the iid Assumption

Assume the X_i 's are iid (independent, identically distributed) with parent rv denoted by X , Bernoulli distributed $\mathcal{B}(p)$, i.e. the loss $L_1 = lX$ occurs with some probability p :

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p \end{cases}$$

Hence the total loss amount $L = l S_{Nn}$ follows a binomial distribution $\mathcal{B}(Nn, p)$. We can then deduce the risk loading for an increasing number N of policies in the portfolio: $R = \eta \left(\frac{\rho(L)}{N} - \ln p \right)$, in order to determinate the risk premium the insurance will ask to a customer if he buys this insurance policy. The relative risk loading per policy is then $\frac{R}{\mathbb{E}[L^{(1)}]} = \eta \left(\frac{\rho(L)}{\ln p} - 1 \right)$.

Numerical application. We choose for instance the number of times one policy is exposed to the risk as $n = 6$ and the unit loss l is fixed to $l = 10$ Euros.

Computing the loss distribution, we obtain the results presented in Table 5. We observe that the probability that the company will turn out paying more than the

Table 5 Distribution of the loss $L = lS_{1n}$ for one policy ($N = 1$) with $n = 6$ and $p = 1/6$

Number of losses k	Policy loss $lX(\omega)$	Probability mass $\mathbb{P}[S_{1n} = k]$	Cdf $\mathbb{P}[S_{1n} \leq k]$
0	0	33.490%	33.490%
1	10	40.188%	73.678%
2	20	20.094%	93.771%
3	30	5.358%	99.130%
4	40	0.804%	99.934%
5	50	0.064%	99.998%
6	60	0.002%	100.000%

Table 6 The Risk loading per policy as a function of the number N of policies in the portfolio (with $n = 6$)

Risk measure ρ	Number N of policies	Risk loading R per policy with probability		
		$p = 1/6$	$p = 1/4$	$p = 1/2$
VaR	1	3.000	3.750	4.500
	5	1.500	1.650	1.800
	10	1.050	1.200	1.350
	50	0.450	0.540	0.600
	100	0.330	0.375	0.420
	1000	0.102	0.117	0.135
	10,000	0.032	0.037	0.043
	TVaR	1	3.226	3.945
5		1.644	1.817	1.963
10		1.164	1.330	1.482
50		0.510	0.707	0.675
100		0.372	0.425	0.476
1000		0.116	0.134	0.154
10,000		0.037	0.042	0.049
$\mathbb{E}[L]/N$		10.00	15.00	30.00

expectation $\mathbb{E}(L) = lnp = 10$, is of more than 26%. It makes then clear why the technical premium cannot be reduced to $\mathbb{E}(L)$.

Now we compute the risk loading per policy as a function of the number N of policies in the portfolio for both risk measures VaR and TVaR, and when taking $p = 1/6$ (fair game), $1/4$ and $1/2$, respectively. We assume that the cost (of capital) is $\eta = 15\%$, and that the risk measure is computed at threshold $\alpha = 99\%$. Results are given in Table 6.

We observe that, in the case of independent risks, the risk loading R is a decreasing function of the number of policies N , even with a biased dice. With 10,000 policies, R is divided by 100, whatever is the choice of the risk measure ρ , with slightly higher values for TVaR than VaR.

(b) Introducing a Structure of Dependence to Reveal a Systematic Risk

We introduce two types of structure of dependence between the risks, in order to explore the occurrence of a systematic risk and, as a consequence, the limits to diversification.

We still consider the sequence $(X_i, i = 1, \dots, Nn)$ to model the occurrence of the risk, with a given severity l , for N policies, but do not assume anymore that the X_i 's are independent (but identically distributed, for sake of simplicity). We assume that the occurrence of the risks X_i 's depends on another phenomenon, represented by a random variable (rv), say U . Depending on the intensity of the phenomenon, i.e. the values taken by U , a risk X_i has more or less chances to occur.

Suppose that the dependence between the risks is totally captured by U that is identified to the occurrence of a state of systematic risk. Consider, w.l.o.g., that U can take two possible values denoted by 1 and 0: $U \stackrel{d}{\sim} \mathcal{B}(\tilde{p}), 0 < \tilde{p} \ll 1$, where \tilde{p} is chosen very small since we want to explore rare events. We present two examples of models (i.e. two types of dependence).

(i) A dependent model, but conditionally independent

The occurrence of the risks $(X_i)_i$ is modeled by a Bernoulli rv whose parameter is chosen depending on U and such that the conditional rv's $X_i | U$ are independent. Since U takes two possible values, the same holds for the parameter of the Bernoulli distribution of the conditionally independent rv's $X_i | U$, namely

$$X_i | (U = 1) \stackrel{d}{\sim} \mathcal{B}(q) \quad \text{and} \quad X_i | (U = 0) \stackrel{d}{\sim} \mathcal{B}(p)$$

where we choose $q \gg p$, so that whenever U occurs (i.e. $U = 1$ (crisis state)), it has a big impact in the sense that there is a higher chance of loss. We include this effect in order to have a systematic risk (non-diversifiable) in our portfolio. Hence the mass probability distribution f_S of the total amount of losses S_{Nn} appears as a mixture of two mass probability distributions $f_{\tilde{S}_q}$ and $f_{\tilde{S}_p}$ of conditional independent rv's $\tilde{S}_q := S_{Nn} | (U = 1) \stackrel{d}{\sim} \mathcal{B}(Nn, q)$ and $\tilde{S}_p := S_{Nn} | (U = 0) \stackrel{d}{\sim} \mathcal{B}(Nn, p)$, respectively:

$$f_S = \tilde{p} f_{\tilde{S}_q} + (1 - \tilde{p}) f_{\tilde{S}_p}.$$

Note that $\tilde{p} = 0$ gives back the normal state.

Numerical application. As for example (a), we take $n = 6$ and $p = 1/n$. Moreover we choose the loss probability during the crisis to be $q = 1/2$, and explore different probabilities \tilde{p} of occurrence of a crisis. In Table 7, the results illustrate well the effect of the non-diversifiable risk. When the probability of occurrence of a crisis is high, the diversification does not play a significant role anymore already with 100 contracts in the portfolio. For $\tilde{p} \geq 1\%$, the risk loading barely changes when there is a large number of policies (starting at

Table 7 The risk loading per policy as a function of the probability of occurrence of a systematic risk in the portfolio using VaR and TVaR measures with $\alpha = 99\%$

Risk measure ρ	Number N of policies	Risk loading R				
		In a normal state	With occurrence of a crisis state			
		$\tilde{p} = 0$	$\tilde{p} = 0.1\%$	$\tilde{p} = 1.0\%$	$\tilde{p} = 5.0\%$	$\tilde{p} = 10.0\%$
Var	1	3.000	2.997	4.469	4.346	5.693
	5	1.500	1.497	2.070	3.450	3.900
	10	1.050	1.047	1.770	3.300	3.450
	50	0.450	0.477	1.410	3.060	3.030
	100	0.330	0.327	1.605	3.000	2.940
	1000	0.102	0.101	2.549	2.900	2.775
	10,000	0.032	0.029	2.837	2.866	2.724
TVaR	1	3.226	3.232	4.711	4.755	5.899
	5	1.644	1.707	2.956	3.823	4.146
	10	1.164	1.266	2.973	3.578	3.665
	50	0.510	0.760	2.970	3.196	3.141
	100	0.372	0.596	2.970	3.098	3.020
	1000	0.116	0.396	2.970	2.931	2.802
	10,000	0.037	0.323	2.970	2.876	2.732
E[L]/N		10.00	10.02	10.20	11.00	12.00

The probability of giving a loss in a state of systematic risk is chosen to be $q = 50\%$

$N = 1000$) in the portfolio, for both VaR and TVaR. The non-diversifiable term dominates the risk. For lower probability \tilde{p} of occurrence of a crisis, the choice of the risk measure matters. For instance, when choosing $\tilde{p} = 0.1\%$, the risk loading, compared to the normal state, is multiplied by 10 in the case of TVaR, for $N = 10,000$ policies, and hardly moves in the case of VaR! This effect remains, but to a lower extend, when diminishing the number of policies. It is clear that the VaR measure does not capture well the crisis state, while TVaR is sensitive to the change of state, even with such a small probability and a high number of policies.

(ii) *A more realistic model setting to introduce a systematic risk*

We adapt further the previous setting to a more realistic description of a crisis. At each of the n exposures to the risk, in a state of systematic risk, the entire portfolio will be touched by the same increased probability of loss, whereas, in a normal state, the entire portfolio will be subject to the same equilibrium probability of loss.

For this modeling, it is more convenient to rewrite the sequence $(X_i, i = 1, \dots, Nn)$ with a vectorial notation, namely $(\mathbf{X}_j, j = 1, \dots, n)$ where the vector \mathbf{X}_j is defined by $\mathbf{X}_j = (X_{1j}, \dots, X_{Nj})^T$. Hence the total loss amount S_{Nn} can be rewritten as

$$S_{Nn} = \sum_{j=1}^n \tilde{S}^{(j)} \quad \text{where} \quad \tilde{S}^{(j)} \text{ is the sum of the components of } \mathbf{X}_j : \tilde{S}^{(j)} = \sum_{i=1}^N X_{ij}.$$

We keep the same notation for the Bernoulli rv U determining the state and for its parameter \tilde{p} . But now, instead of defining a normal ($U = 0$) or a crisis ($U = 1$) state on each element of $(X_i, i = 1, \dots, Nn)$, we do it on each vector $\mathbf{X}_j, 1 \leq j \leq n$.

It comes back to define a sequence of iid rv's $(U_j, j = 1, \dots, n)$ with parent rv U . We deduce that $\tilde{S}^{(j)}$ follows a Binomial distribution whose probability depends on U_j :

$$\tilde{S}^{(j)} \mid (U_j = 1) \stackrel{d}{\sim} \mathcal{B}(N, q) \quad \text{and} \quad \tilde{S}^{(j)} \mid (U_j = 0) \stackrel{d}{\sim} \mathcal{B}(N, p),$$

and these conditional rv's are independent.

Let us introduce the event A_l defined, for $l = 0, \dots, n$, as

$A_l := \{l \text{ vectors } \mathbf{X}_j \text{ are exposed to a crisis state and } n - l \text{ to a normal state}\}$

$$= \left(\sum_{j=1}^n U_j = l \right)$$

whose probability is given by

$$\mathbb{P}(A_l) = \mathbb{P}\left(\sum_{j=1}^n U_j = l\right) = \binom{n}{l} \tilde{p}^l (1 - \tilde{p})^{n-l}.$$

We can then write, with, by conditional independence,

$$\tilde{S}_q^{(l)} = \sum_{j=1}^l \left(\tilde{S}^{(j)} \mid U_j = 1 \right) \stackrel{d}{\sim} \mathcal{B}(Nl, q)$$

and

$$\tilde{S}_p^{(n-l)} = \sum_{j=1}^{n-l} \left(\tilde{S}^{(j)} \mid U_j = 0 \right) \stackrel{d}{\sim} \mathcal{B}(N(n-l), p),$$

that

$$\mathbb{P}(S_{Nn} = k) = \sum_{l=0}^n \mathbb{P}(S_{Nn} = k \mid A_l) \mathbb{P}(A_l) = \sum_{l=0}^n \binom{n}{l} \tilde{p}^l (1 - \tilde{p})^{n-l} \mathbb{P}[\tilde{S}_q^{(l)} + \tilde{S}_p^{(n-l)} = k].$$

Numerical example revisited: In this case, we cannot directly use an explicit expression for the distributions, so we go through Monte-Carlo simulations.

At each of the n exposures to the risk, first choose between a normal or a crisis state. Since, we take here $n = 6$, the chances of choosing a crisis state when

Table 8 The risk loading per policy as a function of the probability of occurrence of a systematic risk in the portfolio using VaR and TVaR measures with $\alpha = 99\%$

Risk measure ρ	Number N of policies	Risk loading R				
		in a normal state	with occurrence of a crisis state			
		$\tilde{p} = 0$	$\tilde{p} = 0.1\%$	$\tilde{p} = 1.0\%$	$\tilde{p} = 5.0\%$	$\tilde{p} = 10.0\%$
VaR	1	3.000	2.997	2.969	4.350	4.200
	5	1.500	1.497	1.470	1.650	1.800
	10	1.050	1.047	1.170	1.350	1.500
	50	0.450	0.477	0.690	0.990	1.200
	100	0.330	0.357	0.615	0.945	1.170
	1000	0.102	0.112	0.517	0.882	1.186
	10,000	0.032	0.033	0.485	0.860	1.196
	100,000	0.010	0.008	0.475	0.853	1.199
TVaR	1	3.226	3.232	4.485	4.515	4.448
	5	1.644	1.792	1.870	2.056	2.226
	10	1.164	1.252	1.342	1.604	1.804
	50	0.510	0.588	0.824	1.183	1.408
	100	0.375	0.473	0.740	1.118	1.358
	1000	0.116	0.348	0.605	1.013	1.295
	10,000	0.037	0.313	0.563	0.981	1.276
	100,000	0.012	0.301	0.550	0.970	1.269
$\mathbb{E}[L]/N$		10.00	10.02	10.20	11.00	12.00

The probability of giving a loss in a state of systematic risk is chosen to be $q = 50\%$

$\tilde{p} = 0.1\%$ is very small. To get enough of the crisis states, we need to do enough simulations, and then average over all the simulations. The results shown in Table 8 are obtained with ten million simulations (we ran it also with 1 and 20 million simulations to check the convergence).

The diversification due to the total number of policies is more effective for this model than for the previous one, but we still experience a part which is not diversifiable. We also computed the case with 100,000 policies (since via Monte Carlo simulations). As expected, the risk loading in the normal state continues to decrease. In this state, it decreases by $\sqrt{10}$. However, except for $\tilde{p} = 0.1\%$ in the VaR case, the decrease becomes very slow when we allow for a crisis state to occur. The behavior of this model is more complex than the previous one, but more realistic, and we reach also the non-diversifiable part of the risk. For a high probability of occurrence of a crisis (1 every 10 years), the limit with VaR is reached already at 100 policies, while, with TVaR, it continues to slowly decrease. Concerning the choice of risk measure, we see a similar behavior as in the previous case for the case $N = 10,000$ and $\tilde{p} = 0.1\%$: VaR is unable to catch the possible occurrence of a crisis state, which shows its limitation as a risk measure. Although we know that there is a part of the risk that is non-diversifiable, VaR does not catch it really when $N = 10,000$ or 100,000 while TVaR does not decrease significantly

Table 9 Summary of the analytical results (expectation and variance per policy) for the three cases of biased games ($L = l S_{Nn}$)

Case	Expectation $\frac{1}{N} \mathbb{E}(L)$	Variance $\frac{1}{N^2} var(L)$
(a)	$\ln q$	$\frac{l^2 n}{N} q(1 - q)$
(b)-(i)	$\ln(\tilde{p} q + (1 - \tilde{p}) p)$	$\frac{l^2 n}{N} (q(1 - q)\tilde{p} + p(1 - p)(1 - \tilde{p})) + l^2 n^2 (q - p)^2 \tilde{p}(1 - \tilde{p})$
(b)-(ii)	$\ln(\tilde{p} q + (1 - \tilde{p}) p)$	$\frac{l^2 n}{N} (q(1 - q)\tilde{p} + p(1 - p)(1 - \tilde{p})) + l^2 n (q - p)^2 \tilde{p}(1 - \tilde{p})$

between 10,000 and 100,000 reflecting the fact that the risk cannot be completely diversified away.

Discussion: Comparison of the Methods

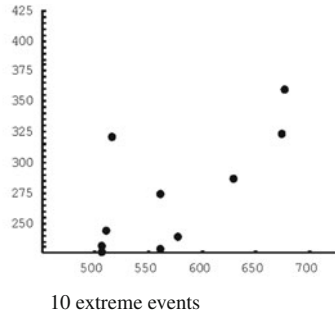
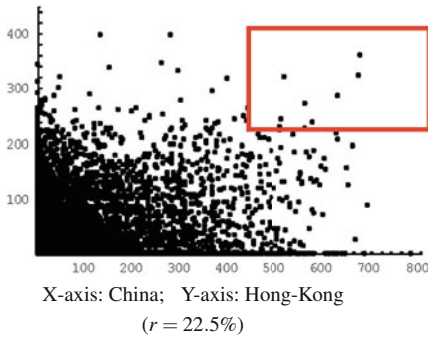
In Table 9, we see in the first case (a) that the variance decreases with increasing N , while both other cases (b) (i and ii) contain a term in the variance that does not depend on N . Those two cases are those containing a systematic risk component that cannot be diversified. Note that the variance $var_2(L)$ of L in the case (b)-(i) contains a non-diversified part that corresponds to n times the non-diversified part of $var_3(L)$ in the case (b)-(ii).

To conclude, we have seen the effect of diversification on the pricing of insurance risk through a simple modeling that allows for a straightforward analytical evaluation of the impact of the non-diversified part. In real life, risk takers have to pay special attention to the effects that can weaken the diversification benefits, hence affect greatly the risk loading of the risk premium (as seen here). Various examples can illustrate this situation, as, for instance, for motor insurance, the appearance of a hail storm may hit a big number of cars at the same time and thus cannot be diversified among the various policies, or for life insurance, pandemic or mortality trend would affect the entire portfolio and cannot be diversified away, or the financial crisis suddenly increases the dependence between risks (systemic risk). There is a saying among traders: “Diversification works the best when you need it the least”.

Understanding the dependence between risks is crucial for solid risk management. For portfolio management, we need to include both the single risk model and the dependence model.

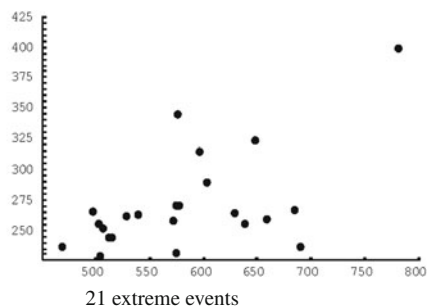
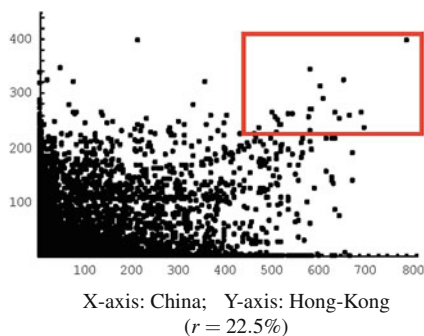
3.1.2 Type of Dependence

Consider a portfolio of political risks, the two largest ones being those of China and Hong-Kong, with 22.5% linear correlation. A customer asks a reinsurer for a cover of those extreme risks, providing him their marginal distributions and the following simulations results:



Applying the reinsurance structure (rectangle) to the customer’s simulations, we find ten relevant events in it. However, in this model, the conditional probability for Hong-Kong to default on the risk, given that China defaults with probability of 1/200 years (i.e. 0.5%), would give a probability less than 5% , which is totally unrealistic, given the political situation of dependence of Hong-Kong on China!

Hence the reinsurer decides to study this portfolio, using the same margins, but suggesting a dependence structure via a Clayton copula, calibrating it to have the same linear correlation of 22.5%. He obtains a much more realistic conditional probability of default of 60%, which gives 21 relevant events in the reinsurance structure. Applying simply a non-linear dependence structure increases by a factor 2 the number of events and by a factor 3 the average loss for the reinsurer. Of course, the price of such a cover would be much higher than what the customer expected given his model.



This example shows that the type of dependence considered for the modeling matters a lot when considering the risk!

3.2 Notion of Dependence

How to analyze a phenomenon in view of understanding it better, then modeling it? Modeling is a simplification but must not be a reduction! It is the fundamental basis of a scientific approach. ‘Everything should be made as simple as possible, but not simpler’ (Saying attributed to Albert Einstein).

We proceed from simplest tools to more elaborated ones, when needed. In terms of dependence, in a multivariate context, it means to look at the rv’s from independence to linear dependence to non-linear dependence. Studying the dependence between risks is essential for understanding their real impacts and consequences. There exists many ways of describing dependence or association between rv’s, e.g. linear correlation coefficient, rank correlations (Kendall’s tau, Spearman’s rho), . . .

Let us present a brief historical overview. Dependence has always been a topic in probability and statistics when looking at what is called a multivariate framework. Notions like linear correlation or copula, for instance, were introduced to treat this problem.

- In 1895: Karl Pearson[30] formalized mathematically the notion of linear correlation (first introduced by Galton in the context of biometric studies). If independence implies linear independence, the converse is false (except in the elliptical case), as illustrated on Fig. 8.
- In 1959: Abe Sklar [35] introduced (in the context of probability theory to solve a theoretical problem posed by Fréchet) the more general concept of dependence structure, called also *copula*, separating this structure from the margins.

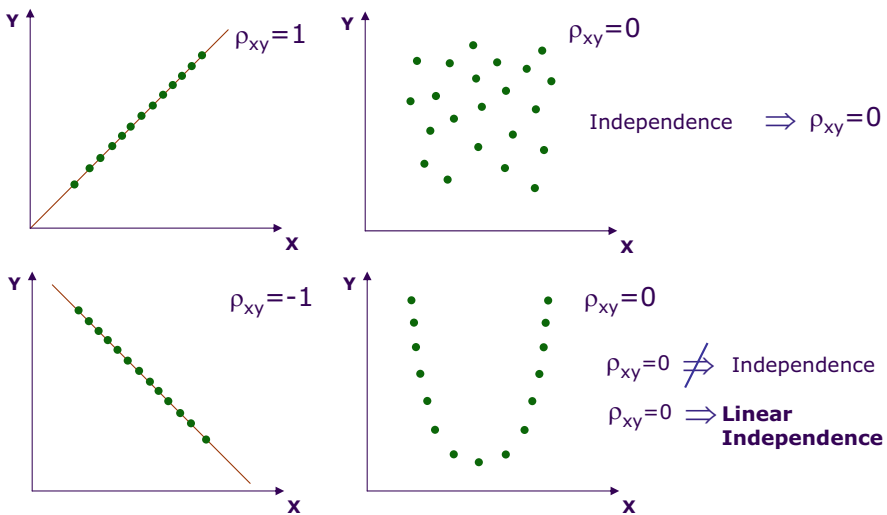


Fig. 8 Linear versus stochastic independence

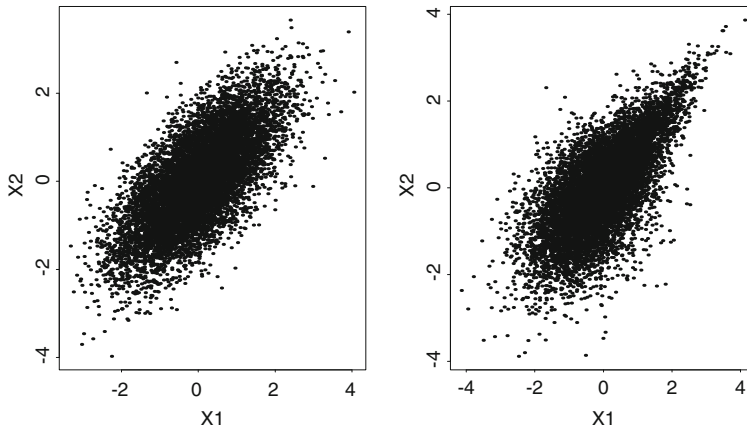


Fig. 9 Scatterplots of (X_1, X_2) with normal margins, linear correlation $\rho = 70\%$, and, respectively, a Gaussian copula (left plot) and a Gumbel copula (right plot)

For instance, consider two random vectors having the same standard normal margins and a linear correlation of 70%, but a different dependence structure, a Gaussian copula and a Gumbel one, respectively. We clearly see in Fig. 9 how different they are.

It emphasizes the fact that knowing the marginal distributions and linear correlation is not enough for determining the joint distribution, except for elliptical distributions (as e.g. the Gaussian ones).

- In 1984: Paul Deheuvels[10] introduced the notion of extreme-value copula.
- From the 1970s, diverse types of dependence have been studied in mathematical statistics and probability.
- From the twenty-first century, those dependence tools have been introduced in the industry: copulas turn out to become an important tool for applications and the evaluation of risks in insurance and reinsurance (and later in finance: non-linear tools cannot/should not be ignored anymore, especially after the second most severe financial crisis starting in 2008)
- After the 2008 financial crisis, Extreme Value Theory (EVT) finally enters the financial world (academics and professionals). The fact that risks are more interdependent in extreme situations led to the development of the notion of systemic risks, risks that would affect the entire system as well as the notion of systematic risks, where components are present in all other risks.

The world has changed a lot, from the end of the nineteenth to early twentieth century, where using the concept of linear correlation (Pearson) was of great help, to nowadays, where world is getting more complex, and more and more interconnected (see [6] for a discussion of this point). In the next few years, research in statistics and probability will have to make significant progress in this area if we want to

master the risk at an aggregate level. We have seen that societal demand goes in this direction, looking for protection at a global level.

3.3 Copulas

Definition 1 A copula is a multivariate distribution function $C : [0, 1]^d \rightarrow [0, 1]$ with standard uniform margins i.e. $C(1, \dots, 1, u_i, 1, \dots, 1) = u_i, \forall i \in \{1, \dots, d\}, u_i \in [0, 1]$.

Sklar showed in [35] how a unique copula C fully describes the dependence of X proving the following theorem.

Theorem 4 (Sklar’s Theorem, 1959) Let F be a joint cdf with margins $(F_i, i = 1, \dots, d)$. There exists a copula C such that

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)), \quad \forall x_i \in \mathbb{R}, \quad i = 1, \dots, d.$$

If the margins are continuous then C is unique.

Conversely, if C is a copula and $(F_i, 1 \leq i \leq d)$ are univariate cdf, then F defined above is a multivariate cdf with margins F_1, \dots, F_d .

Proof as an exercise.

As a consequence, we can give another definition of a copula.

Definition 2 The copula of (X_1, \dots, X_d) (or F) is the cdf C of $(F_1(X_1), \dots, F_d(X_d))$.

We sometimes refer to C as the *dependence structure* of F .

Here is a useful way to express Sklar’s theorem in dimension 2:

Theorem 5 (Sklar—dim 2) Let F be a joint cdf with margins (F_1, F_2) . The copula C associated to F can be written as

$$\begin{aligned} C(u_1, u_2) &= C(F_1(x_1), F_2(x_2)) \\ C(u_1, u_2) &= F(x_1, x_2) \\ C(u_1, u_2) &= F(F_1^{-1}(u_1), F_2^{-1}(u_2)) \end{aligned}$$

If the margins are continuous then C is unique.

Copulas satisfy a property of invariance, very useful in practice, and which is not satisfied by the linear correlation.

Property 1 (Property of Invariance) C is invariant under strictly increasing transformations of the marginals. If T_1, \dots, T_d are strictly increasing, then $(T_1(X_1), \dots, T_d(X_d))$ has the same copula as (X_1, \dots, X_d) .

As for probability distributions, we can define the notion of density function, when existing.

Definition 3 The *density function* c of a copula C is defined by

$$c(u_1, \dots, u_d) = \frac{\partial^d C(u_1, \dots, u_d)}{\partial u_1 \cdots \partial u_d}.$$

The density function of a bivariate distribution can be written in terms of the density function c of the associated copula and in terms of the density functions f_1 and f_2 of the margins:

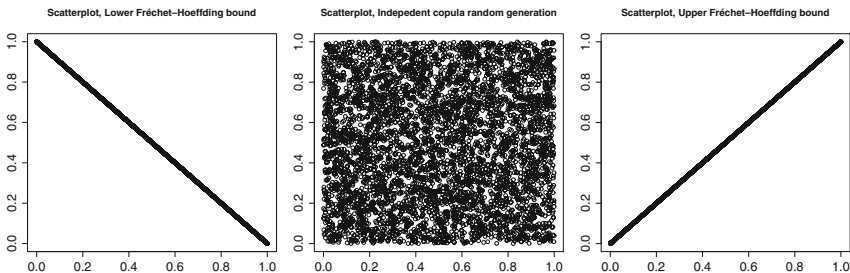
$$f(x_1, x_2) = c(F_1(x_1), F_2(x_2)) f_1(x_1) f_2(x_2).$$

Using this definition, we can prove that: *the product copula characterize the independence between two r.v.* More generally, X_1, \dots, X_d are mutually independent if and only if their copula C satisfies $C(u_1, \dots, u_d) = \prod_{i=1}^d u_i$.

As the linear correlation, which is bounded between -1 and 1 , a copula also admits bounds, named *Fréchet-Hoeffding bounds*:

$$\max\left(\sum_{i=1}^d u_i + 1 - d ; 0\right) \leq C(u) \leq \min_{1 \leq i \leq d} u_i = \mathbb{P}[U \leq u_1, \dots, U \leq u_d]$$

where $u = (u_1, \dots, u_d)$ and U is uniformly distributed on $[0,1]$.



The upper Fréchet-Hoeffding bound $C_u(u_1, \dots, u_d) := \min_{1 \leq i \leq d} u_i$ is a copula for any d . It describes the perfect dependence, named also *comotonicity*:

$$X_i \stackrel{a.s.}{=} T_i(X_1), \text{ with } T_i \text{ strictly increasing function, } i = 2, \dots, d \iff C_u \text{ satisfies } C_u(u_1, \dots, u_d) := \min_{1 \leq i \leq d} u_i.$$

The lower Fréchet-Hoeffding bound $C_l(u) := \max\left(\sum_{i=1}^d u_i + 1 - d ; 0\right)$ is a copula for $d = 2$, but not for all $d > 2$. For $d = 2$, C_l describes the perfect

negative dependence, named also *countercomotonicity*: $X_2 \stackrel{a.s.}{=} T(X_1)$, with T strictly decreasing function $\iff C_l(u_1, u_2) := \max(u_1 + u_2 - 1, 0)$.

Examples of Copulas Since any type of dependence structure can exist, the same can be said about copulas, that is why many new copulas are introduced by researchers. Here let us define three standard classes of copulas, already in use among practitioners, among which the Extreme Value (EV) copulas (which can overlap the two other classes).

- Elliptical or normal mixture copulas, as for instance:
 - The Gaussian copula (often used in financial modeling); in dimension 2, with parameter $\alpha \in (-1, 1)$, it is defined via Sklar’s theorem by

$$C(u, v) = \frac{1}{2\pi\sqrt{1-\alpha^2}} \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \exp\left\{-\frac{x^2 - 2\alpha xy + y^2}{2(1-\alpha^2)}\right\} dx dy$$

where Φ denotes the standard normal distribution.

- The Student- t copula is the distribution of $(T(X_i), i = 1, \dots, d)$ where T is the t -cdf and $(X_i, i = 1, \dots, d)$ has a joint t -distribution. For $d = 2$, it can be expressed (via Sklar’s theorem) as:

$$C(u, v) = \frac{1}{2\pi\sqrt{1-\alpha^2}} \int_{-\infty}^{t_v^{-1}(u)} \int_{-\infty}^{t_v^{-1}(v)} \left(1 + \frac{x^2 - 2\alpha xy + y^2}{2(1-\alpha^2)}\right)^{-(v+2)/2} dx dy$$

for $\alpha \in (-1, 1)$ and degrees of freedom $\nu \geq 2$.

- Archimedean copulas.
Definition. An Archimedean copula C is defined by

$$C(u_1, \dots, u_d) = \psi^{-1}(\psi(u_1) + \dots + \psi(u_d))$$

where $\psi :]0, 1] \rightarrow [0, \infty)$ is continuous, strictly decreasing, convex, and satisfies $\psi(1) = 0$ and $\lim_{t \rightarrow 0} \psi(t) = +\infty$; set $\psi^{-1}(t) = 0$ if $\psi(0) \leq t \leq +\infty$.

We call ψ the *strict generator of C*.

Archimedean copulas are *exchangeable*, i.e. invariant under permutation.

Examples:

- Gumbel copula, defined in dimension 2 by:

$$C_\beta^{Gu}(u, v) = \exp\left\{-\left((-\log u)^\beta + (-\log v)^\beta\right)^{1/\beta}\right\}, \quad \text{with } \beta \geq 1.$$

When $\beta = 1$: $C_1^{Gu}(u, v) = uv$ pointing out the independence of the variables.

When $\beta \rightarrow \infty$, the variables tend to be comonotonic: $C_\infty^{Gu}(u, v) = \min(u, v)$.

– Clayton copula, defined in dimension 2 by:

$$C_{\beta}^{Cl}(u, v) = (u^{-\beta} + v^{-\beta} - 1)^{-1/\beta}, \quad \text{with } \beta > 0.$$

We have $\lim_{\beta \searrow 0} C_{\beta}^{Cl}(u, v) = uv$, independence case, and $\lim_{\beta \rightarrow \infty} C_{\beta}^{Cl}(u, v) = \min(u, v)$, comonotonic case.

- Extreme Value (EV) Copulas

A copula C is said to be an Extreme Value (EV) copula if it satisfies the *max-stability* characteristic property:

$$\forall \gamma > 0, \quad C^{\gamma}(u_1, \dots, u_d) = C(u_1^{\gamma}, \dots, u_d^{\gamma}).$$

An alternative definition is the following, that we state e.g. in dimension 2:

$$C(u, v) = \exp \left\{ (\log u + \log v) A \left(\frac{\log u}{\log u + \log v} \right) \right\}$$

where A , called the dependence (or Pickands) function, is convex on $[0, 1]$ and satisfies $A(0) = A(1) = 1$ and $\max(1 - \omega, \omega) \leq A(\omega) \leq 1, \forall \omega \in [0, 1]$.

The function A can be defined from the EV copula C by setting

$$A(w) = -\ln C(e^{-w}, e^{-(1-w)}), \quad w \in [0, 1].$$

Bounds have also been provided for A . If the upper bound is reached, i.e. if $A(w) = 1, \forall w$, then C is the independence copula. If the lower bound is reached, i.e. if $A(w) = \max(w, 1 - w)$, then it is a comonotonicity copula.

Examples of EV copulas: independence copula, comonotonicity copula, Gumbel copula (it is a parametric EV copula; the Gumbel copula model is sometimes known as the logistic model), Galambos copula (as defined below).

Let us consider the dependence function A , introduced by Galambos and defined, for $0 \leq w \leq 1$, with $0 \leq \alpha, \beta \leq 1$ and $\theta > 0$, by

$$A(w) = 1 - \left((\alpha w)^{-\theta} + (\beta(1 - w))^{-\theta} \right)^{-1/\theta}.$$

We can check that A is a convex function having the right bounds for the definition of an EV copula, so that we can create an EV copula from A . We obtain the bivariate EV copula, named Galambos copula (this copula model is sometimes known as the negative logistic model),

$$C_{\theta, \alpha, \beta}^{Gal}(u, v) = uv \exp \left\{ \left((-\alpha \ln u)^{-\theta} + (-\beta \ln v)^{-\theta} \right)^{-1/\theta} \right\}.$$

It is represented in Fig. 10 [27, p. 313].

To learn more on copulas, refer e.g. to [28] and [20].

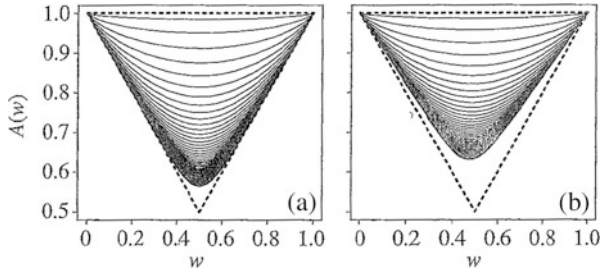


Fig. 10 Plot of dependence function for (a) the symmetric Galambos ($\alpha = \beta = 1$), spanning the whole range from independence to comonotonicity, and (b) the asymmetric Galambos copula with $\alpha = 0.9$ and $\beta = 0.8$; the limit as $\theta \rightarrow 0$ is the independence model, whereas as $\theta \rightarrow \infty$, it is no longer the comonotonicity model. Dashed lines show boundaries of the triangle in which the dependence function must reside; solid lines show dependence functions for a range of θ values running from 0.2 to 5 in steps of size 0.1

3.4 Notion of Rank Correlation

Let us introduce two rank correlations, the Spearman’s rho ρ_S and the Kendall’s tau ρ_τ , which can also be expressed in terms of copulas (see e.g. [27, §5.2]).

Let C denote the copula of (X_1, X_2) , and ρ the Pearson (linear) correlation of X_1 and X_2 .

- The Spearman’s rho ρ_S is defined by

$$\rho_S(X_1, X_2) = \rho(F_1(X_1), F_2(X_2)) = \rho(\text{copula})$$

and also by

$$\rho_S(X_1, X_2) = 12 \int_0^1 \int_0^1 (C(u_1, u_2) - u_1 u_2) du_1 du_2.$$

- The Kendall’s tau ρ_τ is defined by

$$\rho_\tau(X_1, X_2) = 2\mathbb{P}[(X_1 - \tilde{X}_1)(X_2 - \tilde{X}_2) > 0] - 1$$

with $(\tilde{X}_1, \tilde{X}_2)$ an independent copy of (X_1, X_2) , and also by

$$\rho_\tau(X_1, X_2) = 4 \int_0^1 \int_0^1 C(u_1, u_2) dC(u_1, u_2) - 1.$$

Case of elliptical models: Suppose $X = (X_1, X_2)$ has any elliptical distribution (e.g. X has a Student distribution $t_2(\nu, \mu, \Gamma)$). Then $\rho_\tau(X_1, X_2) = \frac{2}{\pi} \arcsin(\rho(X_1, X_2))$.

Note that if X_i has infinite variance, then $\rho(X_1, X_2)$ can be interpreted as $\frac{r_{1,2}}{\sqrt{r_{1,1}r_{2,2}}}$.

3.4.1 Properties of Rank Correlations

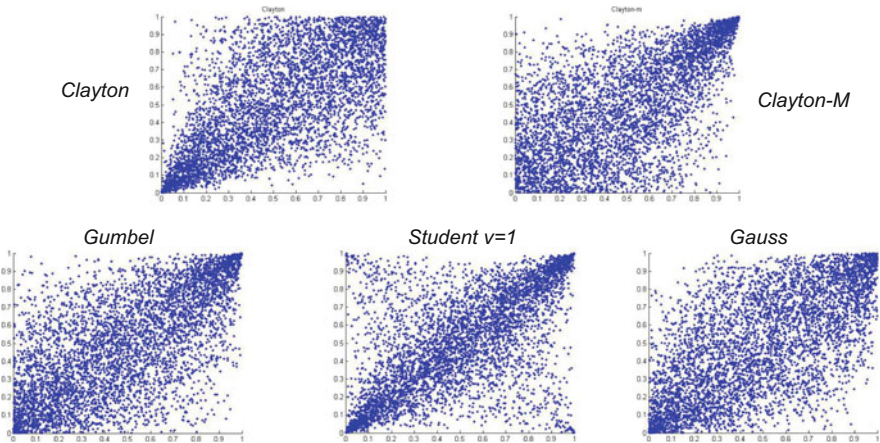
We can enunciate the following properties for the Spearman's rho ρ_S . The same holds true for Kendall's tau ρ_τ . But those properties are not shared by the linear correlation.

1. ρ_S depends only on the copula of (X_1, X_2) ;
2. ρ_S is invariant under strictly increasing transformations of the rv's;
3. $\rho_S(X_1, X_2) = 1 \Leftrightarrow C(X_1, X_2)$ is comonotonic;
4. $\rho_S(X_1, X_2) = -1 \Leftrightarrow C(X_1, X_2)$ is countermonotonic.

3.5 Ranked Scatterplots

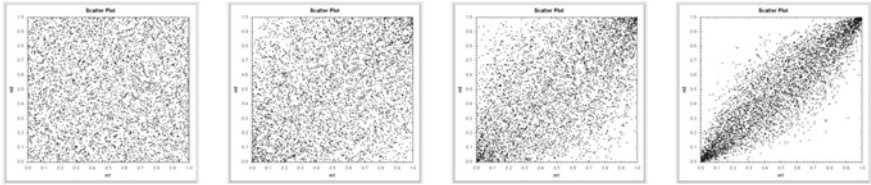
Let us draw ranked scatterplots with different copulas.

First we consider archimedean copulas (with parameter θ), namely Clayton, Clayton-mirror (i.e. when we flip it) and Gumbel, and elliptical ones, namely Student copula with $\nu = 1$ and Gaussian one. For all of them, we choose the Kendall's tau $\rho_\tau = 50\%$.

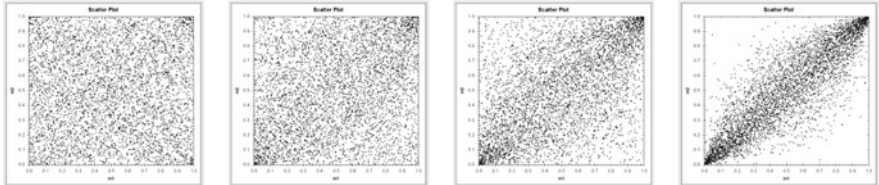


Then we consider the same margins and play with the parameters of the copulas to see their impact.

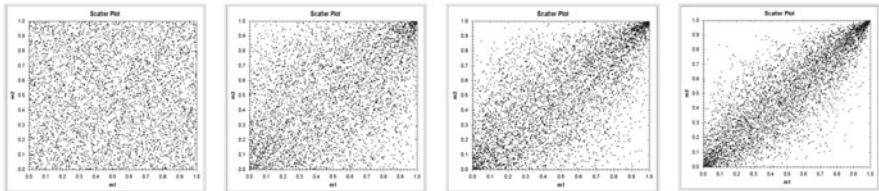
Gaussian copula with $\rho = 0, 30, 60, 90$ % from left to right



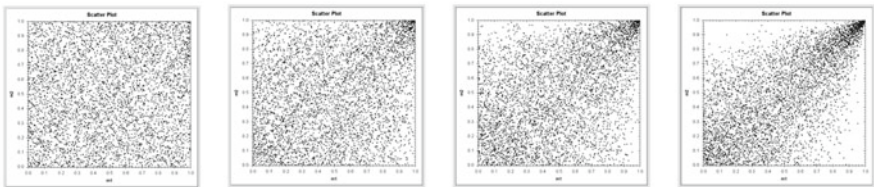
Student t_3 copula with $\rho = 0, 30, 60, 90$ % from left to right



Gumbel copula with $\theta = 1, 1.5, 2, 3$ from left to right



Survival (Mirror) Clayton copula with $\theta = 0.1, 0.5, 1, 2$ from left to right



3.6 Other Type of Dependence: Tail or Extremal Dependence

The objective is to measure the dependence in joint tail of bivariate distribution. Let C denote the copula of the random vector (X_1, X_2) .

- *Coefficient of upper tail dependence.* When the limit exists, it is defined as

$$\lambda_u(X_1, X_2) = \lim_{\alpha \rightarrow 1} \mathbb{P}[X_2 > VaR_\alpha(X_2) \mid X_1 > VaR_\alpha(X_1)]$$

and, as function of the copula C ,

$$\lambda_u(X_1, X_2) = \lim_{\alpha \rightarrow 1} \frac{1 - 2\alpha + C(\alpha, \alpha)}{1 - \alpha}.$$

- *Coefficient of lower tail dependence.* When the limit exists, it is defined as

$$\lambda_l(X_1, X_2) = \lim_{\alpha \rightarrow 0} \mathbb{P}[X_2 \leq VaR_\alpha(X_2) \mid X_1 \leq VaR_\alpha(X_1)]$$

and, as function of C ,

$$\lambda_l(X_1, X_2) = \lim_{\alpha \rightarrow 0} \frac{C(\alpha, \alpha)}{\alpha}.$$

3.6.1 Properties and Terminology

1. $\lambda_u \in [0, 1]$ and $\lambda_l \in [0, 1]$;
2. For elliptical copulas, $\lambda_u = \lambda_l := \lambda$. Note that this is true for all copulas with radial symmetry, i.e. such that $(U_1, U_2) \stackrel{d}{=} (1 - U_1, 1 - U_2)$;
3. If $\lambda_u \in (0, 1]$, then there exists an upper tail dependence and if $\lambda_l \in (0, 1]$, there exists a lower tail dependence;
4. $\lambda_u = 0$ means that there is asymptotic independence in the upper tail and $\lambda_l = 0$ means that there is asymptotic independence in lower tail.

Examples

1. We can prove that a Gaussian copula with parameter ρ is asymptotically independent (i.e. $\lambda = 0$) whenever $|\rho| < 1$;
2. A t -copula with parameter ρ is tail dependent whenever $\rho > -1$, whatever is the number of degrees of freedom ν . Its coefficient of (lower and upper) tail dependence is given by: $\lambda = 2\bar{t}_{\nu+1} \left(\sqrt{1 + \rho} \sqrt{\frac{1 - \rho}{1 + \rho}} \right)$;
3. The Gumbel copula with parameter β is upper tail dependent for $\beta > 1$, and this upper tail dependence is measured by $\lambda_u = 2 - 2^{1/\beta}$;
4. The Clayton copula with parameter β is lower tail dependent for $\beta > 0$, and $\lambda_l = 2^{-1/\beta}$.

The properties of symmetric tail dependence, as well as of asymptotic tail dependence for the Gaussian copula and upper tail dependence for the Student copula, are well illustrated in Fig. 11.

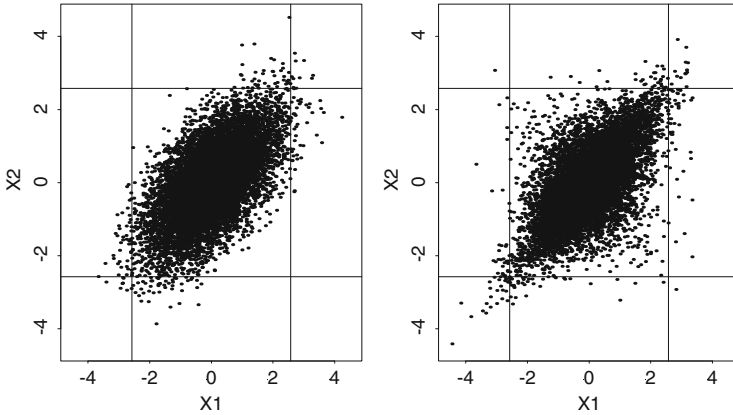


Fig. 11 Gaussian (left) and student t_3 (right) copulas with same margins and parameter $\rho = 70\%$. Quantiles lines are given for 0.5% and 99.5%

Table 10 Left table: joint tail probabilities $\mathbb{P}[X_1 > VaR_\alpha(X_1), X_2 > VaR_\alpha(X_2)]$ for $\alpha = 95, 99, 99.5, 99.9\%$, respectively. Right table: Joint tail probabilities $\mathbb{P}[X_i > VaR_{99\%}(X_i), i = 1, \dots, d]$ for $d = 2, 3, 4, 5$ respectively, when taking equal correlations

ρ	C	Quantile				ρ	C	Dimension d			
		95%	99%	99.5%	99.9%			2	3	4	5
0.5	N	1.21×10^{-2}	1.29×10^{-3}	4.96×10^{-4}	5.42×10^{-5}	0.5	N	1.29×10^{-3}	3.66×10^{-4}	1.49×10^{-4}	7.48×10^{-5}
0.5	t8	1.20	1.65	1.94	3.01	0.5	t8	1.65	2.36	3.09	3.82
0.5	t4	1.39	2.22	2.79	4.86	0.5	t4	2.22	3.82	5.66	7.68
0.5	t3	1.50	2.55	3.26	5.83	0.5	t3	2.55	4.72	7.35	10.34
0.7	N	1.95×10^{-2}	2.67×10^{-3}	1.14×10^{-3}	1.60×10^{-4}	0.7	N	2.67×10^{-3}	1.28×10^{-3}	7.77×10^{-4}	5.35×10^{-4}
0.7	t8	1.11	1.33	1.46	1.86	0.7	t8	1.33	1.58	1.78	1.95
0.7	t4	1.21	1.60	1.82	2.52	0.7	t4	1.60	2.10	2.53	2.91
0.7	t3	1.27	1.74	2.01	2.83	0.7	t3	1.74	2.39	2.97	3.45

For both tables: The copula C of the random vector is either Gaussian (denoted by N) or Student t with three possible degrees of freedom $\nu = 8, 4, 3$ (the smaller is ν , the heavier is the tail) and parameter $\rho = 50\%$ or 70% . Note that for the Student cases, only the factor by which Gaussian (N) joint tail probability must be multiplied, is given

3.6.2 Numerical Example Showing the Impact of the Choice of Copula

We already provided in Sect. 3.1.2 an example with political risks where we observed how much the choice of the dependence structure would impact the results. In that example we considered a Gaussian dependence versus a Clayton one.

Let us give another example where we compare in Table 10 the joint tail probability at finite levels of two copulas which are both elliptical. This is an example developed by McNeil et al. in [27].

Let us illustrate those results, giving the financial interpretation suggested in [27].

Consider daily returns on five financial instruments and suppose that we believe that all correlations between returns are equal to 50%. However, we are unsure about the best multivariate model for these data. On one hand, if returns follow a multivariate Gaussian distribution then the probability that on any day all returns fall below their 1% quantiles is 7.48×10^{-5} . In the long run such an event will

happen once every 13,369 trading days on average, that is roughly once every 51.4 years (assuming 260 trading days in a year). On the other hand, if returns follow a multivariate t distribution with four degrees of freedom then such an event will happen 7.68 times more often, that is roughly once every 6.7 years, which would induce a very different behavior in terms of risk management! During the subprime crisis, this was the problem of the too high rating given to the CDOs (Collateralized Debt Obligation) by the rating agencies, who only considered linear correlation for the dependence between the risks.

4 Multivariate EVT

Let us end those notes by giving a brief idea about the basis on which EVT has been extended in the multivariate setting. It is a research domain which has aroused an increasing interest this past decade, in particular due to its practical use.

4.1 MEV Distribution

Some Notation Let $X_1, \dots, X_i, \dots, X_n$ be iid random vectors in \mathbb{R}^d , each X_i ($i = 1, \dots, n$) having its components denoted by X_{ij} , $j = 1, \dots, d$; they could be interpreted as losses of d different types. Let F be the joint cdf of any random vector X_i and F_1, \dots, F_d be its marginal cdf's. Let $M_{nj} = \max_{1 \leq i \leq n} X_{ij}$, for $j = 1, \dots, d$; it is the maximum of the j th component, and M_n be the d -random vector the vector of componentwise block maxima, i.e. with components M_{nj} , $j = 1, \dots, d$.

The main question that might be asked, when going from univariate EVT to multivariate one, is which underlying multivariate cdf's F are attracted to which MEV distributions H ?

Definition 4 If there exist vectors of normalizing constants (of dimension d) $c_n > 0$ and d_n such that $(M_n - d_n)/c_n$ converges in distribution to a random vector with joint (non-degenerated) cdf H , i.e.

$$\mathbb{P} \left[\frac{M_n - d_n}{c_n} \leq x \right] = F^n(c_n x + d_n) \xrightarrow[n \rightarrow \infty]{} H(x), \quad x \in \mathbb{R}^d,$$

we say that F is in the Maximum Domain of Attraction of H , written $F \in MDA(H)$, and we refer to H as a MEV (Multivariate Extreme Value) distribution.

If H has non-degenerate margins, then

- these margins are univariate EV distributions of one of the three types, by application of univariate EVT;

- via Sklar’s theorem, H has a copula, which is unique if the margins are continuous.

Theorem 6 *If $F \in MDA(H)$ for some F and H with GEV margins, then the unique copula C of H satisfies the scaling property:*

$$C^\gamma(u) = C(u^\gamma), \quad \forall u \in \mathbb{R}^d, \forall \gamma > 0,$$

which means that C is an extreme value (EV) copula (as defined previously); it can then be the copula of a MEV distribution.

4.2 Copula Domain of Attraction

We can enunciate the following asymptotic theorem.

Theorem 7 ([14]) *Let $F_i, i = 1, \dots, d$, be some continuous marginals cdf’s and C some copula. Let define $F(x) = C(F_1(x_1), \dots, F_d(x_d))$ and let $H(x) = C_0(H_1(x_1), \dots, H_d(x_d))$ be a MEV distribution with EV copula C_0 . Then we have*

$$F \in MDA(H) \text{ if and only if } \begin{cases} F_i \in MDA(H_i) \text{ for } i = 1, \dots, d, \\ \text{and} \\ \lim_{t \rightarrow \infty} C^t(u^{1/t}) = C_0(u), \quad u \in [0, 1]^d. \end{cases}$$

Notice that:

- the marginal distributions of F determine the margins of the MEV limit but are irrelevant to the determination of its dependence structure;
- the copula C_0 of the limiting MEV distribution is determined solely by the copula C of the underlying distribution.

Definition 5 *If $\lim_{t \rightarrow \infty} C^t(u^{1/t}) = C_0(u), u \in [0, 1]^d$, for some C and some EV copula C_0 , then we say that C belongs to the copula domain of attraction of C_0 : $C \in CDA(C_0)$.*

4.2.1 Upper Tail Dependence and CDA

Proposition 1 *Let C be a bivariate copula with upper tail-dependence coefficient λ_u . Assume that $C \in MDA(C_0)$ for some EV copula C_0 with Pickands (dependence) function A . Then λ_u is also the upper tail-dependence coefficient of C_0 and is related to its dependence function by $\lambda_u = 2(1 - A(1/2))$.*

Proof First, let us prove that C and C_0 have the same λ_u . To do so, we just need to check that $\lim_{\alpha \rightarrow 1} \frac{1 - C(\alpha, \alpha)}{1 - \alpha} = \lim_{\alpha \rightarrow 1} \frac{1 - C_0(\alpha, \alpha)}{1 - \alpha}$. We have, using the definition of $C \in CDA(C_0)$,

$$\begin{aligned} \lim_{\alpha \rightarrow 1} \frac{1 - C_0(\alpha, \alpha)}{1 - \alpha} &= \lim_{\alpha \rightarrow 1} \frac{\log C_0(\alpha, \alpha)}{1 - \alpha} = \lim_{\alpha \rightarrow 1} \lim_{t \rightarrow \infty} \frac{\log(t[1 - C(\alpha^{1/t}, \alpha^{1/t})])}{1 - \alpha} \\ &= \lim_{\alpha \rightarrow 1} \lim_{s \rightarrow 0^+} \frac{1 - C(\alpha^s, \alpha^s)}{-s \log(\alpha)} = \lim_{\alpha \rightarrow 1} \lim_{s \rightarrow 0^+} \frac{1 - C(\alpha^s, \alpha^s)}{-\log(\alpha^s)} = \lim_{\beta \rightarrow 1^-} \frac{1 - C(\beta, \beta)}{1 - \beta}. \end{aligned}$$

hence the result. The converse is straightforward. \square

Consequence: $\lambda_u = 0 \Rightarrow A(1/2) = 1 \Rightarrow A \equiv 1$ (since A convex function) $\Leftrightarrow C_0$ is the independence copula.

5 Conclusion

In these notes, we have explored the EVT both in the univariate as well as in the multivariate case by looking at dependence between rv's. We have seen that there are mature methods for determining accurately the shape of the tail of the distribution. There are also methods to backtest statistically if the model captures it correctly. This can be done when choosing Expected Shortfall as a risk measure (see e.g. [23] and references therein). We pointed out through examples the importance for good risk management of accounting for extreme risks, but also of correctly modeling the non-linear dependence when present in the data or in the process to be studied. We hope to have shown that there is no excuse anymore to ignore EVT in quantitative risk management.

Acknowledgements I would like to thank the organizers of this workshop, Professors Konstantin Borovkov, Kais Hamza and Alexander Novikov, for the invitation to lecture on this topic.

References

1. Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J.: Statistics of Extremes: Theory and Applications. Wiley, Chichester (2004)
2. Beirlant, J., Fraga Alves, M.I., Gomes, M.I.: Tail fitting for truncated and non-truncated Pareto-type distributions. *Extremes* **19**, 429–462 (2016)
3. Bingham, N., Goldie, C., Teugels, J.: Regular Variation. Cambridge University Press, Cambridge (1989)
4. Busse, M., Dacorogna, M., Kratz, M.: Does risk diversification always work? The answer through simple modelling. In: SCOR Paper **24** (2013)
5. Busse, M., Dacorogna, M., Kratz, M.: The impact of systemic risk on the diversification benefits of a risk portfolio. *Risks* **2**, 260–276 (2014)

6. Dacorogna, M., Kratz, M.: Living in a stochastic world and managing complex risks. Available on SSRN at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2668468 (2015)
7. Davison, A., Smith, R.: Models for exceedances over high thresholds. *J. Royal Stat. Soc. B* **52**(3), 393–442 (1990)
8. de Haan, L., Ferreira A.: *Extreme Value Theory: An Introduction*. Springer, New York (2007)
9. Debbabi, N., Kratz, M., Mboup, M.: A self-calibrating method for heavy tailed data modelling. Application in neuroscience and finance. Submitted Preprint (arXiv1612.03974v2) (2017)
10. Deheuvels, P.: Probabilistic aspects of multivariate extremes. In: Tiago de Oliveira, J. (ed.) *Statistical extremes and applications*. Reidel, Dordrecht (1984)
11. Dekkers, A.L.M., Einmahl, J.H.J., De Haan, L.: A moment estimator for the index of an extreme-value distribution. *Ann. Stat.* **17**(4), 1833–1855 (1989)
12. Embrechts, P., Klüppelberg, C., Mikosch, T.: *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin (1997)
13. Emmer, S., Kratz, M., Tasche, D.: What is the best risk measure in practice? A comparison of standard measures. *J. Risk* **18**(2), 31–60 (2015)
14. Galambos, J.: *The Asymptotic Theory of Extreme Order Statistics*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley, New York (1987)
15. Gnedenko, B.V.: Sur la distribution limite du terme maximum d’une série aléatoire. *Ann. Math.* **44**, 423–453 (1943)
16. Hauksson, H.A., Dacorogna, M.M., Domenig, T., Müller, U., Samorodnitsky, G.: Multivariate extremes, aggregation and risk estimation. *Quant. Finance* **1**, 79–95 (2001)
17. Hill, B.: A simple approach to inference about the tail of a distribution. *Ann. Stat.* **3**, 1163–1174 (1975)
18. Hosking, J.R.M., Wallis, J.R., Wood, E.F.: Estimation of the generalized extreme-value distribution by the method of probability-weighted moments. *Technometrics* **27**(3), 251–261 (1985)
19. Jenkinson, A.F.: The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Q. J. R. Meteorol. Soc.* **81**, 158–171 (1955)
20. Joe, H.: *Dependence Modeling with Copulas*. Chapman and Hall/CRC, New York (2014)
21. Kratz, M.: Normex, a new method for evaluating the distribution of aggregated heavy tailed risks. Application to risk measures. *Extremes* **17**(4), 661–691 (2014). Special issue on *Extremes and Finance* (Guest Ed. P. Embrechts)
22. Kratz, M., Resnick, S.: The QQ-estimator and heavy tails. *Stoch. Models* **12**, 699–724 (1996)
23. Kratz, M., Lok, Y.H., McNeil, A.: Multinomial VaR Backtests: a simple implicit approach to backtesting expected shortfall. *J. Bank. Finance* **88**, 393–407 (2018)
24. Leadbetter, R., Lindgren, G., Rootzén, H.: *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York (1983)
25. Longin, F. (ed.): *Extreme Events in Finance. A Handbook of Extreme Value Theory and its Applications*. Wiley, Hoboken (2016)
26. Markowitz, H.: Portfolio selection. *J. Finance* **7**(1), 77–91 (1952)
27. McNeil, A., Frey, R., Embrechts, P.: *Quantitative Risk Management*. Princeton Series in Finance. Princeton University Press, Princeton (2005; 2016, 2nd Ed.)
28. Nelsen, R.: *An introduction to copulas*. Springer, New York (1999)
29. Novak, S.: *Extreme Value Methods with Applications to Finance*. Chapman & Hall/CRC Press, London (2011)
30. Pearson, K.: Notes on regression and inheritance in the case of two parents. *Proc. R. Soc. Lond.* **58**, 240–242 (1895)
31. Reiss, R.-D., Thomas, M.: *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*. Birkhäuser Verlag, Basel (1997; 2007, 2nd Ed.)
32. Resnick, S.: *Extreme Values, Regular Variation, and Point Processes*. Springer, New York (1987; 2008, 2nd Ed.)
33. Resnick, S.: *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, New York (2006)

34. Sharpe, H.F.: The sharpe ratio. *J. Portf. Manag.* **21**(1), 49–58 (1994)
35. Sklar, A.: Fonctions de répartition à n dimensions et leurs marges, vol. 8, pp. 229–231. Publications de l'Institut de Statistique de l'Université de Paris (ISUP), Paris (1959)
36. von Mises, R.: La distribution de la plus grande de n valeurs (1936). Reprinted in Selected Papers Volume II, pp. 271–294. American Mathematical Society, Providence, RI (1954)

Monotone Sharpe Ratios and Related Measures of Investment Performance



Mikhail Zhitlukhin

Abstract We introduce a new measure of performance of investment strategies, the monotone Sharpe ratio. We study its properties, establish a connection with coherent risk measures, and obtain an efficient representation for using in applications.

1 Introduction

This paper concerns the problem of evaluation of performance of investment strategies. By performance, in a broad sense, we mean a numerical quantity which characterizes how good the return rate of a strategy is, so that an investor typically wants to find a strategy with high performance.

Apparently, the most well-known performance measure is the Sharpe ratio, the ratio of the expectation of a future return, adjusted by a risk-free rate or another benchmark, to its standard deviation. It was introduced by William F. Sharpe in the 1966 paper [23], a more modern look can be also found in [24]. The Sharpe ratio is based on the Markowitz mean-variance paradigm [14], which assumes that investors need to care only about the mean rate of return of assets and the variance of the rate of return: then in order to find an investment strategy with the smallest risk (identified with the variance of return) for a given desired expected return, one just needs to find a strategy with the best Sharpe ratio and diversify appropriately between this strategy and the risk-free asset (see a brief review in Sect. 2 below). Despite its simplicity, as viewed from today's economic science, the Markowitz portfolio theory was a major breakthrough in mathematical finance. Even today, more than 65 years later, analysts still routinely compute Sharpe ratios of investment portfolios and use it, among other tools, to evaluate performance.

In the present paper we look at this theory in a new way, and establish connections with much more recent developments. The main part of the material

M. Zhitlukhin (✉)

Steklov Mathematical Institute of Russian Academy of Sciences, Moscow, Russia

e-mail: mikhailzh@mi.ras.ru

© Springer Nature Switzerland AG 2019

D. R. Wood et al. (eds.), 2017 *MATRIX Annals*, MATRIX Book Series 2,

https://doi.org/10.1007/978-3-030-04161-8_52

637

of the paper developed from a well-known observation that variance is not identical to risk: roughly speaking, one has to distinguish between “variance above mean” (which is good) and “variance below mean” (which is bad). In particular, the Sharpe ratio lacks the property of monotonicity, i.e. there might exist an investment strategy which always yields a return higher than another strategy, but has a smaller Sharpe ratio. The original goal of this work was to study a modification of the Sharpe ratio, which makes it monotone. Some preliminary results were presented in [27, 28]. It turned out, that the modified Sharpe ratio possesses interesting properties and is tightly connected to the theory of risk measures. The study of them is the subject of this paper.

The modification of the Sharpe ratio we consider, which we call the monotone Sharpe ratio, is defined as the maximum of the Sharpe ratios of all probability distributions that are dominated by the distribution of the return of some given investment strategy. In this paper we work only with *ex ante* performance measure, i.e. assume that probability distributions of returns are known or can be modeled, and one needs to evaluate their performance; we leave aside the question how to construct appropriate models and calibrate them from data.

The theory we develop focuses on two aspects: on one hand, to place the new performance measure on a modern theoretical foundation, and, on the other hand, take into account issues arising in applications, like a possibility of fast computation and good properties of numerical results. Regarding the former aspect, we can mention the paper of Cherny and Madan [3], who studied performance measures by an axiomatic approach. The abstract theory of performance measures they proposed is tightly related to the theory of convex and coherent risk measures, which has been a major breakthrough in the mathematical finance in the past two decades. We show that the monotone Sharpe ratio satisfies those axioms, which allows to apply results from the risk measures theory to it through the framework of Cherny and Madan. Also we establish a connection with more recently developed objects, the so-called buffered probabilities, first introduced by Rockafellar and Royset in [21] and now gaining popularity in applications involving optimization under uncertainty. Roughly speaking, they are “nice” alternatives to optimization criteria involving probabilities of adverse events, and lead to solutions of optimization problems which have better mathematical properties compared to those when standard probabilities are used. One of main implications of our results is that the portfolio selection problem with the monotone Sharpe ratio is equivalent to minimization of the buffered probability of loss.

Addressing the second aspect mentioned above, our main result here is a representation of the monotone Sharpe ratio as a solution of some convex optimization problem, which gives a computationally efficient way to evaluate it. Representations of various functionals in such a way are well-known in the literature on convex optimization. For example, in the context of finance, we can mention the famous result of Rockafellar and Uryasev [22] about the representation of the conditional value at risk. That paper also provides a good explanation why such a representation is useful in applications (we also give a brief account on that below).

Our representation also turns out to be useful in stochastic control problems related to maximization of the Sharpe ratio in dynamic trading. Those problems are known in the literature as examples of stochastic control problems where the Bellman optimality principle cannot be directly applied. With our theory, we are able to find the optimal strategies in a shorter and simpler way, compared to the results previously known in the literature.

Finally, we would like to mention, that in the literature a large number of performance measures have been studied. See for example papers [4, 5, 10] providing more than a hundred examples of them addressing various aspects of evaluation of quality of investment strategies. We believe that due to both the theoretical foundation and the convenience for applications, the monotone Sharpe ratio is a valuable contribution to the field.

The paper is organized as follows. In Sect. 2 we introduce the monotone Sharpe ratio and study its basic properties which make it a reasonable performance measure. There we also prove one of the central results, the representation as a solution of a convex optimization problem. In Sect. 3, we generalize the concept of the buffered probability and establish a connection with the monotone Sharpe ratio, as well as show how it can be used in portfolio selection problems. Section 4 contains applications to dynamic problems.

2 The Monotone Sharpe Ratio

2.1 Introduction: Markowitz Portfolio Optimization and the Sharpe Ratio

Consider a one-period market model, where an investor wants to distribute her initial capital between $n + 1$ assets: one riskless asset and n risky assets. Assume that the risky assets yield return R_i , $i = 1, \dots, n$, so that \$1 invested “today” in asset i turns into $\$(1 + R_i)$ “tomorrow”; the rates of return R_i are random variables with known distributions, such that $R_i > -1$ with probability 1 (no bankrupts happen). The rate of return of the riskless asset is constant, $R_0 = r > -1$. We always assume that the probability distributions of R_i are known and given, and, for example, do not consider the question how to estimate them from past data. In other words, we always work with *ex ante* performance measures (see [24]).

An investment portfolio of the investor is identified with a vector $x \in \mathbb{R}^{n+1}$, where x_i is the proportion of the initial capital invested in asset i . In particular, $\sum_i x_i = 1$. Some coordinates x_i may be negative, which is interpreted as short sales ($i = 1, \dots, n$) or loans ($i = 0$). It is easy to see that the total return of the portfolio is $R_x = \langle x, R \rangle := \sum_i x_i R_i$.

The Markowitz model prescribes the investor to choose the optimal investment portfolio in the following way: she should decide what expected return ER_x she wants to achieve, and then find the portfolio x which minimizes the variance of the

return $\text{Var } R_x$. This leads to the quadratic optimization problem:

$$\begin{aligned} & \text{minimize} \quad \text{Var } R_x \text{ over } x \in \mathbb{R}^{n+1} \\ & \text{subject to} \quad \text{E}R_x = \mu \\ & \qquad \qquad \sum_i x_i = 1. \end{aligned} \tag{1}$$

Under mild conditions on the joint distribution of R_i , there exists a unique solution x^* , which can be easily written explicitly in terms of the covariance matrix and the vector of expected returns of R_i (the formula can be found in any textbook on the subject, see, for example, Chapter 2.4 in [18]).

It turns out that points $(\sigma_{x^*}, \mu_{x^*})$, where $\sigma_{x^*} = \sqrt{\text{Var } R_{x^*}}$, $\mu_{x^*} = \text{E}R_{x^*}$ correspond to the optimal portfolios for all possible expected returns $\mu \in [r, \infty)$, lie on the straight line in the plane (σ, μ) , called the efficient frontier. This is the set of portfolios the investor should choose from—any portfolio below this line is inferior to some efficient portfolio (i.e. has the same expected return but larger variance), and there are no portfolios above the efficient frontier.

The slope of the efficient frontier is equal to the Sharpe ratio of any efficient portfolio containing a non-zero amount of risky assets (those portfolios have the same Sharpe ratio). Recall that the Sharpe ratio of return R is defined as the ratio of the expected return adjusted by the risk-free rate to its standard deviation

$$S(R) = \frac{\text{E}(R - r)}{\sqrt{\text{Var } R}}.$$

In particular, to solve problem (1), it is enough to find some efficient portfolio \hat{x} , and then any other efficient portfolio can be constructed by a combination of the riskless portfolio $x_0 = (1, 0, \dots, 0)$ and \hat{x} , i.e. $x^* = (1 - \lambda)x_0 + \lambda\hat{x}$, where $\lambda \in [0, +\infty)$ is chosen to satisfy $\text{E}R_{x^*} = \mu \geq r$. This is basically the statement of the Mutual Fund Theorem. Thus, the Sharpe ratio can be considered as a measure of performance of an investment portfolio and an investor is interested in finding a portfolio with the highest performance. In practice, broad market indices can be considered as quite close to efficient portfolios.

The main part of the material in this paper grew from the observation that the Sharpe ratio is not monotone: for two random variables X, Y the inequality $X \leq Y$ a.s. does not imply the same inequality between their Sharpe ratios, i.e. that $S(X) \leq S(Y)$. Here is an example: let X have the normal distribution with mean 1 and variance 1 and $Y = X \wedge 1$; obviously, $S(X) = 1$ but one can compute that $S(Y) > 1$. From the point of view of the portfolio selection problem, this fact means that it is possible to increase the Sharpe ratio by disposing part of the return (or consuming it). This doesn't agree well with the common sense interpretation of efficiency. Therefore, one may want to look for a replacement of the Sharpe ratio, which will not have such a non-natural property.

In this paper we'll use the following simple idea: if it is possible to increase the Sharpe ratio by disposing a part of the return, let's define the new performance measure as the maximum Sharpe ratio that can be achieved by such a disposal. Namely, define the new functional by

$$\mathbb{S}(X) = \sup_{C \geq 0} S(X - C),$$

where the supremum is over all non-negative random variables C (defined on the same probability space as X), which represent the disposed return. In the rest of this section, we'll study such functionals and how they can be used in portfolio selection problems. We'll work in a more general setting and consider not only the ratio of expected return to standard deviation of return but also ratios of expected return to deviations in L^p . The corresponding definitions will be given below.

2.2 The Definition of the Monotone Sharpe Ratio and Its Representation

In this section we'll treat random variables as returns of some investment strategies, unless otherwise is stated. That is, large values are good, small values are bad. Without loss of generality, we'll assume that the risk-free rate is zero, otherwise one can replace a return X with $X - r$, and all the results will remain valid.

First we give the definition of a deviation measure in L^p , $p \in [1, \infty)$, which will be used in the denominator of the Sharpe ratio instead of the standard deviation (the latter one is a particular case for $p = 2$). Everywhere below, $\|\cdot\|_p$ denotes the norm in L^p , i.e. $\|X\|_p = (E|X|^p)^{\frac{1}{p}}$.

Definition 1 We define the L^p -deviation of a random variable $X \in L^p$ as

$$\sigma_p(X) = \min_{c \in \mathbb{R}} \|X - c\|_p.$$

In the particular case $p = 2$, as is well-known, $\sigma_2(X)$ is the standard deviation, and the minimizer is $c^* = EX$. For $p = 1$, the minimizer $c^* = \text{med}(X)$, the median of the distribution of X , so that $\sigma_1(X)$ is the absolute deviation from the median. It is possible to use other deviation measures to define the monotone Sharpe ratio, for example $\|X - EX\|_p$, but the definition given above seems to be the most convenient for our purposes.

Observe that σ_p obviously satisfies the following properties, which will be used later: (a) it is sublinear; (b) it is uniformly continuous on L^p ; (c) $\sigma_p(X) = 0$ if and only if X is a constant a.s.; (d) for any σ -algebra $\mathcal{G} \subset \mathcal{F}$, where \mathcal{F} is the original σ -algebra on the underlying probability space for X , we have $\sigma_p(E(X | \mathcal{G})) \leq \sigma_p(X)$; (e) if X and Y have the same distributions, then $\sigma_p(X) = \sigma_p(Y)$.

Definition 2 The *monotone Sharpe ratio* in L^p of a random variable $X \in L^p$ is defined by

$$\mathbb{S}_p(X) = \sup_{Y \leq X} \frac{EY}{\sigma_p(Y)}, \tag{2}$$

where the supremum is over all $Y \in L^p$ such that $Y \leq X$ a.s. For $X = 0$ a.s. we set by definition $\mathbb{S}_p(0) = 0$.

One can easily see that if $p > 1$, then $\mathbb{S}_p(X)$ assumes value in $[0, \infty]$. Indeed, if $EX \leq 0$, then $\mathbb{S}_p(X) = 0$ as it is possible to take $Y \leq X$ with arbitrarily large L^p -deviation keeping EY bounded. On the other hand, if $X \geq 0$ a.s. and $P(X > 0) \neq 0$, then $\mathbb{S}_p(X) = +\infty$ as one can consider $Y_\varepsilon = \varepsilon I(X \geq \varepsilon)$ with $\varepsilon \rightarrow 0$ for which $EY_\varepsilon/\sigma_p(Y_\varepsilon) \rightarrow \infty$.

Thus, the main case of interest will be when $EX > 0$ and $P(X < 0) \neq 0$; then $0 < \mathbb{S}_p(X) < \infty$. For this case, the following theorem provides the representation of \mathbb{S}_p as a solution of some convex optimization problem.

Theorem 1 *Suppose $X \in L^p$ and $E(X) > 0$, $P(X < 0) \neq 0$. Then the following representations of the monotone Sharpe ratio are valid.*

1) For $p \in (1, \infty)$ with q such that $\frac{1}{p} + \frac{1}{q} = 1$:

$$(\mathbb{S}_p(X))^q = \max_{a,b \in \mathbb{R}} \left\{ b - E \left(\frac{q-1}{q^p} |(aX + b)_+ - q|^p + (aX + b)_+ \right) \right\}. \tag{3}$$

2) For $p = 1, 2$:

$$\frac{1}{1 + (\mathbb{S}_p(X, r))^p} = \min_{c \in \mathbb{R}} E(1 - cX)_+^p. \tag{4}$$

The main point about this theorem is that it allows to reduce the problem of computing \mathbb{S}_p as the supremum over the set of random variables to the optimization problem with one or two real parameters and the convex objective function. The latter problem is much easier than the former one, since there exist efficient algorithms of numerical convex optimization. This gives a convenient way to compute $\mathbb{S}_p(X)$ (though only numerically, unlike the standard Sharpe ratio). We'll also see that the representation is useful for establishing some theoretical results about \mathbb{S}_p .

For the proof, we need the following auxiliary lemma.

Lemma 1 *Suppose $X \in L^p$, $p \in [1, \infty)$, and q is such that $\frac{1}{p} + \frac{1}{q} = 1$. Then*

$$\sigma_p(X) = \max\{E(RX) \mid R \in L^q, ER = 0, \|R\|_q \leq 1\}.$$

Proof Suppose $\sigma_p(X) = \|X - c^*\|_p$. By Hölder’s inequality, for any $R \in L^q$ with $ER = 0$ and $\|R\|_q \leq 1$ we have

$$E(RX) = E(R(X - c^*)) \leq \|R\|_q \cdot \|X - c^*\|_p \leq \|X - c^*\|_p.$$

On the other hand, the two inequalities turn into equalities for

$$R^* = \frac{\text{sgn}(X - c^*) \cdot |X - c^*|^{p-1}}{\|X - c^*\|_p^{p-1}}$$

and R^* satisfies the above constraints.

Proof (Proof of Theorem 1) Without loss of generality, assume $EX = 1$. First we’re going to show that \mathbb{S}_p can be represented through the following optimization problem:

$$\mathbb{S}_p(X) = \inf_{R \in L^q} \{ \|R\|_q \mid R \leq 1 \text{ a.s., } ER = 0, E(RX) = 1 \}. \tag{5}$$

In (2), introduce the new variables: $c = (EY)^{-1} \in \mathbb{R}$ and $Z = cY \in L^p$. Then

$$\frac{1}{\mathbb{S}_p(X)} = \inf_{\substack{Z \in L^p \\ c \in \mathbb{R}}} \{ \sigma_p(Z) \mid Z \leq cX, EZ = 1 \}.$$

Consider the dual of the optimization problem in the RHS (see the Appendix for a brief overview of duality methods in optimization). Define the dual objective function $g: L^q_+ \times \mathbb{R} \rightarrow \mathbb{R}$ by

$$g(u, v) = \inf_{\substack{Z \in L^p \\ c \in \mathbb{R}}} \{ \sigma_p(Z) + E(u(Z - cX)) - v(EZ - 1) \}.$$

The dual problem consists in maximizing $g(u, v)$ over all $u \in L^q_+, v \in \mathbb{R}$. We want to show that the strong duality takes place, i.e. that the values of the primal and the dual problems are equal:

$$\frac{1}{\mathbb{S}_p(X)} = \sup_{\substack{u \in L^q_+ \\ v \in \mathbb{R}}} g(u, v).$$

To verify the sufficient condition for the strong duality from Theorem 7, introduce the optimal value function $\phi: L^p \times \mathbb{R} \rightarrow [-\infty, \infty)$

$$\phi(a, b) = \inf_{\substack{Z \in L^p \\ c \in \mathbb{R}}} \{ \sigma_p(Z) \mid Z - cX \leq a, EZ - 1 = b \}$$

(obviously, $(\mathbb{S}_p(X))^{-1} = \phi(0, 0)$). Observe that if a pair (Z_1, c_1) satisfies the constraints in $\phi(a_1, b_1)$ then the pair (Z_2, c_2) with

$$c_2 = c_1 + b_2 - b_1 + E(a_1 - a_2), \quad Z_2 = Z_1 + a_2 - a_1 + (c_2 - c_1)X,$$

satisfies the constraints in $\phi(a_2, b_2)$. Clearly, $\|Z_1 - Z_2\|_p + |c_1 - c_2| = O(\|a_1 - a_2\|_p + |b_1 - b_2|)$, which implies that $\phi(a, b)$ is continuous, so the strong duality holds.

Let us now transform the dual problem. It is obvious that if $E(uX) \neq 0$, then $g(u, v) = -\infty$ (minimize over c). For u such that $E(uX) = 0$, using the dual representation of $\sigma_p(X)$, we can write

$$g(u, v) = \inf_{Z \in L^p} \sup_{R \in \mathcal{R}} E(Z(R + u - v) + v) \quad \text{if } E(uX) = 0,$$

where $\mathcal{R} = \{R \in L^q : ER = 0, \|R\|_q \leq 1\}$ is the dual set for σ_p from Lemma 1. Observe that the set \mathcal{R} is compact in the weak-* topology by the Banach-Alaoglu theorem. Consequently, by the minimax theorem (see Theorem 8), the supremum and infimum can be swapped. Then it is easy to see that $g(u, v) > -\infty$ only if there exists $R \in \mathcal{R}$ such that $R + u - v = 0$ a.s., and in this case $g(u, v) = v$. Therefore, the dual problem can be written as follows:

$$\begin{aligned} \frac{1}{\mathbb{S}_p(X)} &= \sup_{\substack{u \in L^q \\ v \in \mathbb{R}}} \{v \mid u \geq 0 \text{ a.s., } E(uX) = 0, v - u \in \mathcal{R}\} \\ &= \sup_{R \in \mathcal{R}} \{E(RX) \mid R \leq E(RX) \text{ a.s.}\} \\ &= \sup_{R \in L^q} \{E(RX) \mid R \leq E(RX) \text{ a.s., } ER = 0, \|R\|_q \leq 1\}, \end{aligned}$$

where in the second equality we used that if $v - u = R \in \mathcal{R}$, then the second constraint imply that $v = E(RX)$ since it is assumed that $EX = 1$. Now by changing the variable R to $R/E(RX)$ in the right-hand side, we obtain representation (5).

From (5), it is obvious that for $p > 1$

$$(\mathbb{S}_p(X))^q = \inf_{R \in L^q} \{E|R|^q \mid R \leq 1 \text{ a.s., } ER = 0, E(RX) = 1\}. \tag{6}$$

We'll now consider the optimization problem dual to this one. Denote its optimal value function by $\phi : L^q \times \mathbb{R} \times \mathbb{R} \rightarrow \overline{\mathbb{R}}$. It will be more convenient to change the optimization variable R here by $1 - R$ (which clearly doesn't change the value of ϕ), so that

$$\phi(a, b, c) = \inf_{R \in L^q} \{E|1 - R|^q \mid R \geq a \text{ a.s., } ER = 1 + b, E(RX) = c\}.$$

Let us show that ϕ is continuous at zero. Denote by $C(a, b, c) \subset L^q$ the set of $R \in L^q$ satisfying the constraints of the problem. It will be enough to show that if $\|a\|_q, |b|, |c|$ are sufficiently small then for any $R \in C(0, 0, 0)$ there exists $\tilde{R} \in C(a, b, c)$ such that $\|R - \tilde{R}\|_q \leq (\|R\|_q + K)(\|a\|_q + |b| + |c|)$ and vice versa. Here K is some fixed constant.

Since $P(X < 0) \neq 0$, there exists $\xi \in L^\infty$ such that $\xi \geq 0$ a.s. and $E(\xi X) = -1$. If $R \in C(0, 0, 0)$, then one can take the required $\tilde{R} \in C(a, b, c)$ in the form

$$\tilde{R} = \begin{cases} a + \lambda_1 R + \lambda_2 \xi, & \text{if } E(aX) \geq 0, \\ a + \mu_1 R + \mu_2, & \text{if } E(aX) < 0, \end{cases}$$

where the non-negative constants $\lambda_1, \lambda_2, \mu_1, \mu_2$ can be easily found from the constraint $\tilde{R} \in C(a, b, c)$, and it turns out that $\lambda_1, \mu_1 = 1 + O(\|a\|_q + |b| + |c|)$ and $\lambda_2, \mu_2 = O(\|a\|_q + |b| + |c|)$. If $R \in C(a, b, c)$, then take

$$\tilde{R} = \begin{cases} \lambda_1(R - a + \lambda_2 \xi), & \text{if } c \geq E(aX), \\ \mu_1(R - a + \mu_2), & \text{if } c < E(aX), \end{cases}$$

with λ_i, μ_i making $\tilde{R} \in C(0, 0, 0)$.

Thus, the strong duality holds in (6) and we have

$$\mathbb{S}_p(X) = \sup_{\substack{u \in L^q_+ \\ v, w \in \mathbb{R}}} g(u, v, w) \tag{7}$$

with the dual objective function $g : L^q_+ \times \mathbb{R} \times \mathbb{R} \rightarrow \overline{\mathbb{R}}$

$$\begin{aligned} g(u, v, w) &= \inf_{R \in L^q} E(|R|^q + R(u + v + wX) - u - w) \\ &= -E\left(\frac{q-1}{q^p} |u + v + wX|^p + u + w\right), \end{aligned}$$

where the second inequality is obtained by choosing R which minimizes the expression under the expectation for every random outcome.

Observe that for any fixed $v, w \in \mathbb{R}$ the optimal $u^* = u^*(v, w)$ in (7) can be found explicitly: $u^* = (v + wX + q)_-$. Then by straightforward algebraic transformation we obtain (3).

For $p = 2$, from (3) we get

$$(\mathbb{S}_2(X))^2 = \max_{a, b \in \mathbb{R}} \left\{ b - \frac{1}{4} E(aX + b)_+^2 - 1 \right\}$$

It is easy to see that it is enough to maximize only over $b \geq 0$. Maximizing over b and introducing the variable $c = -\frac{a}{b}$, we obtain representation (4) for $p = 2$.

To obtain representation (4) for $p = 1$, let's again consider problem (5). Similarly to (7) (the only change will be to use $\|R\|_q$ instead of $E|R|^q$), we can obtain that

$$\mathbb{S}_1(X) = \sup_{\substack{u \in L^{\infty}_+ \\ v, w \in \mathbb{R}}} g(u, v, w),$$

where now we denote

$$g(u, v, w) = \inf_{R \in L^{\infty}} \{ \|R\|_{\infty} + E(R(u + v + wX) - u) - w \}.$$

Observe that a necessary condition for $g(u, v, w) > -\infty$ is that $E|u + v + wX| \leq 1$: otherwise take $\tilde{R} = c(I(u + v + wX \leq 0) - I(u + v + wX > 0))$ and let $c \rightarrow \infty$. Under this condition we have $g(u, v, w) = -Eu - w$ since from Hölder's inequality $|E((\alpha + v + wX)R)| \leq \|R\|_{\infty}$ and therefore the infimum in g is attained at $R = 0$ a.s. Consequently, the dual problem becomes

$$\mathbb{S}_1(X) = - \inf_{\substack{u \in L^{\infty} \\ v, w \in \mathbb{R}}} \{ Eu + w \mid u \geq 0 \text{ a.s., } E|u + v + wX| \leq 1 \}. \tag{8}$$

Observe that the value of the infimum is non-positive, and so it is enough to restrict the values of w to \mathbb{R}_- only. Let's fix $v \in \mathbb{R}$, $w \in \mathbb{R}_-$ and find the optimal $u^* = u^*(v, w)$. Clearly, whenever $v + wX(\omega) \geq 0$, it's optimal to take $u^*(\omega) = 0$. Whenever $v + wX(\omega) < 0$, we should have $u^*(\omega) \leq |v + wX(\omega)|$, so that $u(\omega) + v + wX(\omega) \leq 0$ (otherwise, the choice $u^*(\omega) = |v + wX(\omega)|$ will be better). Thus for the optimal u^*

$$E|u^* + v + wX| = E|v + wX| - Eu^*.$$

In particular, for the optimal u^* the inequality in the second constraint in (8) should be satisfied as the equality, since otherwise it would be possible to find a smaller u^* . Observe that if $E(v + wX)_+ > 1$, then no $u \in L^{\infty}$ exists which satisfies the constraint of the problem. On the other hand, if $E(v + wX)_+ \leq 1$ then at least one such u exists. Consequently, problem (8) can be rewritten as follows:

$$-\mathbb{S}_1(X) = \inf_{v \in \mathbb{R}, w \in \mathbb{R}_-} \{ E|v + wX| + w - 1 \mid E(v + wX)_+ \leq 1 \}.$$

Clearly, $E|v^* + w^*X| \leq 0$ for the optimal pair (v^*, w^*) , so the constraint should be satisfied as the equality (otherwise multiply both v, w by $1/E|v + wX|_+$, which will decrease the value of the objective function). By a straightforward transformation, we get

$$1 + \mathbb{S}_1(X) = \sup_{v \in \mathbb{R}, w \in \mathbb{R}_-} \{ v \mid E(v + wX)_+ = 1 \}$$

and introducing the new variable $c = w/v$, we obtain representation (3).

2.3 Basic Properties

Theorem 2 For any $p \in [1, \infty)$, the monotone Shape ratio in L^p satisfies the following properties.

- (a) (Quasi-concavity) For any $c \in \mathbb{R}$, the set $\{X \in L^p : \mathbb{S}_p(X) \geq c\}$ is convex.
- (b) (Scaling invariance) $\mathbb{S}_p(\lambda X) = \mathbb{S}_p(X)$ for any real $\lambda > 0$.
- (c) (Law invariance) If X and Y have the same distribution, then $\mathbb{S}_p(X) = \mathbb{S}_p(Y)$.
- (d) (2nd order monotonicity) If X dominates Y in the second stochastic order, then $\mathbb{S}_p(X) \geq \mathbb{S}_p(Y)$.
- (e) (Continuity) $\mathbb{S}_p(X)$ is continuous with respect to L^p -norm at any X such that $EX > 0$ and $P(X < 0) \neq 0$.

Before proving this theorem, let us briefly discuss the properties in the context of the portfolio selection problem.

The quasi-concavity implies that the monotone Sharpe ratio favors portfolio diversification: if $\mathbb{S}_p(X) \geq c$ and $\mathbb{S}_p(Y) \geq c$, then $\mathbb{S}_p(\lambda X + (1 - \lambda)Y) \geq c$ for any $\lambda \in [0, 1]$, where $\lambda X + (1 - \lambda)Y$ can be thought of as diversification between portfolios with returns X and Y . Note that the property of quasi-concavity is weaker than concavity; it's not difficult to provide an example showing that the monotone Sharpe ratio is not concave.

The scaling invariance can be interpreted as that the monotone Sharpe ratio cannot be changed by leveraging a portfolio (in the same way as the standard Sharpe ratio). Namely, suppose $X = R_x$, where $R_x = \langle x, R \rangle$ is the return of portfolio $x \in \mathbb{R}^{n+1}$ (as in Sect. 2.1), $\sum_i x_i = 1$. Consider a leveraged portfolio \tilde{x} with $\tilde{x}_i = \lambda x_i, i \geq 1$ and $\tilde{x}_0 = 1 - \sum \tilde{x}_i$, i.e. a portfolio which is obtained from x by proportionally scaling all the risky positions. Then it's easy to see that $R_{\tilde{x}} = \lambda R_x$, and so $\mathbb{S}_p(R_x) = \mathbb{S}_p(R_{\tilde{x}})$.

Law invariance, obviously, states that we are able to evaluate the performance knowing only the distribution of the return. The interpretation of the continuity property is also clear.

The 2nd order monotonicity means that \mathbb{S}_p is consistent with preferences of risk-averse investors. Recall that it is said that the distribution of a random variable X dominates the distribution of Y in the 2nd stochastic order, which we denote by $X \succcurlyeq Y$, if $EU(X) \geq EU(Y)$ for any increasing concave function U such that $EU(X)$ and $EU(Y)$ exist. Such a function U can be interpreted as a utility function, and then the 2nd order stochastic dominance means that X is preferred to Y by any risk averse investor.

Regarding the properties from Theorem 2, let us also mention the paper [3], which studies performance measures by an axiomatic approach in a fashion similar to the axiomatics of coherent and convex risk measures. The authors define a performance measure (also called an acceptability index) as a functional satisfying certain properties, then investigate implications of those axioms, and show a deep connection with coherent risk measures, as well as provide examples of performance measures. The minimal set of four axioms a performance measure should satisfy

consists of the quasi-concavity, monotonicity, scaling invariance and semicontinuity (in the form of the so-called Fatou property in L^∞ , as the paper [3] considers only functionals on L^∞). In particular, the monotone Sharpe ratio satisfies those axioms and thus provides a new example of a performance measure in the sense of this system of axioms. It also satisfies all the additional natural properties discussed in that paper: the law invariance, the arbitrage consistency ($\mathbb{S}_p(X) = +\infty$ iff $X \geq 0$ a.s. and $P(X > 0) \neq 0$) and the expectation consistency (if $EX < 0$ then $\mathbb{S}_p(X) = 0$, and if $EX > 0$ then $\mathbb{S}_p(X) > 0$; this property is satisfied for $p > 1$).

Proof (Proof of Theorem 2) Quasi-concavity follows from that the L^p -Sharpe ratio $S_p(X) = \frac{EX}{\sigma_p(X)}$ is quasi-concave. Indeed, if $S_p(X) \geq c$ and $S_p(Y) \geq c$, then

$$S_p(\lambda X + (1 - \lambda)Y) \geq \frac{\lambda EX + (1 - \lambda)EY}{\lambda \sigma_p(X) + (1 - \lambda)\sigma_p(Y)} \geq c$$

for any $\lambda \in [0, 1]$. Since \mathbb{S}_p is the maximum of $f_Z(X) = S_p(X - Z)$ over $Z \in L^p_+$, the quasi-concavity is preserved.

The scaling invariance is obvious. Since the expectation and the L^p -deviation are law invariant, in order to prove the law invariance of \mathbb{S}_p , it is enough to show that the supremum in the definition of $\mathbb{S}_p(X)$ can be taken over only $Y \leq X$ which are measurable with respect to the σ -algebra generated by X , or, in other words, $Y = f(X)$ for some measurable function f on \mathbb{R} . But this follows from the fact that if for any $Y \leq X$ one considers $\tilde{Y} = E(Y | X)$, then $\tilde{Y} \leq X$, $E(\tilde{Y}) = EY$ and $\sigma_p(Y) \leq \sigma_p(\tilde{Y})$, hence $S_p(\tilde{Y}) \geq S_p(Y)$.

To prove the 2nd order monotonicity, recall that another characterization of the 2nd order stochastic dominance is as follows: $X_1 \preceq X_2$ if and only if there exist random variables X'_2 and Z (which may be defined on a another probability space) such that $X_2 \stackrel{d}{=} X'_2$, $X_1 \stackrel{d}{=} X'_2 + Z$ and $E(Z | X'_2) \leq 0$. Suppose $X_1 \preceq X_2$. From the law invariance, without loss of generality, we may assume that X_1, X_2, Z are defined on the same probability space. Then for any $Y_1 \leq X_1$ take $Y_2 = E(Y_1 | X_2)$. Clearly, $Y_2 \leq X_2$, $EY_2 = EY_1$ and $\sigma_p(Y_2) \leq \sigma_p(Y_1)$. Hence $\sigma_p(X_1) \leq \sigma_p(X_2)$.

Finally, the continuity of $\mathbb{S}_p(X)$ follows from that the expectation and the L^p -deviation are uniformly continuous.

3 Buffered Probabilities

In the paper [21] was introduced the so-called buffered probability, which is defined as the inverse function of the conditional value at risk (with respect to the risk level). The authors of that and other papers (for example, [7]) argue that in stochastic optimization problems related to minimization of probability of adverse events, the buffered probability can serve as a better optimality criterion compared to the usual probability.

In this section we show that the monotone Sharpe ratio is tightly related to the buffered probability, especially in the cases $p = 1, 2$. In particular, this will provide a connection of the monotone Share ratio with the conditional value at risk. We begin with a review of the conditional value at risk and its generalization to the spaces L^p . Then we give a definition of the buffered probability, which will generalize the one in [12, 21] from L^1 to arbitrary L^p .

3.1 A Review of the Conditional Value at Risk

Let X be a random variable, which describes loss. As opposed to the previous section, now large values are bad, small values are good (negative values are profits). For a moment, to avoid technical difficulties, assume that X has a continuous distribution.

Denote by $Q(X, \lambda)$ the λ -th quantile of the distribution of X , $\lambda \in [0, 1]$, i.e. $Q(X, \lambda)$ is a number $x \in \overline{\mathbb{R}}$, not necessarily uniquely defined, such that $P(X \leq x) = \lambda$. The quantile $Q(X, \lambda)$ is also called the value at risk¹ (VAR) of X at level λ , and it shows that in the worst case of probability $1 - \lambda$, the loss will be at least $Q(X, \lambda)$. This interpretation makes VAR a sort of a measure of risk (in a broad meaning of this term), and it is widely used by practitioners.

However, it is well-known that VAR lacks certain properties that one expects from a measure of risk. One of the most important drawbacks is that it doesn't show what happens with probability less than $1 - \lambda$. For example, an investment strategy which loses \$1 million with probability 1% and \$2 million with probability 0.5% is quite different from a strategy which loses \$1 million and \$10 million with the same probabilities, however they will have the same VAR at the 99% level. Another drawback of VAR is that it's not convex—as a consequence, it may not favor diversification of risk, which leads to concentration of risk (above $1 - \lambda$ level).

The conditional value at risk (CVAR; which is also called the average value at risk, or the expected shortfall, or the superquantile) is considered as an improvement of VAR. Recall that if $X \in L^1$ and has a continuous distribution, then CVAR of X at risk level $\lambda \in [0, 1]$ can be defined as the conditional expectation in its right tail of probability $1 - \lambda$, i.e.

$$\text{CVAR}(X, \lambda) = E(X \mid X > Q(X, \lambda)) \tag{9}$$

We will also use the notation $\mathbb{Q}(X, \lambda) = \text{CVAR}(X, \lambda)$ to emphasize the connection with quantiles.

CVAR provides a basic (and the most used) example of a coherent risk measure. The theory of risk measures, originally introduced in the seminal paper [1], plays

¹Some authors use definitions of VAR and CVAR which are slightly different from the ones used here: for example, take $(-X)$ instead of X , or $1 - \lambda$ instead of λ , etc.

now a prominent role in applications in finance. We are not going to discuss all the benefits of using coherent (and convex) risk measures in optimization problems; a modern review of the main results in this theory can be found, for example, in the monograph [8].

Rockafellar and Uryasev [22] proved that CVAR admits the following representation though the optimization problem

$$\mathbb{Q}(X, \lambda) = \min_{c \in \mathbb{R}} \left(\frac{1}{1 - \lambda} \mathbb{E}(X - c)_+ + c \right). \quad (10)$$

Actually, this formula can be used as a general definition for CVAR, which works in the case of any distribution of X , not necessarily continuous. The importance of this representation is that it provides an efficient method to compute CVAR, which in practical applications often becomes much faster than e.g. using formula (9). It also behaves “nicely” when CVAR is used as a constraint or an optimality criterion in convex optimization problems, for example portfolio selection. Details can be found in [22].

Representation (10) readily suggests how CVAR can be generalized to “put more weight” on the right tail of the distribution of X , which provides a coherent risk measure for the space L^p .

Definition 3 For $X \in L^p$, define the L^p -CVAR at level $\lambda \in [0, 1)$ by

$$\mathbb{Q}_p(X, \lambda) = \min_{c \in \mathbb{R}} \left(\frac{1}{1 - \lambda} \|(X - c)_+\|_p + c \right).$$

The L^p -CVAR was studied, for example, in the papers [2, 9]. In particular, in [9], it was argued that higher values of p may provide better results than the standard CVAR ($p = 1$) in certain portfolio selection problems. For us, the cases $p = 1, 2$ will be the most interesting due the direct connection with the monotone Sharpe ratio, as will be shown in the next section.

It is known that the following dual representation holds for L^p -CVAR, which we will use below: for any $X \in L^p$ and $\lambda \in [0, 1)$

$$\mathbb{Q}_p(X, \lambda) = \sup \{ \mathbb{E}(RX) \mid R \in L^q_+, \|R\|_q \leq (1 - \lambda)^{-1}, \mathbb{E}R = 1 \}, \quad (11)$$

where, as usual, $\frac{1}{p} + \frac{1}{q} = 1$. This result is proved in [2].

3.2 The Definition of Buffered Probability and Its Representation

Consider the function inverse to CVAR in λ , that is for a random variable X and $x \in \mathbb{R}$ define $\mathbb{P}(X, x) = \lambda$ where λ is such that $\mathbb{Q}(X, \lambda) = x$ (some care should

be taken at points of discontinuity, a formal definition is given below). In the papers [12, 13, 21], $\mathbb{P}(X, x)$ was called the “buffered” probability that $X > x$; we explain the rationale behind this name below. At this moment, it may seem that from a purely mathematical point of view such a simple operation as function inversion probably shouldn’t deserve much attention. But that’s not the case if we take applications into account. For this reason, before we give any definitions, let us provide some argumentation why studying $\mathbb{P}(X, x)$ may be useful for applications.

In many practical optimization problems one may want to consider constraints defined in terms of probabilities of adverse event, or to use those probabilities as optimization criteria. For example, an investment fund manager may want to maximize the expected return of her portfolio under the constraint that the probability of a loss more than \$1 million should be less than 1%; or an engineer wants to minimize the construction cost of a structure provided that the tension in its core part can exceed a critical threshold with only a very small probability during its lifetime.

Unfortunately, the probability has all the same drawbacks as the value at risk, which were mentioned above: it’s not necessarily convex, continuous and doesn’t provide information about how wrong things can go if an adverse event indeed happens. For those reasons, CVAR may be a better risk measure, which allows to avoid some of the problems. For example, if using CVAR, the above investor can reformulate her problem as maximization of the expected return given that the average loss in the worst 1% of cases doesn’t exceed \$1 million. However, such a setting of the problem may be inconvenient, as CVAR “speaks” in terms of quantiles, but one may need the answer in terms of probabilities. For example, \$1 million may be value of liquid assets of the fund which can be quickly and easily sold to cover a loss; so the manager must ensure that the loss doesn’t exceed this amount. But it is not clear how she can use the information about the average loss which CVAR provides. A similar problem arises in the example with an engineer.

In [21], Rockafellar and Royset proposed the idea that the inverse of CVAR may be appropriate for such cases: since quantiles and probabilities are mutually inverse, and CVAR is a better alternative to quantiles, then one can expect that the inverse of CVAR, the buffered probability, could be a better alternative to probability. Here, we follow this idea.

Note that, in theory, it is possible to invert CVAR as a function in λ , but, in practice, computational difficulty may be a serious problem for doing that: it may take too much time to compute CVAR for a complex system even for one fixed level of risk λ , so inversion, which requires such a computation for several λ , may be not feasible (and this is often the case in complex engineering or financial models). Therefore, we would like to be able to work directly with buffered probabilities, and have an efficient method to compute them. We’ll see that the representation given below turns out to give more than just an efficient method of computation. In particular, in view of Sect. 2, it will show a connection with the monotone Sharpe ratio, a result which is by no means obvious.

The following simple lemma will be needed to show that it is possible to invert CVAR.

Lemma 2 For $X \in L^p$, $p \in [1, \infty)$, the function $f(\lambda) = \mathbb{Q}_p(X, \lambda)$ defined for $\lambda \in [0, 1)$ has the following properties:

1. $f(0) = EX$;
2. $f(\lambda)$ is continuous and non-decreasing;
3. $f(\lambda)$ is strictly increasing on the set $\{\lambda : f(\lambda) < \text{ess sup } X\}$;
4. if $P := P(X = \text{ess sup } X) > 0$, then $f(\lambda) = \text{ess sup } X$ for $\lambda \in [1 - P^{1/p}, 1)$.

Proof The first property obviously follows from the dual representation, and the second one can be easily obtained from the definition. To prove the third property, observe that if $\mathbb{Q}_p(X, \lambda) < \text{ess sup } X$, then the minimum in the definition is attained at some $c^* < \text{ess sup } X$. So, for any $\lambda' < \lambda$ we have $\mathbb{Q}_p(X, \lambda') \leq \frac{1}{1-\lambda'} \|(X - c^*)_+\|_p + c^* < \mathbb{Q}_p(X, \lambda)$ using that $\|(X - c^*)_+\|_p > 0$.

Finally, the fourth property follows from that if $P > 0$, and, in particular, $\text{ess sup } X < \infty$, then $\mathbb{Q}_p(X, \lambda) \leq \text{ess sup } X$ for any $\lambda \in [0, 1)$, as one can take $c = \text{ess sup } X$ in the definition. On the other hand, for $\lambda_0 = 1 - P^{\frac{1}{p}}$ we have that $R = P^{-1}I(X = \text{ess sup } X)$ satisfies the constraint in the dual representation and $E(RX) = \text{ess sup } X$. Hence $\mathbb{Q}(X, \lambda_0) = \text{ess sup } X$, and then $\mathbb{Q}_p(X, \lambda) = \text{ess sup } X$ for any $\lambda \geq \lambda_0$ by the monotonicity.

Definition 4 For $X \in L^p$, $p \in [1, \infty)$, and $x \in \mathbb{R}$, set

$$\mathbb{P}_p(X, x) = \begin{cases} 0, & \text{if } x > \text{ess sup } X, \\ (P(X = \text{sup } X))^{\frac{1}{p}}, & \text{if } x = \text{ess sup } X, \\ 1 - \mathbb{Q}_p^{-1}(X, x), & \text{if } EX < x < \text{ess sup } X, \\ 1, & \text{if } x \leq EX. \end{cases}$$

The “main” case in this definition is the third one. In particular, one can see that for a random variable X which has a distribution with a support unbounded from above, the first and the second cases do not realize. Figure 1 schematically shows the relation between the quantile function, the CVAR, the probability distribution function, and the buffered probability. In particular, it is easy to see that always $\mathbb{P}_p(X, x) \geq P(X > x)$. According to the terminology of [21], the difference between these two quantities is a “safety buffer”, hence the name buffered probability.

Theorem 3 For any $X \in L^p$

$$\mathbb{P}_p(X, x) = \min_{c \geq 0} \|(c(X - x) + 1)_+\|_p. \tag{12}$$

Proof For the case $p = 1$ this result was proved in [12]. Here, we follow the same idea, but for general $p \in [1, \infty)$. Without loss of generality, we can assume $x = 0$, otherwise consider $X - x$ instead of X .

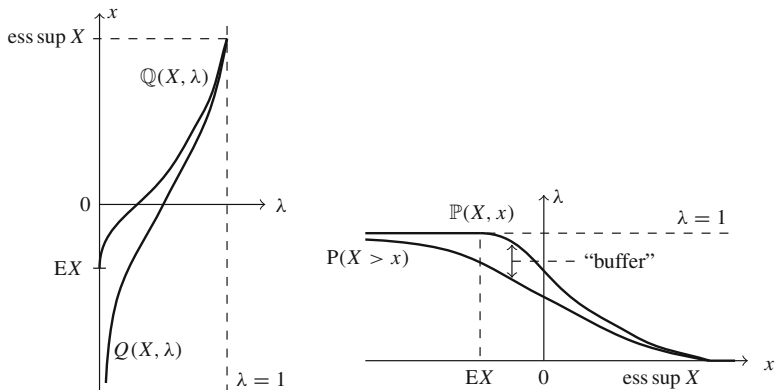


Fig. 1 Left: quantile and distribution functions. Right: complementary probability distribution function and buffered probability $\mathbb{P}(X, x)$. In this example, $\text{ess sup } X < \infty$, but $\mathbb{P}(X = \text{ess sup } X) = 0$, so $\mathbb{P}(X, x)$ is continuous everywhere

Case 1: $EX < 0, \text{ess sup } X > 0$. By Lemma 2 and the definition of \mathbb{Q}_p we have

$$\begin{aligned} \mathbb{P}_p(X, 0) &= \min\{\lambda \in (0, 1) \mid \mathbb{Q}_p(X, 1 - \lambda) = 0\} \\ &= \min_{\lambda \in (0, 1)} \{\lambda \mid \min_{c \in \mathbb{R}} (\frac{1}{\lambda} \|(X + c)_+\|_p - c) = 0\} \\ &= \min_{\substack{\lambda \in (0, 1) \\ c \in \mathbb{R}}} \{\lambda \mid \|(X + c)_+\|_p = \lambda c\}. \end{aligned}$$

Observe that the minimum here can be computed only over $c > 0$ (since for $c \leq 0$ the constraint is obviously not satisfied). Then dividing the both parts of the equality in the constraint by c we get

$$\mathbb{P}_p(X, 0) = \min_{c > 0} \|(c^{-1}X + 1)_+\|_p,$$

which is obviously equivalent to (12).

Case 2: $EX \geq 0$. We need to show that $\min_{c \geq 0} \|(cX + 1)_+\|_p = 1$. This clearly follows from that for any $c \geq 0$ we have $\min_{c \geq 0} \|(cX + 1)_+\|_p \geq \min_{c \geq 0} E(cX + 1) = 1$.

Case 3: $\text{ess sup } X = 0$. Now $\|(cX + 1)_+\|_p \geq \mathbb{P}(X = 0)^{1/p}$ for any $c \geq 0$, while $\|(cX + 1)_+\|_p \rightarrow \mathbb{P}(X = 0)^{1/p}$ as $c \rightarrow +\infty$. Hence $\min_{c \geq 0} \|(cX + 1)_+\|_p = \mathbb{P}(X = 0)^{1/p}$ as claimed.

Case 4: $\text{ess sup } X < 0$. Similarly, $\|(cX + 1)_+\|_p \rightarrow 0$ as $c \rightarrow +\infty$.

From formula (12), one can easily see the connection between the monotone Sharpe ratio and the buffered probability for $p = 1, 2$: for any $X \in L^p$

$$\frac{1}{1 + (\mathbb{S}_p(X))^p} = (\mathbb{P}_p(-X, 0))^p.$$

In particular, if X is as the return of a portfolio, then a portfolio selection problem where one wants to maximize the monotone Sharpe ratio of the portfolio return becomes equivalent to the minimization of the buffered probability that $(-X)$ exceeds 0, i.e. the buffered probability of loss. This is a nice (and somewhat unexpected) connection between the classical portfolio theory and modern developments in risk evaluation!

One can ask a question whether a similar relation between \mathbb{P}_p and \mathbb{S}_p holds for other values of p . Unfortunately, in general, there seems to be no simple formula connecting them. It can be shown that they can be represented as the following optimization problems:

$$\begin{aligned} \mathbb{S}_p(X) &= \min_{R \in L^q_+} \{\|R - 1\|_q \mid ER = 1, E(RX) = 1\}, \\ \mathbb{P}_p(X, 0) &= \min_{R \in L^q_+} \{\|R\|_q \mid ER = 1, E(RX) = 1\}, \end{aligned}$$

which have the same constraint sets but different objective functions. The first formula here easily follows from (5), the second one can be obtained using the dual representation of CVAR (11).

3.3 Properties

In this section we investigate some basic properties of $\mathbb{P}_p(X, x)$ both in X and x , and discuss its usage in portfolio selection problem. One of the main points of this section is that buffered probabilities (of loss) can be used as optimality criteria, similarly to monotone Sharpe ratios (and in the cases $p \neq 1, 2$ they are more convenient due to a simpler representation).

Theorem 4 *Suppose $X \in L^p$, $x \in \mathbb{R}$ and $p \in [1, \infty)$. Then $\mathbb{P}_p(X, x)$ has the following properties.*

1. *The function $x \mapsto \mathbb{P}_p(X, x)$ is continuous and strictly decreasing on $[EX, \text{ess sup}(X))$, and non-increasing on the whole \mathbb{R} .*
2. *The function $X \mapsto \mathbb{P}_p(X, x)$ is quasi-convex, law invariant, 2nd order monotone, continuous with respect to the L^p -norm, and concave with respect to mixtures of distributions.*
3. *The function $p \mapsto \mathbb{P}_p(X, x)$ is non-decreasing in p .*

For $p = 1$, similar results can be found in [12]; the proofs are similar as well (except property 3, but it obviously follows from the Lyapunov inequality), so we do not provide them here.

Regarding the second property note that despite $\mathbb{P}_p(X, x)$ is quasi-convex in X , it's not convex in X as the following simple example shows: consider $X \equiv 2$ and $Y \equiv -1$; then $\mathbb{P}((X + Y)/2, 0) = 1 \not\leq \frac{1}{2} = \frac{1}{2}\mathbb{P}(X, 0) + \frac{1}{2}\mathbb{P}(Y, 0)$.

Also recall that the mixture of two distributions on \mathbb{R} specified by their distribution functions $F_1(x)$ and $F_2(x)$ is defined as the distribution $F(x) = \lambda F_1(x) + (1 - \lambda)F_2(x)$ for any fixed $\lambda \in [0, 1]$. We write $X \stackrel{d}{=} \lambda X_1 \oplus (1 - \lambda)X_2$ if the distribution of a random variable X is the mixture of the distributions of X_1 and X_2 . If ξ is a random variable taking values 1, 2 with probabilities $\lambda, 1 - \lambda$ and independent of X_1, X_2 , then clearly $X \stackrel{d}{=} X_\xi$. Concavity of $\mathbb{P}_p(x, x)$ with respect to mixtures of distributions means that $\mathbb{P}_p(X, x) \geq \lambda \mathbb{P}_p(X_1, x) + (1 - \lambda)\mathbb{P}_p(X_2, x)$.

Now let's look in more details on how a simple portfolio selection problem can be formulated with \mathbb{P}_p . Assume the same setting as in Sect. 2.1: R is a $(n + 1)$ -dimensional vector of asset returns, the first asset is riskless with the rate of return r , and the other n assets are risky with random return in L^p . Let $R_x = (x, R)$ denote the return of a portfolio $x \in \mathbb{R}^{n+1}$, and $\delta > 0$ be a fixed number, a required expected return premium. Consider the following optimization problem:

$$\begin{aligned} &\text{minimize} && \mathbb{P}_p(r - R_x, 0) \text{ over } x \in \mathbb{R}^{n+1} \\ &\text{subject to} && \mathbb{E}(R_x - r) = \delta, \\ &&& \sum_i x_i = 1. \end{aligned} \tag{13}$$

In other words, an investor wants to minimize the buffered probability that the return of her portfolio will be less than the riskless return subject to the constraint on the expected return. Denote the vector of adjusted risky returns $\bar{R} = (R_1 - r, \dots, R_n - r)$, and the risky part of the portfolio $\bar{x} = (x_1, \dots, x_n)$. Using the representation of \mathbb{P}_p , the problem becomes

$$\begin{aligned} &\text{minimize} && \mathbb{E}(1 - \langle \bar{x}, \bar{R} \rangle_+^p) \text{ over } \bar{x} \in \mathbb{R}^n \\ &\text{subject to} && \mathbb{E}\langle \bar{x}, \bar{R} \rangle \geq 0. \end{aligned} \tag{14}$$

If we find a solution \bar{x}^* of this problem, then the optimal portfolio in problem (13) can be found as follows:

$$x_i^* = \frac{\delta \bar{x}_i^*}{\mathbb{E}\langle \bar{x}^*, \bar{R} \rangle}, \quad i = 1, \dots, n, \quad x_0^* = 1 - \sum_{i=1}^n x_i^*.$$

Moreover, observe that the constraint $\mathbb{E}\langle \bar{x}, \bar{R} \rangle \geq 0$ can be removed in (14) since the value of the objective function is not less than 1 in the case if $\mathbb{E}\langle \bar{x}, \bar{R} \rangle < 0$, which is not optimal. Thus, (14) becomes an unconstrained problem.

4 Dynamic Problems

This section illustrates how the developed theory can be used to give new elegant solutions of dynamic portfolio selection problems when an investor can continuously trade in the market. The results we obtain are not entirely new, but their proofs are considerably shorter and simpler than in the literature.

4.1 A Continuous-Time Market Model and Two Investment Problems

Suppose there are two assets traded in the market: a riskless asset with price B_t and a risky asset with price S_t at time $t \in [0, \infty)$. The time runs continuously. Without loss of generality, we assume $B_t \equiv 1$. The price of the risky asset is modeled by a geometric Brownian motion with constant drift μ and volatility σ , i.e.

$$S_t = S_0 \exp\left(\sigma W_t + \left(\mu - \frac{\sigma^2}{2}t\right)\right), \quad t \geq 0,$$

where W_t is a Brownian motion (Wiener process). Without loss of generality, $S_0 = 1$. It is well-known that the process S_t is the unique strong solution of the stochastic differential equation (SDE)

$$dS_t = S_t(\mu dt + \sigma dW_t).$$

We consider the following two problems of choosing an optimal investment strategy in this market model.

Problem 1 Suppose a trader can manage her portfolio dynamically on a time horizon $[0, T]$. A trading strategy is identified with a scalar control process u_t , which is equal to the amount of money invested in the risky asset at time t . The amount of money v_t is invested in the riskless asset. The value $X_t^u = u_t + v_t$ of the portfolio with the starting value $X_0^u = x_0$ satisfies the controlled SDE

$$dX_t^u = u_t(\mu dt + \sigma dW_t), \quad X_0^u = x_0. \quad (15)$$

This equation is well-known and it expresses the assumption that the trading strategy is self-financing, i.e. it has no external inflows or outflows of capital. Note that v_t doesn't appear in the equation since it can be uniquely recovered as $v_t = X_t^u - u_t$.

To have X^u correctly defined, we'll assume that u_t is predictable with respect to the filtration generated by W_t and $E \int_0^T u_t^2 dt < \infty$. We'll also need to impose the

following mild technical assumption:

$$E \exp\left(\frac{\sigma^2 p^2}{2} \int_0^T \frac{u_t^2}{(1 - X_t^u)^2} dt\right) < \infty. \tag{16}$$

The class of all the processes u_t satisfying these assumptions will be denoted by \mathcal{U} . Actually, it can be shown that (16) can be removed without changing the class of optimal strategies in the problem formulated below, but to keep the presentation simple, we will require it to hold.

The problem consists in minimizing the buffered probability of loss by time T . So the goal of the trader is to solve the following control problem with some fixed $p \in (1, \infty)$:

$$V_1 = \inf_{u \in \mathcal{U}} \mathbb{P}_p(x_0 - X_T^u, 0). \tag{17}$$

For $p = 2$ this problem is equivalent to the problem of maximization of the monotone Sharpe ratio $\mathbb{S}_p(X_T^u - x_0)$. Moreover, we'll also show that the same solution is obtained in the problem of maximization of the standard Sharpe ratio, $S(X_T^u - x_0)$. Note that we don't consider the case $p = 1$.

From (15) and (17), it is clear that without loss of generality we can (and will) assume $x_0 = 0$. It is also clear that there is no unique solution of problem (17): if some u^* minimizes $\mathbb{P}_p(x_0 - X_T^u, 0)$ then so does any $u_t = cu_t^*$ with a constant $c > 0$. Hence, additional constraints have to be imposed if one wants to have a unique solution, for example a constraint on the expected return like $EX_T^u = x_0 + \delta$. This is similar to the standard Markowitz portfolio selection problem, as discussed in Sect. 2.1.

Problem 2 Suppose at time $t = 0$ a trader holds one unit of the risky asset with the starting price $S_0 = 1$ and wants to sell it better than some goal price $x \geq 1$. The asset is indivisible (e.g. a house) and can be sold only at once.

A selling strategy is identified with a Markov time of the process S_t . Recall that a random variable τ with values in $[0, \infty]$ is called a Markov time if the random event $\{\tau \leq t\}$ is in the σ -algebra $\sigma(S_r; r \leq t)$ for any $t \geq 0$. The notion of a stopping time reflects the idea that no information about the prices in the future can be used at the moment when the trader decides to sell the asset. The random event $\{\tau = \infty\}$ is interpreted as the situation when the asset is never sold, and we set $S_\infty := 0$. We'll see that the optimal strategy in the problem we formulate below will not sell the asset with positive probability.

Let \mathcal{M} we denote the class of all Markov times of the process S_t . We consider the following optimal stopping problem for $p \in (1, \infty)$:

$$V_2 = \inf_{\tau \in \mathcal{M}} \mathbb{P}_p(x - S_\tau, 0), \tag{18}$$

i.e. minimization of the buffered probability to sell worse than for the goal price x . Similarly to Problem 1, in the case $p = 2$, it'll be shown that this problem is equivalent to maximization of the monotone Sharpe ratio $\mathbb{S}_2(S_\tau - x)$, and the optimal strategy also maximizes the standard Sharpe ratio.

4.2 A Brief Literature Review

Perhaps, the most interesting thing to notice about the two problems is that they are “not standard” from the point of view of the stochastic control theory for diffusion processes and Brownian motion. Namely, they don't directly reduce to solutions of some PDEs, which for “standard” problems is possible via the Hamilton–Jacobi–Bellman equation. In Problems 1 and 2 (and also in the related problems of maximization of the standard Sharpe ratio), the HJB equation cannot be written because the objective function is not in the form of the expectation of some functional of the controlled process, i.e. not $EF(X_r^u; r \leq T)$ or $EF(S_r; r \leq \tau)$. Hence, another approach is needed to solve them.

Dynamic problems of maximization of the standard Sharpe ratio and related problems with mean-variance optimality criteria have been studied in the literature by several authors. We just briefly mention several of them.

Richardson [19] was, probably, the first who solved a dynamic portfolio selection problem under a mean-variance criterion (the earlier paper of White [26] can also be mentioned); he used “martingale” methods. Li and Ng [11] studied a multi-asset mean-variance selection problem, which they solved through auxiliary control problems in the standard setting. A similar approach was also used in the recent papers by Pedersen and Peskir [15, 16]. The first of them provides a solution for the optimal selling problem (an analogue of our Problem 2), the second paper solves the portfolio selection problem (our Problem 1). There are other results in the literature, a comprehensive overview can be found in the above-mentioned papers by Pedersen and Peskir, and also in the paper [6].

It should be also mentioned that a large number of papers study the so-called problem of time inconsistency of the mean-variance and similar optimality criteria, which roughly means that at a time $t > t_0$ it turns out to be not optimal to follow the strategy, which was optimal at time t_0 . Such a contradiction doesn't happen in standard control problems for Markov processes, where the Bellman principle can be applied, but it is quite typical for non-standard problems. Several approaches to redefine the notion of an optimal strategy that would take into account time inconsistency are known: see, for example, the already mentioned papers [6, 15, 16] and the references therein. We will not deal with the issue of time inconsistency (our solutions are time inconsistent).

Compared to the results in the literature, the solutions of Problems 1 and 2 in the case $p = 2$ readily follows from earlier results (e.g. from [15, 16, 19]); the other cases can be also studied by previously known methods. Nevertheless, the value of this paper is in the new approach to solve them through the monotone Sharpe

ratio and buffered probabilities. This approach seems to be simpler than previous ones (the reader can observe how short the solutions presented below compared to [15, 16]) and promising for more general settings.

4.3 Solution of Problem 1

Theorem 5 *The class of optimal control strategies in Problem 1 is given by*

$$u_t^c = \frac{\mu}{\sigma^2(p - 1)}(c - X_t^{u^c}),$$

where $c > 0$ can be an arbitrary constant. The process $Y_t^{u^c} = c - X_t^{u^c}$ is a geometric Brownian motion satisfying the SDE

$$\frac{dY_t^{u^c}}{Y_t^{u^c}} = -\frac{\mu^2}{\sigma^2(p - 1)}dt - \frac{\mu}{\sigma(p - 1)}dW_t, \quad Y_0^{u^c} = c.$$

Proof Assuming $x_0 = 0$, from the representation of $\mathbb{P}_p(X, x)$ we have

$$V_1 = \min_{c \geq 0} \min_{u \in \mathcal{U}} \|(1 - cX_T^u)_+\|_p = \min_{u \in \mathcal{U}} \|(1 - X_T^u)_+\|_p, \tag{19}$$

where in the second equality we used that the constant c can be included in the control, since $cX^u = X^{cu}$. Denote $\tilde{X}_t^u = 1 - X_t^u$, so that the controlled process \tilde{X}^u satisfies the equation

$$d\tilde{X}_t^u = -\mu u_t dt - \sigma u_t dW_t, \quad \tilde{X}_0^u = 1.$$

Then

$$V_1^p = \min_{u \in \mathcal{U}} E|\tilde{X}_T^u|^p, \tag{20}$$

where $(\cdot)_+$ from (19) was removed since it is obvious that as soon as \tilde{X}_t^u reaches zero, it is optimal to choose $u \equiv 0$ afterwards, so the process stays at zero.

Let $v_t = v_t(u) = -u_t/\tilde{X}_t^u$. Then for any $u \in \mathcal{U}$ we have

$$E|\tilde{X}_T^u|^p = E\left\{Z_T \exp\left(\int_0^T (\mu p v_s + \frac{1}{2}\sigma^2(p^2 - p)v_s^2)ds\right)\right\},$$

where Z is the stochastic exponent process $Z = \mathcal{E}(\sigma p v)$. From Novikov’s condition, which holds due to (16), Z_t is a martingale and $E Z_T = 1$. By introducing the new measure Q on the σ -algebra $\mathcal{F}_T = \sigma(W_t, t \leq T)$ with the density

$dQ = Z_T dP$ we obtain

$$E|\tilde{X}_T^u|^p = E^Q \left\{ \exp\left(\int_0^T (\mu p v_s + \frac{1}{2}\sigma^2(p^2 - p)v_s^2) ds\right) \right\}.$$

Clearly, this expression can be minimized by minimizing the integrand for each t , i.e. by

$$v_t^* = -\frac{\mu}{\sigma^2(p - 1)} \text{ for all } t \in [0, T].$$

Obviously, it satisfies condition (16), so the corresponding control process

$$u_t^* = \frac{\mu}{\sigma^2(p - 1)} \tilde{X}_t^u = \frac{\mu}{\sigma^2(p - 1)} (1 - X_t^u)$$

is optimal in problem (20). Consequently, any control process $u_t^c = cu_t^*$, $c > 0$, will be optimal in (17). Since $X_t^{u^c} = cX_t^{u^*}$, we obtain the first claim of the theorem. The representation for $Y_t^{u^c}$ follows from straightforward computations.

Corollary 1 *Let $u^* = \frac{\mu}{\sigma^2}(c - X_t^u)$, with some $c > 0$, be an optimal control strategy in problem (17) for $p = 2$. Then the standard Sharpe ratio of $X_T^{u^*}$ is equal to its monotone Sharpe ratio, $S(X_T^{u^*}) = \mathbb{S}_2(S_T^{u^*})$.*

In particular, u^ also maximizes the standard Sharpe ratio of the return X_T^u , i.e. $S(X_T^u) \leq S(X_T^{u^*})$ for any $u \in \mathcal{U}$.*

Proof Suppose there is $Y \in L^2$ such that $Y \leq X_T^{u^*}$ and $S(Y) > S(X_T^{u^*})$. It is well-known that the market model we consider is no-arbitrage and complete. This implies that there exists $y_0 < 0$ and a control u_t such that $X_0^u = y_0$ and $X_T^u = Y$. The initial capital y_0 is negative, because otherwise an arbitrage opportunity can be constructed. But then the capital process $\tilde{X}_t = X_t^u - y_0$ would have a higher Sharpe ratio than Y and hence a higher monotone Sharpe ratio than $X_T^{u^*}$. A contradiction. This proves the first claim of the corollary, and the second one obviously follows from it.

4.4 Solution of Problem 2

We'll assume that $x \geq 1$, $\mu \in \mathbb{R}$, $\sigma > 0$, $p > 1$ are fixed throughout and use the following auxiliary notation:

$$\gamma = \frac{2\mu}{\sigma^2}, \quad C(b) = \left(\frac{b}{1 + \frac{x}{b-x}(1 - b^{1-\gamma})^{\frac{1}{p-1}}} - x \right)^{-1} \text{ for } b \in [x, \infty).$$

Theorem 6 *The optimal selling time τ^* in problem (18) is as follows.*

1. *If $\mu \leq 0$, then τ^* can be any Markov time: $\mathbb{P}_p(x - S_\tau, 0) = 1$ for any $\tau \in \mathcal{M}$.*
2. *If $\mu \geq \frac{\sigma^2}{2}$, then S_t reaches any level $x' > x$ with probability 1 and any stopping time of the form $\tau^* = \inf\{t \geq 0 : S_t = x'\}$ is optimal.*
3. *If $0 < \mu < \frac{\sigma^2}{2}$, then the optimal stopping time is*

$$\tau^* = \inf\{t \geq 0 : S_t \geq b^*\},$$

where $b^* \in [x, \infty)$ is the point of minimum of the function

$$f(b) = ((1 + C(b)(x - b))^p b^{\gamma-1} + (1 + C(b)x)^p (1 - b^{\gamma-1})), \quad b \in [x, \infty),$$

and we set $\tau^* = +\infty$ on the random event $\{S_t < b \text{ for all } t \geq 0\}$.

Observe that if $0 < \mu < \frac{\sigma^2}{2}$, i.e. $\gamma \in (0, 1)$, then the function $f(b)$ attains its minimum on $[x, \infty)$, since it is continuous with the limit values $f(x) = f(\infty) = 1$.

Proof From the representation for \mathbb{P}_p we have

$$V_2^p = \inf_{c \geq 0} \inf_{\tau \in \mathcal{M}} E|(1 + c(x - S_\tau))_+|^p.$$

Let $Y_t^c = 1 + c(x - S_t)$. Observe that if $\mu \leq 0$, then Y_t^c is a submartingale for any $c \geq 0$, and so by Jensen's inequality $|(Y_t)_+|^p$ is a submartingale as well. Hence for any $\tau \in \mathcal{M}$ we have $E|(1 + c(x - S_\tau))_+|^p \geq 1$, and then $V_2 = 1$.

If $\mu \geq \frac{\sigma^2}{2}$, then from the explicit representation $S_t = \exp(\sigma W_t + (\mu - \frac{\sigma^2}{2})t)$ one can see that S_t reaches any level $x' \geq 1$ with probability 1 (as the Brownian motion W_t with non-negative drift does so). Then for any $x' > x$ we have $\mathbb{P}_p(x - S_{\tau_{x'}}, 0) = 0$, where $\tau_{x'}$ is the first moment of reaching x' .

In the case $\mu \in (0, \frac{\sigma^2}{2})$, for any $c \geq 0$, consider the optimal stopping problem

$$V_{2,c} = \inf_{\tau \in \mathcal{M}} E|(1 + c(x - S_\tau))_+|^p.$$

This is an optimal stopping problem for a Markov process S_t . From the general theory (see e.g. [17]) it is well known that the optimal stopping time here is of the threshold type:

$$\tau_c^* = \inf\{t \geq 0 : S_t \geq b_c\},$$

where $b_c \in [x, x + \frac{1}{c}]$ is some optimal level, which has to be found. Then the distribution of $S_{\tau_c^*}$ is binomial: it assumes only two values b_c and 0 with probabilities p_c and $1 - p_c$, where $p_c = b_c^{\gamma-1}$ as can be easily found from the general formulas

for boundary crossing probabilities for a Brownian motion with drift. Consequently,

$$V_2^p = \inf_{b \geq x} \inf_{c \leq \frac{1}{(b-x)}} \left((1 + c(x - b))^p b^{\gamma-1} + (1 + cx)^p (1 - b^{\gamma-1}) \right).$$

It is straightforward to find that for any $b \geq x$ the optimal $c^*(b)$ is given by $c^*(b) = C(b)$, which proves the claim of the theorem.

Corollary 2 *Assume $\mu \in (0, \frac{\sigma^2}{2})$ and $p = 2$. Let τ^* denote the optimal stopping time from Theorem 6. Then the standard Sharpe ratio of $S_{\tau^*} - x$ is equal to its monotone Sharpe ratio, $S(S_{\tau^*} - x) = \mathbb{S}_2(S_{\tau^*} - x)$. In particular, τ^* also maximizes the standard Sharpe ratio of $S_{\tau} - x$, i.e. $S(S_{\tau} - x) \leq S(S_{\tau^*} - x)$ for any $\tau \in \mathcal{M}$.*

Moreover, in this case the optimal threshold b^ can be found as the point of maximum of the function*

$$g(b) = \frac{b^{\gamma} - x}{b^{\frac{\gamma+1}{2}} (1 - b^{\gamma-1})^{\frac{1}{2}}}.$$

Proof Suppose $Y \leq S_{\tau^*} - x$. As shown above, it is enough to consider only Y which are measurable with respect to the σ -algebra generated by the random variable S_{τ^*} . Since S_{τ^*} has a binomial distribution, then Y should also have a binomial distribution, assuming values $y_1 \leq b^* - x$ and $y_2 \leq -x$ with the same probabilities $(b^*)^{\gamma-1}$ and $1 - (b^*)^{\gamma-1}$ as S_{τ^*} assumes the values b^* and 0. Using this, it is now not difficult to see that $S(Y) \leq S(S_{\tau^*} - x)$, which proves the first claim.

The second claim follows from that for any stopping time of the form $\tau_b = \{t \geq 0 : S_t = b\}$, $b \in [x, \infty)$ we have $S(S_{\tau_b} - x) = g(b)$.

Appendix

This appendix just reminds some facts from convex optimization and related results which were used in the paper.

Duality in Optimization

Let \mathcal{Z} be a topological vector space and $f(z)$ a real-valued function on \mathcal{Z} . Consider the optimization problem

$$\text{minimize } f(z) \text{ over } z \in \mathcal{Z}. \tag{21}$$

A powerful method to analyze such an optimization problem consists in considering its dual problem. To formulate it, suppose that $f(z)$ can be represented in the form

$f(z) = F(z, 0)$ for all $z \in \mathcal{Z}$, where $F(z, a): \mathcal{Z} \times \mathcal{A} \rightarrow \mathbb{R}$ is some function, and \mathcal{A} is another topological vector space (a convenient choice of F and \mathcal{A} plays an important role).

Let \mathcal{A}^* denote the topological dual of \mathcal{A} . Define the Lagrangian $L: \mathcal{Z} \times \mathcal{A}^* \rightarrow \overline{\mathbb{R}}$ and the dual objective function $g: \mathcal{A}^* \rightarrow \overline{\mathbb{R}}$ by

$$L(z, u) = \inf_{a \in \mathcal{A}} \{F(z, a) + \langle a, u \rangle\}, \quad g(u) = \inf_{z \in \mathcal{Z}} L(z, u).$$

Then the dual problem is formulated as the optimization problem

$$\text{maximize } g(u) \text{ over } u \in \mathcal{A}^*.$$

If we denote by V_P and V_D the optimal values of the primal and dual problems respectively (i.e. the infimum of $f(z)$ and the supremum of $g(u)$ respectively), then it is easy to see that $V_P \geq V_D$ always.

We are generally interested in the case when the strong duality takes place, i.e. $V_P = V_D$, or, explicitly,

$$\min_{z \in \mathcal{Z}} f(z) = \max_{u \in \mathcal{A}^*} g(u). \tag{22}$$

Introduce the optimal value function $\phi(a) = \inf_{z \in \mathcal{Z}} F(z, a)$. The following theorem provides a sufficient condition for the strong duality (22) (see Theorem 7 in [20]).

Theorem 7 *Suppose F is convex in (z, a) and $\phi(0) = \liminf_{a \rightarrow 0} \phi(a)$. Then (22) holds.*

Let us consider a particular case of problem (21) which includes constraints in the form of equalities and inequalities. Assume that $\mathcal{Z} = L^p$ for some $p \in [1, \infty)$ and two functions $h_i: L^p \rightarrow L^{r_i}(\mathbb{R}^{n_i}), i = 1, 2$ are given (the spaces L^p and L^{r_i} are not necessarily defined on the same probability space). Consider the problem

$$\begin{aligned} &\text{minimize } f(z) \text{ over } z \in L^p \\ &\text{subject to } g(z) \leq 0 \text{ a.s.} \\ &h(z) = 0 \text{ a.s.} \end{aligned}$$

This problem can be formulated as a particular case of the above abstract setting by defining

$$F(z, a_1, a_2) = \begin{cases} f(z), & \text{if } g(z) \leq a_1 \text{ and } h(z) = a_2 \text{ a.s.,} \\ +\infty, & \text{otherwise.} \end{cases}$$

The Lagrangian of this problem is

$$\begin{aligned} L(z, u_1, u_2) &= \inf_{a_1, a_2} \{F(z, a_1, a_2) + \langle a_1, u_1 \rangle + \langle a_2, u_2 \rangle\} \\ &= \begin{cases} f(z) + \langle g(z), u_1 \rangle + \langle h(z), u_2 \rangle, & \text{if } u_1 \geq 0 \text{ a.s.}, \\ -\infty, & \text{otherwise,} \end{cases} \end{aligned}$$

where we denote $\langle a, u \rangle = E(\sum_i a_i u_i)$.

So the dual objective function

$$g(u, v) = \inf_{z \in L^p} \{f(z) + \langle g(z), u \rangle + \langle h(z), v \rangle\} \quad \text{for } u \geq 0 \text{ a.s.},$$

and the dual optimization problem

$$\begin{aligned} &\text{maximize } g(u, v) \text{ over } u \in L^{r'}, v \in L^{w'} \\ &\text{subject to } u \geq 0. \end{aligned}$$

The strong duality equality:

$$\min_z \{f(z) \mid g(z) \leq 0, h(z) = 0\} = \max_{u, v} \{g(u, v) \mid u \geq 0\}$$

The Minimax Theorem

Theorem 8 (Sion's Minimax Theorem, Corollary 3.3 in [25]) *Suppose X, Y are convex spaces such that one of them is compact, and $f(x, y)$ is a function on $X \times Y$, such that $x \mapsto f(x, y)$ is quasi-concave and u.s.c. for each fixed y and $y \mapsto f(x, y)$ is quasi-convex and l.s.c. for each fixed x . Then*

$$\sup_{x \in X} \inf_{y \in Y} f(x, y) = \inf_{y \in Y} \sup_{x \in X} f(x, y).$$

References

1. Artzner, P., Delbaen, F., Eber, J.M., Heath, D.: Coherent measures of risk. *Math. Financ.* **9**(3), 203–228 (1999)
2. Cheridito, P., Li, T.: Risk measures on Orlicz hearts. *Math. Financ.* **19**(2), 189–214 (2009)
3. Cherny, A., Madan, D.: New measures for performance evaluation. *Rev. Financ. Stud.* **22**(7), 2571–2606 (2008)
4. Cogneau, P., Hübner, G.: The (more than) 100 ways to measure portfolio performance. Part 2: special measures and comparison. *J. Perform. Meas.* **14**, 56–69 (2009)

5. Cogneau, P., Hübner, G.: The (more than) 100 ways to measure portfolio performance. Part 1: standardized risk-adjusted measures. *J. Perform. Meas.* **13**, 56–71 (2009)
6. Cui, X., Li, D., Wang, S., Zhu, S.: Better than dynamic mean-variance: time inconsistency and free cash flow stream. *Math. Financ.* **22**(2), 346–378 (2012)
7. Davis, J.R., Uryasev, S.: Analysis of tropical storm damage using buffered probability of exceedance. *Nat. Hazards* **83**(1), 465–483 (2016)
8. Föllmer, H., Schied, A.: *Stochastic Finance: An Introduction in Discrete Time*. Walter de Gruyter, Berlin (2011)
9. Krokmal, P.A.: Higher moment coherent risk measures. *Quant. Financ.* **7**, 373–387 (2007)
10. Le Sourd, V.: Performance measurement for traditional investment. *Financ. Anal. J.* **58**(4), 36–52 (2007)
11. Li, D., Ng, W.L.: Optimal dynamic portfolio selection: multiperiod mean-variance formulation. *Math. Financ.* **10**(3), 387–406 (2000)
12. Mafusalov, A., Uryasev, S.: Buffered probability of exceedance: mathematical properties and optimization algorithms. Risk Management and Financial Engineering Lab, Department of Industrial and Systems Engineering, University of Florida, Research Report, 1, 2015–5 (2014)
13. Mafusalov, A., Shapiro, A., Uryasev, S.: Estimation and asymptotics for buffered probability of exceedance. *Eur. J. Oper. Res.* **270**(3), 826–836 (2018)
14. Markowitz, H.: Portfolio selection. *J. Financ.* **7**(1), 77–91 (1952)
15. Pedersen, J.L., Peskir, G.: Optimal mean-variance selling strategies. *Math. Financ. Econ.* **10**(2), 203–220 (2016)
16. Pedersen, J.L., Peskir, G.: Optimal mean-variance portfolio selection. *Math. Financ. Econ.* **11**(2), 137–160 (2017)
17. Peskir, G., Shiryaev, A.: *Optimal Stopping and Free-Boundary Problems*. Springer, Berlin (2006)
18. Pliska, S.: *Introduction to Mathematical Finance*. Blackwell Publishers, Oxford (1997)
19. Richardson, H.R.: A minimum variance result in continuous trading portfolio optimization. *Manag. Sci.* **35**(9), 1045–1055 (1989)
20. Rockafellar, R.T.: *Conjugate Duality and Optimization*, vol. 16. SIAM, Philadelphia (1974)
21. Rockafellar, R.T., Royset, J.O.: On buffered failure probability in design and optimization of structures. *Reliab. Eng. Syst. Saf.* **95**(5), 499–510 (2010)
22. Rockafellar, R.T., Uryasev, S.: Optimization of conditional value-at-risk. *J. Risk* **2**, 21–42 (2000)
23. Sharpe, W.F.: Mutual fund performance. *J. Bus.* **39**(1), 119–138 (1966)
24. Sharpe, W.F.: The sharpe ratio. *J. Portf. Manag.* **21**(1), 49–58 (1994)
25. Sion, M.: On general minimax theorems. *Pac. J. Math.* **8**(1), 171–176 (1958)
26. White, D.: Dynamic programming and probabilistic constraints. *Oper. Res.* **22**(3), 654–664 (1974)
27. Zhitlukhin, M.: A second-order monotone modification of the sharpe ratio. In: *Recent Advances in Financial Engineering 2014: Proceedings of the TMU Finance Workshop 2014*, pp. 217–226. World Scientific, Singapore (2016)
28. Zhitlukhin, M.: On maximization of the expectation-to-deviation ratio of a random variable. *Russ. Math. Surv.* **72**(4), 765 (2017)

On Chernoff's Test for a Fractional Brownian Motion



Alexey Muravlev and Mikhail Zhitlukhin

Abstract We construct a sequential test for the sign of the drift of a fractional Brownian motion. We work in the Bayesian setting and assume the drift has a prior normal distribution. The problem reduces to an optimal stopping problem for a standard Brownian motion, obtained by a transformation of the observable process. The solution is described as the first exit time from some set, whose boundaries satisfy certain integral equations, which are solved numerically.

1 Introduction

This paper provides an overview of the results obtained in the paper [22].

Suppose one observes a fractional Brownian motion process (fBm) with linear drift and unknown drift coefficient. We are interested in sequentially testing the hypotheses that the drift coefficient is positive or negative. We consider a Bayesian setting where the drift coefficient has a prior normal distribution, and we use an optimality criteria of a test which consists of a linear penalty for the duration of observation and a penalty for a wrong decision proportional to the true value of the drift coefficient. The main result of this paper describes the structure of the exact optimal test in this problem, i.e. specifies a time to stop observation and a rule to choose between the two hypotheses.

The main novelty of our work compared to the large body of literature related to sequential tests (for an overview of the field, see e.g. [10, 20]) is that we work with fBm. To the best of our knowledge, this is the first non-asymptotic solution of a continuous-time sequential testing problem for this process. It is well-known that a fBm is not a Markov process, neither a semimartingale except the particular case when it is a standard Brownian motion (standard Bm). As a consequence, many standard tools of stochastic calculus and stochastic control (Itô's formula, the HJB equation, etc.) cannot be directly applied in models based on fBm. Fortunately, in

A. Muravlev · M. Zhitlukhin (✉)

Steklov Mathematical Institute of Russian Academy of Sciences, Moscow, Russia

e-mail: almurav@mi.ras.ru; mikhailzh@mi.ras.ru

the problem we consider it turns out to be possible to change the original problem for fBm so that it becomes tractable. One of the key steps is a general transformation outlined in the note [13], which allows to reduce sequential testing problems for fBm to problems for diffusion processes.

In the literature, the result which is most closely related to ours is the sequential test proposed by Chernoff [3], which has exactly the same setting and uses the same optimality criterion, but considers only standard Bm. For a prior normal distribution of the drift coefficient, Chernoff and Breakwell [1, 4] found asymptotically optimal sequential tests when the variance of the drift goes to zero or infinity.

Let us mention two other recent results in the sequential analysis of fBm, related to estimation of its drift coefficient. Çetin et al. [2] considered a sequential estimation problem assuming a normal prior distribution of the drift with a quadratic or a δ -function penalty for a wrong estimate and a linear penalty for observation time. They proved that in their setting the optimal stopping time is non-random. Gapeev and Stoev [8] studied sequential testing and changepoint detection problems for Gaussian processes, including fBm. They showed how those problems can be reduced to optimal stopping problems and found asymptotics of optimal stopping boundaries. There are many more results related to fixed-sample (i.e. non-sequential) statistical analysis of fBm. See, for example, Part II of the recent monograph [19], which discusses statistical methods for fBm in details.

The remaining part of our paper is organized as follows. Section 2 formulates the problem. Section 3 describes a transformation of the original problem to an optimal stopping problem for a standard Bm and introduces auxiliary processes which are needed to construct the optimal sequential test. The main result of the paper—the theorem which describes the structure of the optimal sequential test—is presented in Sect. 4, together with a numerical solution.

2 Decision Rules and Their Optimality

Recall that the fBm B_t^H , $t \geq 0$, with Hurst parameter $H \in (0, 1)$ is a zero-mean Gaussian process with the covariance function

$$\text{cov}(B_t^H, B_s^H) = \frac{1}{2}(s^{2H} + t^{2H} - |t - s|^{2H}), \quad t, s \geq 0.$$

In the particular case $H = 1/2$ this process is a standard Brownian motion (standard Bm) and has independent increments; its increments are positively correlated in the case $H > 1/2$ and negatively correlated in the case $H < 1/2$.

Suppose one observes the stochastic process

$$Z_t = \theta t + B_t^H, \quad t \geq 0,$$

where $H \in (0, 1)$ is known, and θ is a random variable independent of B^H and having a normal distribution with known mean $\mu \in \mathbb{R}$ and known variance $\sigma^2 > 0$.

It is assumed that neither the value of θ , nor the value of B_t^H can be observed directly, but the observer wants to determine whether the value of θ is positive or negative based on the information conveyed by the combined process Z_t . We will look for a sequential test for the hypothesis $\theta > 0$ versus the alternative $\theta \leq 0$. By a sequential test we call a pair $\delta = (\tau, d)$, which consists of a stopping time τ of the filtration \mathcal{F}_t^Z , generated by Z , and an \mathcal{F}_τ^Z -measurable function d assuming values ± 1 . The stopping time is the moment of time when observation is terminated and a decision about the hypotheses is made; the value of d shows which of them is accepted.

We will use the criterion of optimality of a decision rule consisting in minimizing the linear penalty for observation time and the penalty for a wrong decision proportional to the absolute value of θ . Namely, with each decision rule δ we associate the risk

$$R(\delta) = E(\tau + |\theta|I(d \neq \text{sgn}(\theta))).$$

The problem consists in finding δ^* that minimizes $R(\delta)$ over all decision rules.

This problem was proposed by Chernoff in [3] for standard Bm, and we refer the reader to that paper for a rationale for this setting. The subsequent papers [1, 4, 5] include results about the asymptotics of the optimal test and other its properties, including a comparison with Wald’s sequential probability ratio test. Our paper [21] contains a result which allows to find the exact (non-asymptotic) optimal test by a relatively simple numerical procedure.

3 Reduction to an Optimal Stopping Problem

From the relation $|\theta|I(d \neq \text{sgn}(\theta)) = \theta^+I(d = -1) + \theta^-I(d = 1)$, where $\theta^+ = \max(\theta, 0)$, $\theta^- = -\min(\theta, 0)$, and from that d is \mathcal{F}_τ^Z -measurable, one can see that the optimal decision rule should be looked for among rules (τ, d) with $d = \min(E(\theta^- | \mathcal{F}_\tau^Z), E(\theta^+ | \mathcal{F}_\tau^Z))$. Hence, it will be enough to solve the optimal stopping problem which consists in finding a stopping time τ^* such that $R(\tau^*) = \inf_\tau R(\tau)$, where

$$R(\tau) = E(\tau + \min(E(\theta^- | \mathcal{F}_\tau^Z), E(\theta^+ | \mathcal{F}_\tau^Z)))$$

(for brevity, we’ll use the same notation R for the functional associated with a decision rule, and the functional associated with a stopping time).

We will transform the expression inside the expectation in $R(\tau)$ to the value of some process, constructed from a standard Bm. Introduce the process $X_t, t \geq 0$, by

$$X_t = C_H \int_0^t K_H(t, s)dZ_s$$

with the integration kernel $K_H(t, s) = (t-s)^{\frac{1}{2}-H} {}_2F_1(\frac{1}{2}-H, \frac{1}{2}-H, \frac{3}{2}-H, \frac{s-t}{t})$, where ${}_2F_1$ is the Gauss (ordinary) hypergeometric function, and the constant $C_H = \left(\frac{\Gamma(2-2H)}{2H\Gamma(\frac{1}{2}+H)(\Gamma(\frac{3}{2}-H))^3}\right)^{\frac{1}{2}}$; as usual, Γ denotes the gamma function.

As follows from [9] (see also earlier results [12, 14]), $B_t = C_H \int_0^t K_H(t, s) dB_s^H$ is a standard Bm, and by straightforward computation we obtain the representation

$$dX_t = B_t + \theta L_H t^{\frac{1}{2}-H} dt,$$

and the filtrations of the processes Z_t and X_t coincide. The constant L_H in the above formula is defined by $L_H = (2H(\frac{3}{2}-H)B(\frac{1}{2}+H, 2-2H))^{-\frac{1}{2}}$, where B is the beta function.

Now one can find that the conditional distribution $\text{Law}(\theta \mid \mathcal{F}_\tau^X)$ is normal and transform the expression for the risk to

$$R(\tau) = E\left(\tau - \frac{1}{2}Y_\tau\right) + const,$$

where *const* denotes some constant (depending on μ, σ, H), the value of which is not essential for what follows, and $Y_t, t \geq 0$, is the process satisfying the SDE

$$dY_t = t^{\frac{1}{2}-H}((\sigma L_H)^{-2} + t^{2-2H}/(2-2H))d\tilde{B}_t, \quad Y_0 = \mu L_H,$$

where \tilde{B} is another standard Bm, the innovation Brownian motion (see e.g. Chapter 7.4 in [11]). For brevity, denote $\gamma = (2-2H)^{-1}$. Then under the following monotone change of time

$$t(r) = \left(\frac{(2-2H)r}{(\sigma L_H)^2(1-r)}\right)^\gamma, \quad r \in [0, 1),$$

where t runs through the half-interval $[0, \infty)$ when r runs through $[0, 1)$, the process

$$W_r = (\sigma L_H)^{-1}Y_{t(r)} - \mu\sigma^{-1}$$

is a standard Bm in $r \in [0, 1)$, and the filtrations \mathcal{F}_r^W and $\mathcal{F}_{t(r)}^X$ coincide. Denote

$$M_{\sigma,H} = \frac{2}{\sigma} \left(\frac{2-2H}{\sigma^2 L_H^2}\right)^\gamma.$$

Then the optimal stopping problem for X in t -time is equivalent to the following optimal stopping problem for W in r -time:

$$V = \inf_{\rho < 1} E\left(M_{\sigma,H} \left(\frac{\rho}{1-\rho}\right)^\gamma - \left|W_\rho + \frac{\mu}{\sigma}\right|\right). \tag{1}$$

Namely, if ρ^* is the optimal stopping time for V , then the optimal decision rule $\delta^* = (\tau^*, d^*)$ is given by

$$\tau^* = t(\rho^*), \quad d^* = I(a_{\tau^*} > 0) - I(a_{\tau^*} \leq 0). \tag{2}$$

4 The Main Results

In this section we formulate a theorem about the solution of problem (1), which gives an optimal sequential test through transformation (2). Throughout we will assume that σ and H are fixed and will denote the function

$$f(t) = M_{\sigma,H} \left(\frac{t}{1-t} \right)^\gamma.$$

It is well-known that under general conditions the solution of an optimal stopping problem for a Markov process can be represented as the first time when the process enters a certain set (a stopping set). Namely, let us first rewrite our problem in the Markov setting by allowing the process W_t to start from any point $(t, x) \in [0, 1) \times \mathbb{R}$:

$$V(t, x) = \inf_{\rho < 1-t} E(f(t + \rho) - |W_\rho + x|) - f(t). \tag{3}$$

For example, for the quantity V from (1) we have $V = V(0, \frac{\mu}{\sigma})$. We subtract $f(t)$ in the definition of $V(t, x)$ to make the function $V(t, x)$ bounded. For $t = 1$ we define $V(1, x) = -|x|$.

The following theorem describes the structure of the optimal stopping time in problem (3). In its statement, we set

$$t_0 = t_0(H) := \max\left(0, \frac{1 - 2H}{4(1 - H)}\right).$$

Obviously, $t_0 > 0$ for $H < \frac{1}{2}$ and $t_0 = 0$ for $H \geq \frac{1}{2}$.

Theorem 1

1) *There exists a function $A(t)$ defined on $(t_0, 1]$, which is continuous, strictly decreasing, and strictly positive for $t < 1$ with $A(1) = 0$, such that for any $t > t_0$ and $x \in \mathbb{R}$ the optimal stopping time in the problem (3) is given by*

$$\rho^*(t, x) = \inf\{s \geq 0 : |W_s + x| \geq A(t + s)\}.$$

Moreover, for any $t \in (t_0, 1]$ the function $A(t)$ satisfies the inequality

$$A(t) \leq \frac{(1-t)^\gamma}{2M_{\sigma,H}t^{\gamma-1}}. \tag{4}$$

- 2) The function $A(t)$ is the unique function which is continuous, non-negative, satisfies (4), and solves the integral equation

$$G(t, A(t)) = \int_t^1 F(t, A(t), s, A(s)) ds, \quad t \in (t_0, 1), \quad (5)$$

with the functions $G(t, x) = E|\zeta\sqrt{1-t} + x| - x$ and $F(t, x, s, y) = f'(s)P(|\zeta\sqrt{s-t} + x| \leq y)$ for a standard normal random variable ζ .

Regarding the statement of this theorem, first note that for $H < \frac{1}{2}$, the stopping boundary $A(t)$ is described only for $t > t_0 > 0$. The main reason is that the method of proof we use to show that $A(t)$ is continuous and satisfies the integral equation requires it to be of bounded variation (at least, locally). In particular, this is a sufficient condition for applicability of the Itô formula with local time on curves [16], which is used in the proof. In the case $H \geq \frac{1}{2}$ and for $t \geq t_0$ in the case $H < \frac{1}{2}$ by a direct probabilistic argument we can prove that $A(t)$ is monotone and therefore has bounded variation; this argument however doesn't work for $t < t_0$ in the case $H < \frac{1}{2}$, and, as a formal numerical solution shows, the boundary $A(t)$ seems to be indeed not monotone in that case. Of course, the assumption of bounded variation can be relaxed while the Itô formula can still be applicable (see e.g. [6, 7, 16]), however verification of weaker sufficient conditions is problematic. Although the general scheme to obtain integral equations of type (5) and prove uniqueness of their solutions has been discovered quite a while ago (the first full result was obtained by Peskir for the optimal stopping problem for American options, see [17]), and has been used many times in the literature for various optimal stopping problems (a large number of examples can be found in [18]), we are unaware of any of its applications in the case when stopping boundaries are not monotone and/or cannot be transformed to monotone ones by space and/or time change. Nevertheless, a formal numerical solution of the integral equation produces stopping boundaries which "look smooth", but we admit that a rigor proof of this fact in the above-mentioned cases remains an open question.

Note also, that in the case $H \geq \frac{1}{2}$ the space-time transformation we apply to pass from the optimal stopping problem for the process $a(t)$ to the problem for W_r is essential from this point of view, because the boundaries in the problem for $a(t)$ are not monotone. Moreover, they are not monotone even in the case $H = \frac{1}{2}$, when $a(t)$ is obtained by simply shifting X_r in time and space, see [3, 21].

The second remark we would like to make is that in the case $H > \frac{1}{2}$ we do not know whether $A(0)$ is finite. In the case $H = \frac{1}{2}$ the finiteness of $A(0)$ follows from inequality (4), which is proved by a direct argument based on comparison with a simpler optimal stopping problem (one can see that it extends to $t = 0$). It seems that a deeper analysis may be required for the case $H > \frac{1}{2}$, which is beyond this paper.

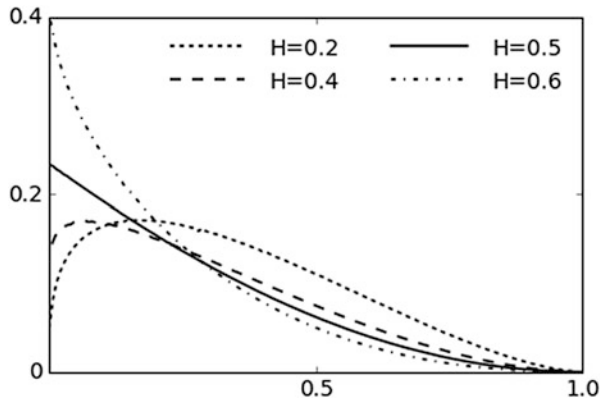


Fig. 1 The stopping boundary $A(t)$ for different values of H and $\sigma = 1$

Figure 1 shows the stopping boundary $A(t)$ for different values H computed by solving Eq. (5) numerically. A description of the numerical method, based on “backward induction”, can be found, for example, in [15].

References

1. Breakwell, J., Chernoff, H.: Sequential tests for the mean of a normal distribution II (large t). *Ann. Math. Stat.* **35**, 162–173 (1964)
2. Çetin, U., Novikov, A., Shiryaev, A.N.: Bayesian sequential estimation of a drift of fractional Brownian motion. *Seq. Anal.* **32**, 288–296 (2013)
3. Chernoff, H.: Sequential tests for the mean of a normal distribution. In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 79–91 (1961)
4. Chernoff, H.: Sequential test for the mean of a normal distribution III (small t). *Ann. Math. Stat.* **36**, 28–54 (1965)
5. Chernoff, H.: sequential tests for the mean of a normal distribution IV (discrete case). *Ann. Math. Stat.* **36**, 55–68 (1965)
6. Eisenbaum, N.: Local timespace stochastic calculus for Lévy processes. *Stoch. Process. Appl.* **116**, 757–778 (2006)
7. Föllmer, H., Protter, P., Shiryaev, A.N.: Quadratic covariation and an extension of Itô’s formula. *Bernoulli* **1**, 149–169 (1995)
8. Gapeev, P.V., Stoev, Y.I.: On the Laplace transforms of the first exit times in one-dimensional non-affine jumpdiffusion models. *Stat. Probab. Lett.* **121**, 152–162 (2017)
9. Jost, C.: Transformation formulas for fractional Brownian motion. *Stoch. Process. Appl.* **116**, 1341–1357 (2006)
10. Lai, T.L.: On optimal stopping problems in sequential hypothesis testing. *Stat. Sin.* **7**, 33–51 (1997)
11. Liptser, R.S., Shiryaev, A.N.: *Statistics of Random Processes I: General Theory*, 2nd edn. Springer, Berlin (2001)
12. Molchan, G.M., Golosov, Y.I.: Gaussian stationary processes with asymptotical power spectrum. *Dokl. Akad. Nauk SSSR* **184**, 546–549 (1969)

13. Muravlev, A.A.: Methods of sequential hypothesis testing for the drift of a fractional Brownian motion. *Russ. Math. Surv.* **68**, 577–579 (2013)
14. Norros, I., Valkeila, E., Virtamo, J.: An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motions. *Bernoulli* **5**, 571–587 (1999)
15. Pedersen, J.L., Peskir, G.: On nonlinear integral equations arising in problems of optimal stopping. In: *Functional Analysis VII: proceedings, Dubrovnik, 17–26 Sept 2001* (2002)
16. Peskir, G.: A change-of-variable formula with local time on curves. *J. Theor. Probab.* **18**, 499–535 (2005)
17. Peskir, G.: On the American option problem. *Math. Financ.* **15**, 169–181 (2005)
18. Peskir, G., Shiryaev, A.: *Optimal Stopping and Free-Boundary Problems*. Birkhäuser, Basel (2006)
19. Tanaka, K.: *Time Series Analysis: Nonstationary and Noninvertible Distribution Theory*. Wiley, Hoboken (2017)
20. Tartakovsky, A., Nikiforov, I., Basseville, M.: *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. CRC Press, Boca Raton (2014)
21. Zhitlukhin, M.V., Muravlev, A.A.: On Chernoff's hypotheses testing problem for the drift of a Brownian motion. *Theory Probab. Appl.* **57**, 708–717 (2013)
22. Zhitlukhin, M.V., Muravlev, A.A.: A Bayesian sequential test for the drift of a fractional Brownian motion (2018). arXiv:1804.02757

How to Sheafify an Elliptic Quantum Group



Yaping Yang and Gufang Zhao

Abstract These lecture notes are based on Yang's talk at the MATRIX program *Geometric R-Matrices: from Geometry to Probability*, at the University of Melbourne, Dec. 18–22, 2017, and Zhao's talk at Perimeter Institute for Theoretical Physics in January 2018. We give an introductory survey of the results in Yang and Zhao (Quiver varieties and elliptic quantum groups, 2017. arxiv1708.01418). We discuss a sheafified elliptic quantum group associated to any symmetric Kac-Moody Lie algebra. The sheafification is obtained by applying the equivariant elliptic cohomological theory to the moduli space of representations of a preprojective algebra. By construction, the elliptic quantum group naturally acts on the equivariant elliptic cohomology of Nakajima quiver varieties. As an application, we obtain a relation between the sheafified elliptic quantum group and the global affine Grassmannian over an elliptic curve.

1 Motivation

The parallelism of the following three different kinds of mathematical objects was observed in [11] by Ginzburg-Kapranov-Vasserot.

Y. Yang (✉)

School of Mathematics and Statistics, The University of Melbourne, Parkville, VIC, Australia
e-mail: yaping.yang1@unimelb.edu.au

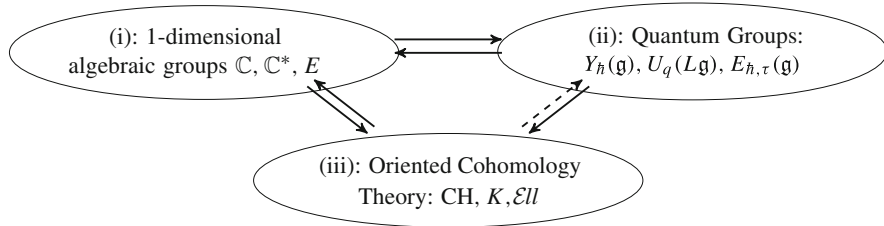
G. Zhao

Institute of Science and Technology Austria, Klosterneuburg, Austria
e-mail: gufang.zhao@ist.ac.at

© Springer Nature Switzerland AG 2019

D. R. Wood et al. (eds.), *2017 MATRIX Annals*, MATRIX Book Series 2,
https://doi.org/10.1007/978-3-030-04161-8_54

675



Here, the correspondence (i) \leftrightarrow (iii) is well known in algebraic topology, goes back to the work of Quillen, Hirzebruch, et al. Similar correspondence also exists in algebraic geometry thanks to the oriented cohomology theories (OCT) of Levine and Morel [15]. The algebraic OCT associated to \mathbb{C} , \mathbb{C}^* and E are, respectively, the intersection theory (Chow groups) CH, the algebraic K-theory K and the algebraic elliptic cohomology Ell .

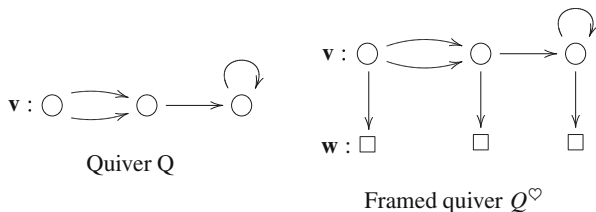
The correspondence (i) \leftrightarrow (ii) was introduced to the mathematical community by Drinfeld [6]. Roughly speaking, the quantum group in (ii) quantizes of the Lie algebra of maps from the algebraic group in (i) to a Lie algebra \mathfrak{g} . The quantization is obtained from the solutions to the quantum Yang-Baxter equation

$$R_{12}(u)R_{13}(u + v)R_{23}(v) = R_{23}(v)R_{13}(u + v)R_{12}(u). \tag{QYBE}$$

The Yangian $Y_h(\mathfrak{g})$, quantum loop algebra $U_q(L\mathfrak{g})$, and elliptic quantum group $E_{h,\tau}(\mathfrak{g})$ are respectively obtained from the rational, trigonometric and elliptic solutions of the QYBE. There is a dynamical elliptic quantum group associated to a general symmetrizable Kac-Moody Lie algebra, which is obtained from solutions to the dynamical Yang-Baxter equation.

The correspondence (ii) \leftrightarrow (iii) is the main subject of this note, and is originated from the work of Nakajima [21], Varagnolo [24], Maulik-Okounkov [17], and many others. Without going to the details, the quantum group in (ii) acts on the corresponding oriented cohomology of the Nakajima quiver varieties recalled below.

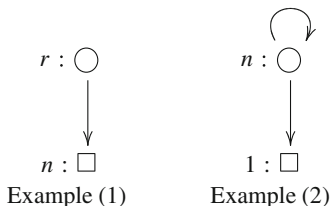
Let $Q = (I, H)$ be a quiver, with I being the set of vertices, and H being the set of arrows. Let Q^\heartsuit be the framed quiver, schematically illustrated below. For any dimension vector (\mathbf{v}, \mathbf{w}) of Q^\heartsuit , we have the Nakajima quiver variety $\mathfrak{M}(\mathbf{v}, \mathbf{w})$.



Denote $\mathfrak{M}(\mathbf{w}) = \coprod_{\mathbf{v} \in \mathbb{N}^l} \mathfrak{M}(\mathbf{v}, \mathbf{w})$. We follow the terminologies and notations in [21] on quiver varieties, hence refer the readers to *loc. cit.* for the details. Nevertheless, we illustrate by the following examples, and fix some conventions in Sect. 2.

Example 1

1. Let Q be the quiver with one vertex, no arrows. Let $(\mathbf{v}, \mathbf{w}) = (r, n)$, with $0 \leq r \leq n$. Then, $\mathfrak{M}(r, n) \cong T^*\text{Gr}(r, n)$, the cotangent bundle of the Grassmannian.
2. Let Q be the Jordan quiver with one vertex, and one self-loop. Let $(\mathbf{v}, \mathbf{w}) = (n, 1)$, with $n \in \mathbb{N}$. Then, $\mathfrak{M}(n, 1) \cong \text{Hilb}^n(\mathbb{C}^2)$, the Hilbert scheme of n -points on \mathbb{C}^2 .



When Q has no edge-loops, Nakajima in [21] constructed an action of $U_q(\mathfrak{L}\mathfrak{g})$ on the equivariant K-theory of $\mathfrak{M}(\mathbf{w})$, with \mathfrak{g} being the symmetric Kac-Moody Lie algebra associated to Q . Varagnolo in [24] constructed an action of $Y_{\hbar}(\mathfrak{g})$ on the equivariant cohomology of the Nakajima quiver varieties. For a general quiver Q , Maulik-Okounkov in [17] constructed a bigger Yangian \mathbb{Y}_{MO} acting on the equivariant cohomology of $\mathfrak{M}(\mathbf{w})$ using a geometric R-matrix formalism. See [1, 22] for the geometric R-matrix construction in the trigonometric (K-theoretic stable envelope), and elliptic case (elliptic stable envelope).

The goal of the present notes is to explain a direct construction of the correspondence from (iii) to (ii) above, using cohomological Hall algebra (CoHA) following [28]. Most of the constructions are direct generalizations of Schiffmann and Vasserot [23]. Closely related is the CoHA of Kontsevich-Soibelman [14], defined to be the critical cohomology (cohomology valued in a vanishing cycle) of the moduli space of representations of a quiver with potential. A relation between the CoHA in the present note and the CoHA from [14] is explained in [25].

This approach of studying quantum groups has the following advantages.

- The construction works for any oriented cohomology theory, beyond CH, K , $\mathcal{E}ll$. One interesting example is the Morava K-theory. The new quantum groups obtained via this construction are expected to be related the Lusztig’s character formulas [28, §6].
- In the case of $\mathcal{E}ll$, the construction gives a sheafified elliptic quantum group, as well as an action of $E_{\hbar, \tau}(\mathfrak{g})$ on the equivariant elliptic cohomology of Nakajima quiver varieties, as will be explained in Sect. 3.

2 Construction of the Cohomological Hall Algebras

For illustration purpose, in this section we take the OCT to be the intersection theory CH. Most statements have counterparts in an arbitrary oriented cohomology theory.

2.1 The Theorem

Let $Q = (I, H)$ be an arbitrary quiver. Denote by \mathfrak{g}_Q the corresponding symmetric Kac-Moody Lie algebra associated to Q . The preprojective algebra, denoted by Π_Q , is the quotient of the path algebra $\mathbb{C}(Q \cup Q^{op})$ by the ideal generated by the relation $\sum_{x \in H} [x, x^*] = 0$, where $x^* \in H^{op}$ is the reversed arrow of $x \in H$. Fix a dimension vector $\mathbf{v} \in \mathbb{N}^I$, let $\text{Rep}(Q, \mathbf{v})$ be the affine space parametrizing representations of the path algebra $\mathbb{C}Q$ of dimension \mathbf{v} , and let $\text{Rep}(\Pi_Q, \mathbf{v})$ be the affine algebraic variety parametrizing representations of Π_Q of dimension \mathbf{v} , with an action of $\text{GL}_{\mathbf{v}} := \prod_{i \in I} \text{GL}_{v_i}$. Here $\mathbf{v} = (v^i)_{i \in I}$. Let $\mathfrak{R}\text{ep}(\Pi_Q, \mathbf{v}) := \text{Rep}(\Pi_Q, \mathbf{v})/\text{GL}_{\mathbf{v}}$ be the quotient stack.



Example 2 Let Q be the Jordan quiver: \circlearrowleft^x . The preprojective algebra $\Pi_Q = \mathbb{C}[x, x^*]$ is the free polynomial ring in two variables. We have

$$\mathfrak{R}\text{ep}(\Pi_Q, n) = \{(A, B) \in (\mathfrak{gl}_n)^2 \mid [A, B] = 0\}/\text{GL}_n.$$

Consider the graded vector space

$$\mathcal{P}(\text{CH}, Q) := \bigoplus_{\mathbf{v} \in \mathbb{N}^I} \text{CH}_{\mathbb{C}^*}(\mathfrak{R}\text{ep}(\Pi_Q, \mathbf{v})) = \bigoplus_{\mathbf{v} \in \mathbb{N}^I} \text{CH}_{\text{GL}_{\mathbf{v}} \times \mathbb{C}^*}(\text{Rep}(\Pi_Q, \mathbf{v})). \quad (1)$$

The torus \mathbb{C}^* acts on $\text{Rep}(\Pi_Q, \mathbf{v})$ the same way as in [21, (2.7.1) and (2.7.2)]. More explicitly, let a be the number of arrows in Q from vertex i to j ; We enumerate these arrows as h_1, \dots, h_a . The corresponding reversed arrows in Q^{op} are enumerated as h_1^*, \dots, h_a^* . We define an action of \mathbb{C}^* on $\text{Rep}(\Pi_Q, \mathbf{v})$ in the following way. For $t \in \mathbb{C}^*$ and $(B_p, B_p^*) \in \text{Rep}(\Pi_Q, \mathbf{v})$ with $h_p \in H$, we define

$$t \cdot B_p := t^{a+2-2p} B_p, \quad t \cdot B_p^* := t^{-a+2p} B_p^*$$

Theorem A ([27, 28, Yang-Zhao])

1. The vector space $\mathcal{P}(\text{CH})$ is naturally endowed with a product \star and a coproduct Δ , making it a bialgebra.
2. On $D(\mathcal{P}) := \mathcal{P} \otimes \mathcal{P}$, there is a bialgebra structure obtained as a Drinfeld double of $\mathcal{P}(\text{CH})$. For any $\mathbf{w} \in \mathbb{N}^I$, the algebra $D(\mathcal{P})$ acts on $\text{CH}_{\text{GL}_{\mathbf{w}}}(\mathfrak{M}(\mathbf{w}))$.

3. Assume Q has no edge loops. There is a certain spherical subalgebra $D(\mathcal{P}^{\text{sph}}) \subseteq D(\mathcal{P}(\text{CH}))$, such that

$$D(\mathcal{P}^{\text{sph}}) \cong Y_h(\mathfrak{g}_Q).$$

Furthermore, the induced Yangian action from (2) is compatible with the action constructed by Varagnolo in [24], and, in the case when Q is of ADE type, the action of Maulik-Okounkov in [17].

Remark 1 The construction of the Drinfeld double involves adding a Cartan subalgebra and defining a bialgebra pairing, as can be found in detail in [27, §3]. The Cartan subalgebra can alternatively be replaced by a symmetric monoidal structure as in [26, §5]. The definition of the spherical subalgebra can be found in [27, §3.2].

2.2 Constructions

The Hall multiplication \star of $\mathcal{P}(\text{CH})$ is defined using the following correspondence.

$$\mathfrak{Rep}(\Pi_Q, \mathbf{v}_1) \times \mathfrak{Rep}(\Pi_Q, \mathbf{v}_2) \begin{array}{c} \xleftarrow{\phi} \mathfrak{Ext} \xrightarrow{\psi} \\ \end{array} \mathfrak{Rep}(\Pi_Q, \mathbf{v}_1 + \mathbf{v}_2)$$

where \mathfrak{Ext} is the moduli space of extensions $\{0 \rightarrow V_1 \rightarrow V \rightarrow V_2 \rightarrow 0 \mid \dim(V_i) = \mathbf{v}_i, i = 1, 2\}$. The map $\phi : (0 \rightarrow V_1 \rightarrow V \rightarrow V_2 \rightarrow 0) \mapsto (V_1, V_2)$ is smooth, and $\psi : (0 \rightarrow V_1 \rightarrow V \rightarrow V_2 \rightarrow 0) \mapsto V$ is proper. The Hall multiplication \star is defined to be

$$\star = \psi_* \circ \phi^*.$$

Here the stacks $\mathfrak{Rep}(\Pi_Q, \mathbf{v})$ for $\mathbf{v} \in \mathbb{N}^I$ are endowed with obstruction theories obtained from the embeddings $\mathfrak{Rep}(\Pi_Q, \mathbf{v}) \hookrightarrow T^* \text{Rep}(Q, \mathbf{v})/\text{GL}_{\mathbf{v}}$, and \mathfrak{Ext} has a similar obstruction theory described in detail in [28, §4.1]. Similar for the construction of the action below.

Now we explain the action in Theorem A (2). Let $\mathfrak{Rep}^{\text{fr}}(\Pi_Q, \mathbf{v}, \mathbf{w})$ be the moduli space of framed representations of Π_Q with dimension vector (\mathbf{v}, \mathbf{w}) , which is constructed as a quotient stack $\text{Rep}^{\text{fr}}(\Pi_Q, \mathbf{v}, \mathbf{w})/\text{GL}_{\mathbf{v}}$. Imposing a suitable semistability condition, explained in detail in [21], we get an open subset $\mathfrak{Rep}^{\text{fr}, ss}(\Pi_Q, \mathbf{v}, \mathbf{w}) \subset \mathfrak{Rep}^{\text{fr}}(\Pi_Q, \mathbf{v}, \mathbf{w})$. There is an isomorphism

$$\text{CH}(\mathfrak{Rep}^{\text{fr}, ss}(\Pi_Q, \mathbf{v}, \mathbf{w})) = \text{CH}_{\text{GL}_{\mathbf{w}}}(\mathfrak{M}(\mathbf{v}, \mathbf{w})).$$

We have the following general correspondence [28, §4.1]:

$$\mathfrak{Rep}^{\text{fr}}(\Pi_Q, \mathbf{v}_1, \mathbf{w}_1) \times \mathfrak{Rep}^{\text{fr}}(\Pi_Q, \mathbf{v}_2, \mathbf{w}_2) \xrightarrow[\overline{\psi}]{\overline{\phi}} \mathfrak{Ext}^{\text{fr}} \xrightarrow{\overline{\psi}} \mathfrak{Rep}^{\text{fr}}(\Pi_Q, \mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}_1 + \mathbf{w}_2) \quad (2)$$

The action in Theorem A (2) is defined as $(\overline{\psi}^{ss})_* \circ (\overline{\phi}^{ss})^*$ by taking $\mathbf{w}_1 = \mathbf{0}$, and imposing a suitable semistability condition on the correspondence (2).

2.3 Shuffle Algebra

Notations as before, let $Q = (I, H)$ be a quiver. Following the proof of [28, Proposition 3.4], we have the following shuffle description of $(\overline{\psi})_* \circ (\overline{\phi})^*$ in (2).

Let SH be an $\mathbb{N}^I \times \mathbb{N}^I$ -graded $\mathbb{C}[t_1, t_2]$ -algebra. As a $\mathbb{C}[t_1, t_2]$ -module, we have

$$\text{SH} = \bigoplus_{\mathbf{v} \in \mathbb{N}^I, \mathbf{w} \in \mathbb{N}^I} \text{SH}_{\mathbf{v}, \mathbf{w}},$$

where

$$\text{SH}_{\mathbf{v}, \mathbf{w}} := \mathbb{C}[t_1, t_2] \otimes \mathbb{C}[\lambda_s^i]_{i \in I, s=1, \dots, v^i}^{\mathfrak{S}_{\mathbf{v}}} \otimes \mathbb{C}[z_t^j]_{j \in I, t=1, \dots, w^j}^{\mathfrak{S}_{\mathbf{w}}}.$$

Here $\mathfrak{S}_{\mathbf{v}} = \prod_{i \in I} \mathfrak{S}_{v^i}$ is the product of symmetric groups, and $\mathfrak{S}_{\mathbf{v}}$ naturally acts on the variables $\{\lambda_s^i\}_{i \in I, s=1, \dots, v^i}$ by permutation. For any $(\mathbf{v}_1, \mathbf{w}_1)$ and $(\mathbf{v}_2, \mathbf{w}_2) \in \mathbb{N}^I \times \mathbb{N}^I$, we consider $\text{SH}_{\mathbf{v}_1, \mathbf{w}_1} \otimes_{\mathbb{C}[t_1, t_2]} \text{SH}_{\mathbf{v}_2, \mathbf{w}_2}$ as a subalgebra of

$$\mathbb{C}[t_1, t_2][\lambda_s^i, z_t^j]_{\left\{ \begin{array}{l} i \in I, s=1, \dots, (v_1^i + v_2^i), \\ j \in I, t=1, \dots, (w_1^j + w_2^j) \end{array} \right\}}$$

by sending $(\lambda_s^i, z_t^j) \in \text{SH}_{\mathbf{v}_1, \mathbf{w}_1}$ to (λ_s^i, z_t^j) , and $(\lambda_s^i, z_t^j) \in \text{SH}_{\mathbf{v}_2, \mathbf{w}_2}$ to $(\lambda_{s+v_1^i}^i, z_{t+w_1^j}^j)$.

We define the shuffle product $\text{SH}_{\mathbf{v}_1, \mathbf{w}_1} \otimes_{\mathbb{C}[t_1, t_2]} \text{SH}_{\mathbf{v}_2, \mathbf{w}_2} \rightarrow \text{SH}_{\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}_1 + \mathbf{w}_2}$,

$$f(\lambda_{\mathbf{v}_1}, z_{\mathbf{w}_1}) \otimes g(\lambda_{\mathbf{v}_2}, z_{\mathbf{w}_2}) \mapsto \sum_{\sigma \in \text{Sh}(\mathbf{v}_1, \mathbf{v}_2) \times \text{Sh}(\mathbf{w}_1, \mathbf{w}_2)} \sigma \left(f(\lambda'_{\mathbf{v}_1}, z'_{\mathbf{w}_1}) \cdot g(\lambda''_{\mathbf{v}_2}, z''_{\mathbf{w}_2}) \cdot \text{fac}_{\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}_1 + \mathbf{w}_2} \right), \quad (3)$$

with $\text{fac}_{\mathbf{v}_1 + \mathbf{v}_2, \mathbf{w}_1 + \mathbf{w}_2}$ specified as follows.

Let

$$\text{fac}_1 := \prod_{i \in I} \prod_{s=1}^{v_1^i} \prod_{t=1}^{v_2^i} \frac{\lambda_s^i - \lambda_t^i + t_1 + t_2}{\lambda_t^i - \lambda_s^i}. \quad (4)$$

Let $m : H \amalg H^{\text{op}} \rightarrow \mathbb{Z}$ be a function, which for each $h \in H$ provides two integers m_h and m_{h^*} . We define the torus $T = (\mathbb{C}^*)^2$ action on $\text{Rep}(Q \cup Q^{\text{op}})$ according to the function m , satisfying some technical conditions spelled out in [27, Assumption 1.1]. The T -equivariant variables are denoted by t_1, t_2 . Define

$$\begin{aligned} \text{fac}_2 := & \prod_{h \in H} \left(\prod_{s=1}^{v_1^{\text{out}(h)}} \prod_{t=1}^{v_2^{\text{in}(h)}} (\lambda_t^{\text{in}(h)} - \lambda_s^{\text{out}(h)} + m_h t_1) \prod_{s=1}^{v_1^{\text{in}(h)}} \prod_{t=1}^{v_2^{\text{out}(h)}} (\lambda_t^{\text{out}(h)} - \lambda_s^{\text{in}(h)} + m_{h^*} t_2) \right) \\ & \cdot \prod_{i \in I} \left(\prod_{s=1}^{v_1^i} \prod_{t=1}^{w_2^i} (z_t^{\text{in}(h)} - \lambda_s^i + t_1) \prod_{s=1}^{w_1^i} \prod_{t=1}^{v_2^i} (\lambda_t^{\text{in}(h)} - z_s^i + t_2) \right) \end{aligned} \tag{5}$$

Let

$$\text{fac}_{\mathbf{v}_1+\mathbf{v}_2, \mathbf{w}_1+\mathbf{w}_2} := \text{fac}_1 \cdot \text{fac}_2. \tag{6}$$

Proposition 1 *Under the identification*

$$\begin{aligned} & \text{CH}_{\text{GL}_{\mathbf{v}} \times \text{GL}_{\mathbf{w}} \times T}(\text{Rep}^{\text{fr}}(\Pi_Q, \mathbf{v}, \mathbf{w})) \\ & \cong \mathbb{C}[t_1, t_2] \otimes \mathbb{C}[\lambda_s^i]_{i \in I, s=1, \dots, v^i}^{\otimes \mathbf{v}} \otimes \mathbb{C}[z_t^j]_{j \in I, s=1, \dots, w^j}^{\otimes \mathbf{w}} =: \text{SH}_{\mathbf{v}, \mathbf{w}}, \end{aligned}$$

the map $(\overline{\psi})_* \circ (\overline{\phi})^*$ is equal to the multiplication (3) of the shuffle algebra $\text{SH} = \bigoplus_{\mathbf{v} \in \mathbb{N}^I, \mathbf{w} \in \mathbb{N}^I} \text{SH}_{\mathbf{v}, \mathbf{w}}$.

Proof The proof follows from the same proof as [28, Proposition 3.4] replacing the quiver Q by the framed quiver Q^\heartsuit .

Remark 2

1. For an arbitrary cohomology theory, Proposition 1 is still true when $A + B$ is replaced by $A +_F B$ in the formula (6) of $\text{fac}_{\mathbf{v}_1+\mathbf{v}_2, \mathbf{w}_1+\mathbf{w}_2}$, where F is the formal group law associated to this cohomology theory.
2. Restricting to the open subset $\mathfrak{R}\text{ep}^{\text{fr}, ss}(\Pi_Q, \mathbf{v})$ of $\mathfrak{R}\text{ep}^{\text{fr}}(\Pi_Q, \mathbf{v})$ induces a surjective map $\text{SH}_{\mathbf{v}, \mathbf{w}} \rightarrow \text{CH}_{\text{GL}_{\mathbf{w}}}(\mathfrak{M}(\mathbf{v}, \mathbf{w}))$, for $\mathbf{v} \in \mathbb{N}^I, \mathbf{w} \in \mathbb{N}^I$. The surjectivity follows from [18]. This map is compatible with the shuffle product of the left hand side, and the multiplication on the right hand side induced from (2).

2.4 Drinfeld Currents

Let $\mathbf{v} = e_i$ be the dimension vector valued 1 at vertex $i \in I$ and zero otherwise. Then, $\mathcal{P}_{e_i} = \text{CH}_{\text{GL}_{e_i} \times \mathbb{C}^*}(\text{Rep}(\Pi_Q, e_i)) = \mathbb{C}[\hbar][x_i]$. Let $(x_i)^k \in \mathcal{P}_{e_i}, k \in \mathbb{N}$, be the

natural monomial basis of \mathcal{P}_{e_i} . One can form the Drinfeld currents

$$\mathfrak{X}_i^+(u) := \sum_{k \in \mathbb{N}} (x_i)^k u^{-k-1}, \quad i \in I.$$

By Theorem A, the generating series $\{\mathfrak{X}_i^+(u) \mid i \in I\}$ satisfy the relations of $Y_h^+(\mathfrak{g})[[u]]$ [28, § 7].

3 Sheafified Elliptic Quantum Groups

In this section, we applying the equivariant elliptic cohomology to the construction in Sect. 2. It gives a sheafified elliptic quantum group, as well as its action on $\mathfrak{M}(\mathfrak{w})$.

3.1 Equivariant Elliptic Cohomology

There is a sheaf-theoretic definition of the equivariant elliptic cohomology theory Ell_G in [11] by Ginzburg-Kapranov-Vasserot. It was investigated by many others later on, including Ando [2, 3], Chen [5], Gepner [10], Goerss-Hopkins [12], Lurie [16].

Let \mathcal{X}_G be the moduli scheme of semisimple semistable degree 0 G -bundles over an elliptic curve. For a G -variety X , the G -equivariant elliptic cohomology $Ell_G(X)$ of X is a quasi-coherent sheaf of $\mathcal{O}_{\mathcal{X}_G}$ -module, satisfying certain axioms. In particular, $Ell_G(\text{pt}) \cong \mathcal{O}_{\mathcal{X}_G}$.

Example 3

1. Let $G = S^1$, then $Ell_{S^1}(X)$ is a coherent sheaf on $\text{Pic}(E) \cong E$. This fact leads to the following patten.
 - $\text{CH}_{S^1}(X)$ is a module over $\text{CH}_{S^1}(\text{pt}) = \mathcal{O}_{\mathbb{C}}$.
 - $K_{S^1}(X)$ is a module over $K_{S^1}(\text{pt}) = \mathcal{O}_{\mathbb{C}^*}$.
 - $Ell_{S^1}(X)$ is a module over $Ell_{S^1}(\text{pt}) = \mathcal{O}_E$.
2. Let $G = \text{GL}_n$, then $Ell_{\text{GL}_n}^*(X)$ is a coherent sheaf over $E^{(n)} = E^n / \mathfrak{S}_n$.

There is a subtlety for pushforward in the equivariant elliptic cohomology theory. Let $f : X \rightarrow Y$ be a proper, G -equivariant homomorphism. The pushforward f_* is the following map

$$f_* : Ell_G(\text{Th}(f)) \rightarrow Ell_G(Y),$$

where $\mathcal{E}ll_G(\text{Th}(f))$, depending on the Thom bundle of the relative tangent bundle of f , is a rank 1 locally free module over $\mathcal{E}ll_G(X)$. The appearance of this twist can be illustrated in the following simple example. The general case is discussed in detail in [11, § 2] and [29, § 2].

Example 4 Let $f : \{0\} \hookrightarrow \mathbb{C}$ be the inclusion. The torus S^1 acts on \mathbb{C} by scalar multiplication. Denote by D the disc around 0. We have the Thom space $\text{Th} := D/S^1$. There is an exact sequence

$$0 \rightarrow \mathcal{E}ll_{S^1}(\text{Th}) \rightarrow \mathcal{E}ll_{S^1}(D) \rightarrow \mathcal{E}ll_{S^1}(S^1) \rightarrow 0.$$

As $\mathcal{E}ll_{S^1}(D) \cong \mathcal{O}_E$, since D is contractible, and $\mathcal{E}ll_{S^1}(S^1)$ is the skyscraper sheaf \mathbb{C}_0 at 0, we have the isomorphism $\mathcal{E}ll_{S^1}(\text{Th}) \cong \mathcal{O}(-\{0\})$.

3.2 The Sheafified Elliptic Quantum Group

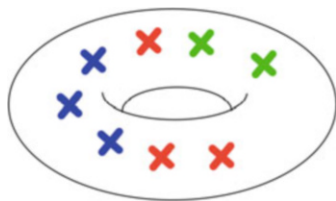
Recall the elliptic cohomological Hall algebra (see (1)) is defined as

$$\mathcal{P}(\mathcal{E}ll, Q) := \bigoplus_{\mathbf{v} \in \mathbb{N}^I} \mathcal{E}ll_{\text{GL}_{\mathbf{v}} \times \mathbb{C}^*}(\text{Rep}(\Pi_Q, \mathbf{v})).$$

By the discussion in Sect. 3.1, $\mathcal{P}(\mathcal{E}ll, Q)$ is a sheaf on

$$\mathcal{H}_{E \times I} \times E_{\hbar} := \coprod_{\{\mathbf{v}=(v^i)_{i \in I} \in \mathbb{N}^I\}} (E^{(v^1)} \times E^{(v^2)} \times \dots \times E^{(v^n)}) \times E_{\hbar},$$

where \hbar comes from the \mathbb{C}^* -action, and $\mathcal{H}_{E \times I}$ is the moduli space of I -colored points on E .



Due to the subtlety of pushing-forward in the elliptic cohomology, there is no product on $\mathcal{P}(\mathcal{E}ll, Q)$ in the usual sense. We illustrate the structure of the Hall multiplication \star of $\mathcal{P}(\mathcal{E}ll, Q)$ in the following example.

Example 5 Let Q be the quiver \bigcirc with one vertex, no arrows. In this case, we have $\mathfrak{g}_Q = \mathfrak{sl}_2$. The elliptic CoHA $\mathcal{P}(\mathcal{E}ll, \mathfrak{sl}_2)$ associated to the Lie algebra \mathfrak{sl}_2 consists of:

- A coherent sheaf $\mathcal{P}(\mathcal{E}ll, \mathfrak{sl}_2) = (\mathcal{P}_n)_{n \in \mathbb{N}}$ on $\mathcal{H}_E \times E_{\hbar} = \coprod_{n \in \mathbb{N}} E^{(n)} \times E_{\hbar}$.
- For any $n, m \in \mathbb{N}$, a morphism of sheaves on $E^{(n+m)} \times E_{\hbar}$:

$$\star : (\Sigma_{n,m})_*((\mathcal{P}_n \boxtimes \mathcal{P}_m) \otimes \mathcal{L}_{n,m}) \rightarrow \mathcal{P}_{n+m},$$

where $\Sigma_{n,m}$ the symmetrization map $E^{(n)} \times E^{(m)} \rightarrow E^{(n+m)}$, and $\mathcal{L}_{n,m}$ is some fixed line bundle on $E^{(n)} \times E^{(m)} \times E_{\hbar}$, depending on the parameter \hbar .

- The above morphisms are associative in the obvious sense.

Let \mathcal{C} be the category of coherent sheaves on $\mathcal{H}_{E \times I} \times E_{\hbar}$. Motivated by the Hall multiplication \star of $\mathcal{P}(\mathcal{E}ll, Q)$, we define a tensor structure \otimes_{\hbar} on \mathcal{C} : for $\{\mathcal{F}\}, \{\mathcal{G}\} \in \text{Obj}(\mathcal{C})$, $\mathcal{F} \otimes_{\hbar} \mathcal{G}$ is defined as

$$(\mathcal{F} \otimes_{\hbar} \mathcal{G})_{\mathbf{v}} := \bigoplus_{\mathbf{v}_1 + \mathbf{v}_2 = \mathbf{v}} (\Sigma_{\mathbf{v}_1, \mathbf{v}_2})_*((\mathcal{F}_{\mathbf{v}_1} \boxtimes \mathcal{G}_{\mathbf{v}_2}) \otimes \mathcal{L}_{\mathbf{v}_1, \mathbf{v}_2}). \tag{7}$$

Theorem B ([26, Yang-Zhao])

1. The category $(\mathcal{C}, \otimes_{\hbar})$ is a symmetric monoidal category, with the braiding given by Yang and Zhao [26, Theorem 3.3].
2. The elliptic CoHA $(\mathcal{P}(\mathcal{E}ll, Q), \star, \Delta)$, endowed with the Hall multiplication \star , and coproduct Δ , is a bialgebra object in $(\mathcal{C}^{loc}, \otimes_{\hbar})$.
3. The Drinfeld double $D(\mathcal{P}(\mathcal{E}ll, Q))$ of $\mathcal{P}(\mathcal{E}ll, Q)$ acts on $\mathcal{E}ll_{G_{\mathbf{w}}}(\mathfrak{M}(\mathbf{w}))$, for any $\mathbf{w} \in \mathbb{N}^I$.
4. After taking a certain space of meromorphic sections Γ_{rat} , the bialgebra $\Gamma_{\text{rat}}(D(\mathcal{P}_{\lambda}^{\text{sph}}(\mathcal{E}ll, Q)))$ becomes the elliptic quantum group given by the dynamical elliptic R-matrix of Felder [7], Gautam-Toledano Laredo [9].

Remark 3

1. In Theorem B(4), a version of the sheafified elliptic quantum group with dynamical twist $D(\mathcal{P}_{\lambda}^{\text{sph}}(\mathcal{E}ll, Q))$ is needed in order to recover the R-matrix of Felder-Gautam-Toledano Laredo. This twist is explained in detail in [26, § 10.2]. Below we illustrate the flavour of this dynamical twist in Sect. 3.4.

In particular, the abelian category of representations of $D(\mathcal{P}(\mathcal{E}ll, Q))$ and $D(\mathcal{P}_{\lambda}^{\text{sph}}(\mathcal{E}ll, Q))$ are both well-defined (see [26, § 9] for the details). Furthermore, it is proved in *loc. cit.* that these two representation categories are equivalent as abelian categories.

2. The details of the space of meromorphic sections are explained in [26, § 6].
3. Based on the above theorem, we define the sheafified elliptic quantum group to be the Drinfeld double of $\mathcal{P}^{\text{sph}}(\mathcal{E}ll, Q)$.

3.3 The Shuffle Formulas

The shuffle formula of the elliptic quantum group is given by (3), with the factor $A + B$ replaced by $\vartheta(A + B)$. The shuffle formula gives an explicit description of the elliptic quantum group, as well as its action on the Nakajima quiver varieties. In this section, we illustrate the shuffle description of the elliptic quantum group $E_{\tau, \hbar}(\mathfrak{sl}_2)$. Furthermore, we show (3) applied to special cases coincides with the shuffle formulas in [8, Proposition 3.6] and [13, Definition 5.9].

When $\mathfrak{g} = \mathfrak{sl}_2$, the corresponding quiver is \circ , with one vertex, no arrows. Let $\text{SH}_{\mathbf{w}=0} := \mathcal{O}_{\mathcal{H}_E \times E_{\hbar}} = (\bigoplus_{n \in \mathbb{N}} \mathcal{O}_{E^{(n)}}) \boxtimes \mathcal{O}_{E_{\hbar}}$. For any local sections $f \in \mathcal{O}_{E^{(n)}}$, $g \in \mathcal{O}_{E^{(m)}}$, by (3) and (6), we have

$$f \star g = \sum_{\sigma \in \text{Sh}(n, m)} \sigma \left(fg \prod_{1 \leq s \leq n, n+1 \leq t \leq n+m} \frac{\vartheta(x_s - x_t + \hbar)}{\vartheta(x_s - x_t)} \right),$$

where $f \star g \in \mathcal{O}_{E^{(n+m)}}$ and $\text{Sh}(n, m)$ consists of (n, m) -shuffles, i.e., permutations of $\{1, \dots, n+m\}$ that preserve the relative order of $\{1, \dots, n\}$ and $\{n+1, \dots, n+m\}$. The elliptic quantum group $E_{\tau, \hbar}^+(\mathfrak{sl}_2)$ is a subalgebra of $(\text{SH}_{\mathbf{w}=0}, \star)$.

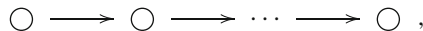
In the following examples we consider SH with general \mathbf{w} .

Example 6 Assume the quiver is \circ , with one vertex, no arrows. Let $\mathbf{v}_1 = k'$, $\mathbf{v}_2 = k''$, and $\mathbf{w}_1 = n'$, $\mathbf{w}_2 = n''$. Choose $t_1 = \hbar$, and $t_2 = 0$. Applying the formula (6) to this case, we have

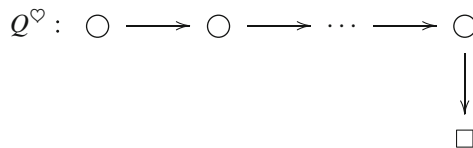
$$\text{fac}_{k'+k'', n'+n''} = \prod_{s=1}^{k'} \prod_{t=k'+1}^{k'+k''} \frac{\vartheta(\lambda_s - \lambda_t + \hbar)}{\vartheta(\lambda_t - \lambda_s)} \cdot \prod_{s=1}^{k'} \prod_{t=n'+1}^{n'+n''} \vartheta(z_t - \lambda_s + \hbar) \prod_{s=1}^{n'} \prod_{t=k'+1}^{k'+k''} \vartheta(\lambda_t - z_s),$$

where $\vartheta(z)$ is the odd Jacobi theta function, normalized such that $\vartheta'(0) = 1$. This is exactly the same formula as [8, Proposition 3.6].

Example 7 When $\mathfrak{g} = \mathfrak{sl}_N$, the corresponding quiver is



with $N - 1$ vertices, and $N - 2$ arrows. Consider the framed quiver



Label the vertices of Q^\heartsuit by $\{1, 2, \dots, N - 1, N\}$. Let $\mathbf{v}_1 = (v_1^{(l)})_{l=1, \dots, N-1}$, $\mathbf{v}_2 = (v_2^{(l)})_{l=1, \dots, N-1}$ be a pair of dimension vectors, and take the framing to be $\mathbf{w}_1 = (0, \dots, 0, n)$, $\mathbf{w}_2 = (0, \dots, 0, m)$. To simplify the notations, let $v_1^{(N)} = n$, $v_2^{(N)} = m$, and denote the variables $\mathbf{z}_{\mathbf{w}_1}$ by $\{\lambda_s^{(N)}\}_{s=1, \dots, v_1^{(N)}}$, and $\mathbf{z}_{\mathbf{w}_2}$ by $\{\lambda_t^{(N)}\}_{t=1, \dots, v_2^{(N)}}$. Applying the formula (6) to this case, we then have

$$\begin{aligned} & \text{fac}_{\mathbf{v}_1+\mathbf{v}_2, \mathbf{w}_1+\mathbf{w}_2} \\ &= \prod_{l=1}^{N-1} \left(\prod_{s=1}^{v_1^{(l)}} \prod_{t=1}^{v_2^{(l)}} \frac{\vartheta(\lambda_s^{(l)} - \lambda_t^{(l)} + t_1 + t_2)}{\vartheta(\lambda_t^{(l)} - \lambda_s^{(l)})} \right. \\ & \quad \cdot \prod_{s=1}^{v_1^{(l)}} \prod_{t=1}^{v_2^{(l+1)}} \vartheta(\lambda_t^{(l+1)} - \lambda_s^{(l)} + t_1) \prod_{t=1}^{v_1^{(l+1)}} \prod_{s=1}^{v_2^{(l)}} \vartheta(\lambda_s^{(l)} - \lambda_t^{(l+1)} + t_2) \left. \right). \end{aligned}$$

Following [13, §5 (5.3)], we denote by $H_{\mathbf{v}+\mathbf{w}}(\lambda_{\mathbf{v}}, \mathbf{z}_{\mathbf{w}})$ the following element

$$H_{\mathbf{v}+\mathbf{w}}(\lambda_{\mathbf{v}}, \mathbf{z}_{\mathbf{w}}) = \prod_{l=1}^{N-1} \prod_{s=1}^{v^{(l)}} \prod_{t=1}^{v^{(l+1)}} \vartheta(\lambda_t^{(l+1)} - \lambda_s^{(l)} + t_1).$$

Define $H_{\text{cross}} = \frac{H_{\mathbf{v}_1+\mathbf{v}_2+\mathbf{w}_1+\mathbf{w}_2}(\lambda_{\mathbf{v}_1 \cup \mathbf{v}_2}, \mathbf{z}_{\mathbf{w}_1 \cup \mathbf{w}_2})}{H_{\mathbf{v}_1+\mathbf{w}_1}(\lambda_{\mathbf{v}_1}, \mathbf{z}_{\mathbf{w}_1}) \cdot H_{\mathbf{v}_2+\mathbf{w}_2}(\lambda_{\mathbf{v}_2}, \mathbf{z}_{\mathbf{w}_2})}$. We have

$$\begin{aligned} & H_{\text{cross}} \\ &= \prod_{l=1}^{N-1} \left(\prod_{s=1}^{v_1^{(l)}} \prod_{t=1}^{v_2^{(l+1)}} \vartheta(\lambda_t^{(l+1)} - \lambda_s^{(l)} + t_1) \prod_{t=1}^{v_1^{(l+1)}} \prod_{s=1}^{v_2^{(l)}} \vartheta(\lambda_t^{(l+1)} - \lambda_s^{(l)} + t_1) \right) \\ &= \prod_{l=1}^{N-1} \left(\prod_{s=1}^{v_1^{(l)}} \prod_{t=v_1^{(l+1)}+1}^{v_1^{(l+1)}+v_2^{(l+1)}} \vartheta(\lambda_t^{(l+1)} - \lambda_s^{(l)} + t_1) \prod_{t=1}^{v_1^{(l+1)}} \prod_{s=v_1^{(l)}+1}^{v_1^{(l)}+v_2^{(l)}} \vartheta(\lambda_t^{(l+1)} - \lambda_s^{(l)} + t_1) \right). \end{aligned} \tag{8}$$

Divide $\text{fac}_{\mathbf{v}_1+\mathbf{v}_2, \mathbf{w}_1+\mathbf{w}_2}$ by H_{cross} , we obtain

$$\begin{aligned} \frac{\text{fac}_{\mathbf{v}_1+\mathbf{v}_2, \mathbf{w}_1+\mathbf{w}_2}}{H_{\text{cross}}} &= (-1)^{v_1^{(l+1)}+v_2^{(l)}} \prod_{l=1}^{N-1} \left(\prod_{s=1}^{v_1^{(l)}} \prod_{t=1}^{v_2^{(l)}} \frac{\vartheta(\lambda_s^{(l)} - \lambda_t^{(l)} + t_1 + t_2)}{\vartheta(\lambda_t^{(l)} - \lambda_s^{(l)})} \right. \\ & \quad \cdot \prod_{t=1}^{v_1^{(l+1)}} \prod_{s=1}^{v_2^{(l)}} \frac{\vartheta(\lambda_t^{(l+1)} - \lambda_s^{(l)} - t_2)}{\vartheta(\lambda_t^{(l+1)} - \lambda_s^{(l)} + t_1)} \left. \right). \end{aligned}$$

This coincides with the formula in [13, Definition 5.9] when $t_1 = -1, t_2 = 0$.

In other words, consider the map

$$\bigoplus_{\mathbf{v} \in \mathbb{N}^l, \mathbf{w} \in \mathbb{N}^l} \text{SH}_{\mathbf{v}, \mathbf{w}} \rightarrow \bigoplus_{\mathbf{v} \in \mathbb{N}^l, \mathbf{w} \in \mathbb{N}^l} (\text{SH}_{\mathbf{v}, \mathbf{w}})_{\text{loc}}, \text{ given by}$$

$$f(\lambda_{\mathbf{v}}, z_{\mathbf{w}}) \mapsto \frac{f(\lambda_{\mathbf{v}}, z_{\mathbf{w}})}{H_{\mathbf{v}+\mathbf{w}}(\lambda_{\mathbf{v}}, z_{\mathbf{w}})}.$$

It intertwines the shuffle product (3) (with theta-functions), and the shuffle product of [13, Definition 5.9].

Remark 4

1. In the work of [8] in the \mathfrak{sl}_2 case, and [13] in the \mathfrak{sl}_N case, the shuffle formulas are used to obtain an inductive formula for the elliptic weight functions. Proposition 1 provides a way to define the elliptic weight functions associated to a general symmetric Kac-Moody Lie algebra \mathfrak{g} .
2. In the above cases for \mathfrak{sl}_N and the special framing, the elliptic weight functions are expected to be related to the elliptic stable basis in [1] (see also [8, 13]). Therefore, it is reasonable to expect an expression of the elliptic stable basis in terms of the shuffle product (3) (with theta-functions) for general quiver varieties.

3.4 Drinfeld Currents

We now explain the Drinfeld currents in the elliptic case. The choice of an elliptic curve E gives rise to the dynamical elliptic quantum group.

Let $\mathcal{M}_{1,2}$ be the open moduli space of 2 pointed genus 1 curves. We write a point in $\mathcal{M}_{1,2}$ as (E_τ, λ) , where $E_\tau = \mathbb{C}/\mathbb{Z} \oplus \tau\mathbb{Z}$, and λ gives a line bundle $\mathbb{L}_\lambda \in \text{Pic}(E_\tau)$. Let E be the universal curve on $\mathcal{M}_{1,2}$. There is a Poincare line bundle \mathbb{L} on E , which has a natural rational section

$$\frac{\vartheta(z + \lambda)}{\vartheta(z)\vartheta(\lambda)}$$

where z is the coordinate of E_τ , and $\vartheta(z)$ is the Jacobi odd theta function, normalized such that $\vartheta'(0) = 1$.

We can twist the equivariant elliptic cohomology $\mathcal{E}ll_G$ by the Poincare line bundle \mathbb{L} . For each simple root e_k , after twisting, we have $\text{SH}_{e_k}^{\mathbb{L}} = \mathcal{O}_{E^{(e_k)}} \otimes \mathbb{L}$. A basis of the meromorphic sections $\Gamma_{\text{rat}}(\text{SH}_{e_k}^{\mathbb{L}})$ consists of $\left\{ g_{\lambda_k}^{(i)}(z_k) := \frac{\partial^i}{\partial z_k^i} \left(\frac{\vartheta(z_k + \lambda_k)}{\vartheta(z_k)\vartheta(\lambda_k)} \right) \right\}_{i \in \mathbb{N}}$.

Consider $\lambda = (\lambda_k)_{k \in I}$, and let

$$\mathfrak{X}_k^+(u, \lambda) := \sum_{i=0}^{\infty} g_{\lambda_k}^{(i)}(z_k) u^i = \frac{\vartheta(z_k + \lambda_k + u)}{\vartheta(z_k + u)\vartheta(\lambda_k)} \in \Gamma_{\text{rat}}(\text{SH}^{\mathbb{L}})[[u]], \quad k \in I.$$

Similarly, we define series $\mathfrak{X}_k^-(u, \lambda)$, $\Phi_k(u)$. The series $\mathfrak{X}_k^+(u, \lambda)$, $\mathfrak{X}_k^-(u, \lambda)$, and $\Phi_k(u)$ satisfy the relations of the elliptic quantum group of Gautam-Toledano Laredo [9].

4 Relation with the Affine Grassmannian

In this section, we explain one unexpected feature of the sheafified elliptic quantum group, namely, its relation with the global loop Grassmannians over an elliptic curve E . We assume the quiver $Q=(I, H)$ is of type ADE in this section.

We collect some facts from [19, 20]. Let C be a curve. An I -colored local space Z over C , defined in [19, Section 2], see also [20, Section 4.1], is a space $Z \rightarrow \mathcal{H}_{C \times I}$ over the I -colored Hilbert scheme of points of C , together with a consistent system of isomorphisms

$$Z_D \times Z_{D'} \rightarrow Z_{D \sqcup D'}, \quad \text{for } D, D' \in \mathcal{H}_{C \times I} \text{ with } D \cap D' = \emptyset.$$

Similarly, for a coherent sheaf \mathcal{F} on $\mathcal{H}_{C \times I}$, a locality structure of \mathcal{F} on $\mathcal{H}_{C \times I}$ is a compatible system of identifications

$$\iota_{D, D'} : \mathcal{F}_D \boxtimes \mathcal{F}_{D'} \cong \mathcal{F}_{D \sqcup D'}, \quad \text{for } D, D' \in \mathcal{H}_{C \times I} \text{ with } D \cap D' = \emptyset.$$

An example of local spaces is the zastava space $\mathcal{Z} \rightarrow \mathcal{H}_{C \times I}$, recollected in detail in [19, Section 3] and [20, Section 4.2.3]. Here C is a smooth curve. In Mirković (personal communication), a modular description of \mathcal{Z} is given along the lines of Drinfeld’s compactification. Let G be the semisimple simply-connected group associated to Q , with the choice of opposite Borel subgroups $B = TN$ and $B^- = TN^-$ with the joint Cartan subgroup T . Consider the Drinfeld’s compactification \mathcal{Y}_G of a point:

$$\mathcal{Y}_G = G \backslash [(G/N^+)^\text{aff} \times (G/N^-)^\text{aff}] / T.$$

The zastava space $\mathcal{Z} \rightarrow \mathcal{H}_{C \times I}$ for G is defined as the moduli of generic maps from C to \mathcal{Y}_G . Gluing the zastava spaces, one get a loop Grassmannian $\mathcal{G}r$ as a local space over $\mathcal{H}_{C \times I}$, which is a refined version of the Beilinson-Drinfeld Grassmannian, see [19, Section 3] and [20, Section 4.2.3].

Fix a point $c \in C$, and a dimension vector $\mathbf{v} \in \mathbb{N}^I$, let $\{c\} \in C^{(\mathbf{v})} \subseteq \mathcal{H}_{C \times I}$ be the special divisor supported on $\{c\}$. The fiber $\mathcal{Z}_{\{c\}}$ is the Mirković-Vilonen scheme

associated to the root vector \mathbf{v} , i.e., the intersection of closures of certain semi-infinite orbits in $G_{\mathbb{C}((z))}/G_{\mathbb{C}[[z]]}$ [19, Section 3].

The maximal torus $T \subset G$ acts on \mathcal{Z} . There is a certain component in a torus-fixed loci $(\mathcal{Z})^T$, which gives a section $\mathcal{H}_{C \times I} \subset (\mathcal{Z})^T$. We denote this component by $(\mathcal{Z})^{T^\circ}$. The tautological line bundle $\mathcal{O}_{\mathcal{G}_r}(1)|_{\mathcal{Z}^{T^\circ}}$ has a natural locality structure, and is described in [19, Theorem 3.1].

We now take the curve C to be the elliptic curve E . Let $\mathcal{G}_r \rightarrow \mathcal{H}_{E \times I}$ be the global loop Grassmannian over $\mathcal{H}_{E \times I}$. The following theorem relies on the description of the local line bundle on $\mathcal{H}_{C \times I}$ in [19] and [20, Section 4.2.1].

Theorem C ([26] Yang-Zhao)

1. The classical limit $\mathcal{P}^{\text{sph}}(\mathcal{E}ll, Q)|_{\hbar=0}$ is isomorphic to $\mathcal{O}_{\mathcal{G}_r}(-1)|_{\mathcal{Z}^{T^\circ}}$ as sheaves on $\mathcal{H}_{E \times I}$.
2. The Hall multiplication \star on $\mathcal{P}^{\text{sph}}(\mathcal{E}ll, Q)|_{\hbar=0}$ is equivalent to the locality structure on $\mathcal{O}_{\mathcal{G}_r}(1)|_{\mathcal{Z}^{T^\circ}}$.

Remark 5

1. Theorem C is true when the curve E is replaced by \mathbb{C} (and \mathbb{C}^*), while the corresponding cohomological Hall algebra is modified to $\mathcal{P}^{\text{sph}}(\text{CH}, Q)|_{\hbar=0}$ (and $\mathcal{P}^{\text{sph}}(K, Q)|_{\hbar=0}$ respectively). The sheaf $\mathcal{P}^{\text{sph}}(\mathcal{E}ll, Q)$ deforms the local line bundle $\mathcal{O}_{\mathcal{G}_r}(-1)|_{\mathcal{Z}^{T^\circ}}$. In the classification of local line bundles in [19], $\mathcal{O}_{\mathcal{G}_r}(1)|_{\mathcal{Z}^{T^\circ}}$ is characterized by certain diagonal divisor of $\mathcal{H}_{E \times I}$ [20, Section 4.2.1]. As a consequence of Theorem C, the shuffle formula of $\mathcal{P}^{\text{sph}}(\mathcal{E}ll, Q)$ gives the \hbar -shifting of the diagonal divisor of $\mathcal{H}_{E \times I}$ that appears in $\mathcal{O}_{\mathcal{G}_r}(1)|_{\mathcal{Z}^{T^\circ}}$.
2. When the base is $\mathcal{H}_{C \times I}$, and cohomology theory is the Borel-Moore homology H_{BM} , Theorem C (1) has a similar flavour as [4, Theorem 3.1]. Here, we only consider $\mathcal{P}^{\text{sph}}(H_{\text{BM}}, Q)|_{\hbar=0}$ and $\mathcal{O}_{\mathcal{G}_r}(-1)|_{\mathcal{Z}^{T^\circ}}$ as sheaves of abelian groups. By Theorem A, $(\mathcal{P}^{\text{sph}}(H_{\text{BM}}, Q), \star)$ is isomorphic to the positive part of the Yangian, which is in turn related to $\mathcal{O}_{\mathcal{G}_r}(-1)|_{\mathcal{Z}^{T^\circ} \rightarrow \mathcal{H}_{C \times I}}$ by Theorem C.

Acknowledgements Y.Y. would like to thank the organizers of the MATRIX program *Geometric R-Matrices: from Geometry to Probability* for their kind invitation, and many participants of the program for useful discussions, including Vassily Gorbounov, Andrei Okounkov, Allen Knutson, Hitoshi Konno, Paul Zinn-Justin. Proposition 1 and Sect. 3.3 are new, for which we thank Hitoshi Konno for interesting discussions and communications. These notes were written when both authors were visiting the Perimeter Institute for Theoretical Physics (PI). We are grateful to PI for the hospitality.

References

1. Aganagic, M., Okounkov, A.: Elliptic stable envelopes (2016, preprint). [arXiv:1604.00423](https://arxiv.org/abs/1604.00423)
2. Ando, M.: Power operations in elliptic cohomology and representations of loop groups. *Trans. Am. Math. Soc.* **352**(12), 5619–5666 (2000). [MR1637129](https://doi.org/10.2307/2688719)

3. Ando, M.: The Sigma-orientation for circle-equivariant elliptic cohomology. *Geom. Topol.* **7**, 91–153 (2003). [MR1988282](#)
4. Braverman, A., Finkelberg, M., Nakajima, H.: Coulomb branches of $3d$, $\mathcal{N} = 4$ quiver gauge theories and slices in the affine Grassmannian (with appendices by Alexander Braverman, Michael Finkelberg, Joel Kamnitzer, Ryosuke Kodera, Hiraku Nakajima, Ben Webster, and Alex Weekes). [arXiv:1604.03625](#)
5. Chen, H.-Y.: Torus equivariant elliptic cohomology and sigma orientation. Ph.D. Thesis, University of Illinois at Urbana-Champaign, 109pp. (2010). [MR2873496](#)
6. Drinfeld, V.: Quasi-Hopf algebras. *Algebra i Analiz* **1**(6), 114–148 (1989). [MR1047964](#)
7. Felder, G.: Elliptic quantum groups. In: XIth International Congress of Mathematical Physics (Paris, 1994), pp. 211–218. Int. Press, Cambridge (1995). [MR1370676](#)
8. Felder, G., Rimanyi, R., Varchenko, A.: Elliptic dynamical quantum groups and equivariant elliptic cohomology (2017, preprint). [arXiv:1702.08060](#)
9. Gautam, S., Toledano Laredo, V.: Elliptic quantum groups and their finite-dimensional representations (2017, preprint). [arXiv:1707.06469](#)
10. Gepner, D.: Equivariant elliptic cohomology and homotopy topoi. Ph.D. thesis, University of Illinois (2006)
11. Ginzburg, V., Kapranov, M., Vasserot, E.: Elliptic algebras and equivariant elliptic cohomology. (1995, Preprint). [arXiv:9505012](#)
12. Goerss, P., Hopkins, M.: Moduli spaces of commutative ring spectra. In: *Structured Ring Spectra*, London Mathematical Society Lecture Note Series, vol. 315, pp. 151–200. Cambridge University Press, Cambridge (2004)
13. Konno, H.: Elliptic weight functions and elliptic q-KZ equation. *J. Integr. Syst.* **2**(1), xyx011 (2017)
14. Kontsevich, M., Soibelman, Y.: Cohomological Hall algebra, exponential Hodge structures and motivic Donaldson-Thomas invariants. *Commun. Number Theory Phys.* **5**(2), 231–352 (2011). [MR2851153](#)
15. Levine, M., Morel, F.: *Algebraic Cobordism Theory*. Springer, Berlin (2007). [MR2286826](#)
16. Lurie, J.: A survey of elliptic cohomology. In: *Algebraic Topology*. Abel Symposia, vol. 4, pp. 219–277. Springer, Berlin (2009). [MR2597740](#)
17. Maulik, D., Okounkov, A.: Quantum groups and quantum cohomology (preprint). [arXiv:1211.1287v1](#)
18. McGerty, K., Nevins, T.: Kirwan surjectivity for quiver varieties. *Invent. Math.* **212**, 161–187 (2018)
19. Mirkovic, I.: The loop Grassmannians in the framework of local spaces over a curve . In: *Recent Advances in Representation Theory, Quantum Groups, Algebraic Geometry, and Related Topics*. Contemporary Mathematics, vol. 623, pp. 215–226. American Mathematical Society, Providence (2014). [MR3288629](#)
20. Mirković, I.: Some extensions of the notion of loop Grassmannians. *Rad Hrvat. Akad. Znan. Umjet. Mat. Znan.* **21**(532), 53–74 (2017). [MR3697895](#)
21. Nakajima, H.: Quiver varieties and finite dimensional representations of quantum affine algebras. *J. Am. Math. Soc.* **14**(1), 145–238 (2001). [MR1808477](#)
22. Okounkov, A., Smirnov, A.: Quantum difference equation for Nakajima varieties. [arXiv:1602.09007](#)
23. Schiffmann, O., Vasserot, E.: The elliptic Hall algebra and the K-theory of the Hilbert scheme of \mathbb{A}^2 . *Duke Math. J.* **162**(2), 279–366 (2013). [MR3018956](#)
24. Varagnolo, M.: Quiver varieties and yangians. *Lett. Math. Phys.* **53**(4), 273–283 (2000). [MR1818101](#)
25. Yang, Y., Zhao, G.: On two cohomological Hall algebras. *Proc. R. Soc. Edinb. Sect. A* (to appear). [arXiv:1604.01477](#)
26. Yang, Y., Zhao, G.: Quiver varieties and elliptic quantum groups (2017, preprint). [arxiv1708.01418](#)

27. Yang, Y., Zhao, G.: Cohomological Hall algebras and affine quantum groups. *Sel. Math.* **24**(2), 1093–1119 (2018). [arXiv:1604.01865](#)
28. Yang, Y., Zhao, G.: The cohomological Hall algebra of a preprojective algebra. *Proc. Lond. Math. Soc.* **116**, 1029–1074. [arXiv:1407.7994](#)
29. Zhao, G., Zhong, C.: Elliptic affine Hecke algebra and its representations (2015, preprint). [arXiv:1507.01245](#)