LECTURE NOTES IN COMPUTATIONAL SCIENCE AND ENGINEERING 102

Stephan Dahlke · Wolfgang Dahmen
Michael Griebel · Wolfgang Hackbusch
Klaus Ritter · Reinhold Schneider
Christoph Schwab · Harry Yserentant *Editors*

Extraction of Quantifiable Information from Complex Systems

Springer

Lecture Notes
in Computational Science
and Engineering

102

Editors:

More information about this series at
http://www.springer.com/series/3527

Stephan Dahlke • Wolfgang Dahmen •
Michael Griebel • Wolfgang Hackbusch •
Klaus Ritter • Reinhold Schneider •
Christoph Schwab • Harry Yserentant
Editors

Extraction of Quantifiable Information from Complex Systems

Springer

Editors
Stephan Dahlke
Fachbereich Mathematik und Informatik
Philipps-Universität Marburg
Marburg, Germany

Wolfgang Dahmen
Institut für Geometrie und Praktische Mathematik
RWTH Aachen
Aachen, Germany

Michael Griebel
Institut für Numerische Simulation
Universität Bonn
Bonn, Germany

Wolfgang Hackbusch
Max-Planck-Institut für Mathematik in den Naturwissenschaften
Leipzig, Germany

Klaus Ritter
Fachbereich Mathematik
Technische Universität Kaiserslautern
Kaiserslautern, Germany

Reinhold Schneider
Institut für Mathematik
Technische Universität Berlin
Berlin, Germany

Christoph Schwab
Seminar für Angewandte Mathematik
ETH Zürich
Zürich, Switzerland

Harry Yserentant
Institut für Mathematik
Technische Universität Berlin
Berlin, Germany

ISSN 1439-7358 ISSN 2197-7100 (electronic)
ISBN 978-3-319-08158-8 ISBN 978-3-319-08159-5 (eBook)
DOI 10.1007/978-3-319-08159-5
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2014954584

Mathematics Subject Classification (2010): 35-XX, 41-XX, 60-XX, 65-XX

Printed on acid-free paper

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

Foreword

In April 2007, the Deutsche Forschungsgemeinschaft (DFG) approved the Priority Program 1324 "Mathematical Methods for Extracting Quantifiable Information from Complex Systems". The objective of this volume is to offer a comprehensive overview of the scientific highlights obtained in the course of this priority program.

Mathematical models of complex systems are gaining rapidly increasing importance in driving fundamental developments in various fields such as science and engineering at large but also in new areas such as computational finance. Ever-increasing hardware capacities and computing power encourage and foster the development of more and more realistic models. On the other hand, the necessarily growing complexity of such models keeps posing serious and even bigger challenges to their numerical treatment.

Principal obstructions such as the *curse of dimensionality* suggest that a proper response to these challenges cannot be based solely on further increasing computing power. Instead, recent developments in mathematical sciences indicate that significant progress can only be achieved by contriving novel and much more powerful numerical solution strategies by systematically exploiting synergies and conceptual interconnections between the various relevant research areas. Needless to stress that this requires a deeper understanding of the mathematical foundations as well as exploring new and efficient algorithmic concepts. Fostering such well-balanced developments has been a central objective of this priority program.

The understanding and numerical treatment of spatially high-dimensional systems is clearly one of the most challenging tasks in applied mathematics. The problem of spatial high dimensionality is encountered in numerous application contexts such as machine learning, design of experiments, parameter-dependent models and their optimization, mathematical finance, PDEs in high-dimensional phase space, to name only a few, which already reflect the conceptual breadth. It is this seeming variability that makes a substantial impact of better exploiting conceptual and methodological synergies conceivable and in fact likely. It seems

that to be really successful, theoretical research and practical applications should go hand in hand. In fact, this volume reflects an attempt to realize a proper balance between research with a primary methodological focus and challenging concrete application areas, although these two regimes can, of course, not be strictly separated. To that end, it has appeared to be necessary to combine different fields of mathematics such as numerical analysis and computational stochastics. On the other hand, to keep the whole programme sufficiently focused, it seemed advisable to concentrate on specific but related fields of application that share some common characteristics that allow one to benefit from conceptual similarities.

On the methodological side, several important new numerical approximation methods have been developed and/or further investigated in the course of the priority program. First of all, as one of the central techniques, let us mention tensor approximations. New tensor formats have been developed, and efficient tensor approximation schemes for various applications, e.g. in quantum dynamics and computational finance, have been studied; see Chaps. 2, 10, 12, 16 and 19. Adaptive strategies with all their facets have been employed in most of the projects; see, e.g., Chaps. 2, 4, 5, 9, 10, 14 and 16. Closely related with adaptivity is of course the concept of sparsity/compressed sensing; see Chaps. 14 and 18. As further techniques, sparse grids (Chap. 9), ANOVA decompositions (Chap. 11) and Fourier methods (Chap. 17) have been investigated. As a quite new technique, the reduced basis methods also came into play (see Chap. 2), in particular in the second period of SPP 1324. Of course, tensor methods as well as model order reduction concepts such as the reduced basis method address spatially high-dimensional problems. Both paradigms use the separation of variables as the central means to reduce computational complexity. Moreover, they can be viewed as trying to exploit sparsity by determining specific problem- and solution-dependent dictionaries that are able to approximate the searched object by possibly few terms. Moreover, Chaps. 1, 6 and 20 are concerned with Monte Carlo and Multilevel Monte Carlo methods in the context of stochastic applications.

One of the major themes within SPP 1324 has been high-dimensional problems in physics. Chapter 21 is concerned with the regularity of the solution to the electronic Schrödinger equation. Chapter 19 studies problems in quantum dynamics, the chemical master equation is one of the topics in Chap. 15, and Chap. 11 is concerned with electronic structure problems. Another very important issue within SPP 1324 has been differential equations with random or parameter-dependent coefficients and their various applications. The theory and numerical treatment of these problems are discussed in Chaps. 2 and 7. Closely related with this topic are stochastic differential equations and stochastic partial differential equations. The adaptive numerical treatment of SPDEs is studied in Chap. 5. SDEs with their various applications such as stochastic filtering are discussed in Chaps. 1, 6 and 8. Additional fields of application have been computational finance (see Chap. 16) and inverse problems (see Chaps. 3 and 18).

Overall, the network of SPP 1324 comprised more than 60 scientists, and 20 projects were funded in two periods. Up to now, more than 170 papers have been published by the participants of SPP 1324. The aim of this volume is of course not

to give a complete presentation of all these results but rather to collect the scientific highlights in order to demonstrate the impact of SPP 1324 on further researches. The editors and authors hope that this volume will arouse interest in the reader in the various new mathematical concepts and numerical algorithms that have been developed in the priority program. For further information concerning SPP 1324, please visit http://www.dfg-spp1324.de/.

Marburg, Germany — Stephan Dahlke
Aachen, Germany — Wolfgang Dahmen
Bonn, Germany — Michael Griebel
Leipzig, Germany — Wolfgang Hackbusch
Kaiserslautern, Germany — Klaus Ritter
Berlin, Germany — Reinhold Schneider
Zürich, Switzerland — Christoph Schwab
Berlin, Germany — Harry Yserentant
June 2014

Acknowledgements

First of all, the authors and the editors of this book thank the Deutsche Forschungsgemeinschaft (DFG) for the support within the DFG Priority Program 1324 “Extraction of Quantifiable Information from Complex Systems”. Moreover, we thank the referees Folkmar Bornemann, Joachim Buhmann, Hans Georg Feichtinger, Ursula Gather, Markus Hegland, Des Higham, George Karniadakis, Claudia Klüppelberg, Stig Larsson, Claude Le Bris, Gilles Pagès, Otmar Scherzer, Ian H. Sloan and Endre Süli for their hard work and conscientious refereeing. We also feel very grateful to our SPP fellows Kyeong-Hun Kim, Mihály Kovács, Stig Larsson, Kijung Lee, Raul Tempone and Henryk Woźniakowski, who contributed a lot to the success of SPP 1324. Special thanks are devoted to the DFG representatives Frank Kiefer and Carsten Balleier for the very productive cooperation. Last but not least, we thank Frank Eckhardt for assistance and support during the production of this book.

Contents

1 Solving Stochastic Dynamic Programs by Convex Optimization and Simulation 1
Denis Belomestny, Christian Bender, Fabian Dickmann, and Nikolaus Schweizer
1.1 Introduction 1
1.2 The Primal-Dual Approach to Convex Dynamic Programs 3
1.3 Construction of Lower Bounds via Martingale Basis Functions ... 8
1.4 Construction of Upper Bounds: Multilevel Monte Carlo and Sieve Optimization 11
1.5 Numerical Experiments 17
References 22

2 Efficient Resolution of Anisotropic Structures 25
Wolfgang Dahmen, Chunyan Huang, Gitta Kutyniok, Wang-Q Lim, Christoph Schwab, and Gerrit Welper
2.1 Introduction 26
2.2 Anisotropic Approximations 28
2.3 Well-Conditioned Stable Variational Formulations 30
2.4 Reduced Basis Methods 41
2.5 Sparse Tensor Approximation for Radiative Transfer 46
References 50

3 Regularity of the Parameter-to-State Map of a Parabolic Partial Differential Equation 53
Rudolf Ressel, Patrick Dülk, Stephan Dahlke, Kamil S. Kazimierski, and Peter Maass
3.1 Introduction 53
3.2 Function Spaces 55

3.3 The Model PDE as an Evolution Equation 56
3.4 The Parameter-to-State Map 61
References 67

4 Piecewise Tensor Product Wavelet Bases by Extensions and Approximation Rates 69
Nabi G. Chegini, Stephan Dahlke, Ulrich Friedrich, and Rob Stevenson
4.1 Introduction 69
4.2 Approximation by Tensor Product Wavelets on the Hypercube ... 71
4.3 Construction of Riesz Bases by Extension 73
4.4 Approximation by – Piecewise – Tensor Product Wavelets 76
4.5 Numerical Results 79
References 80

5 Adaptive Wavelet Methods for SPDEs 83
Petru A. Cioica, Stephan Dahlke, Nicolas Döhring, Stefan Kinzel, Felix Lindner, Thorsten Raasch, Klaus Ritter, and René L. Schilling
5.1 Introduction 84
5.2 Preliminaries 85
5.3 Regularity Analysis in Besov Spaces 91
5.4 Nonlinear Approximation for Elliptic Equations 97
References 106

6 Constructive Quantization and Multilevel Algorithms for Quadrature of Stochastic Differential Equations 109
Martin Altmayer, Steffen Dereich, Sangmeng Li, Thomas Müller-Gronbach, Andreas Neuenkirch, Klaus Ritter, and Larisa Yaroslavtseva
6.1 Introduction 110
6.2 Constructive Quantization of Systems of SDEs 111
6.3 Multilevel Methods for Discontinuous Payoffs in the Generalized Heston Model 118
6.4 Multilevel Methods for Lévy-Driven SDEs 124
References 130

7 Bayesian Inverse Problems and Kalman Filters 133
Oliver G. Ernst, Björn Sprungk, and Hans-Jörg Starkloff
7.1 Introduction 133
7.2 Bayesian Approach to Inverse Problems 135
7.3 Analysis of Kalman Filters for Bayesian Inverse Problems 146
7.4 Numerical Example: 1D Elliptic Boundary Value Problem 152
7.5 Conclusions 156
References 157

8 Robustness in Stochastic Filtering and Maximum Likelihood Estimation for SDEs 161
Joscha Diehl, Peter K. Friz, Hilmar Mai, Harald Oberhauser, Sebastian Riedel, and Wilhelm Stannat
8.1 Introduction 162
8.2 Robustness of the Stochastic Filter 165
8.3 Maximum Likelihood Estimation for SDEs 175
8.4 Practical Implications 177
References 177

9 Adaptive Sparse Grids in Reinforcement Learning 179
Jochen Garcke and Irene Klompmaker
9.1 Introduction 179
9.2 Reinforcement Learning 181
9.3 Planning with Sparse Grids 184
9.4 Sparse Grid Based Scheme for Reinforcement Learning 190
9.5 Experiments 191
9.6 Conclusion 192
References 193

10 A Review on Adaptive Low-Rank Approximation Techniques in the Hierarchical Tensor Format 195
Jonas Ballani, Lars Grasedyck, and Melanie Kluge
10.1 Introduction 195
10.2 Low-Rank Tensor Representations 197
10.3 Adaptive Tensor Sampling 202
10.4 Non-adaptive Tensor Sampling 206
10.5 Tensor Completion 207
References 209

11 A Bond Order Dissection ANOVA Approach for Efficient Electronic Structure Calculations 211
Michael Griebel, Jan Hamaekers, and Frederik Heber
11.1 Introduction 212
11.2 Schrödinger Equation in the Born-Oppenheimer Approximation 215
11.3 ANOVA Decomposition Scheme 216
11.4 Numerical Results 225
Concluding Remarks 232
References 233

12 Tensor Spaces and Hierarchical Tensor Representations 237
Wolfgang Hackbusch and Reinhold Schneider
12.1 Introduction 237
12.2 Subspace Approximation and Tucker Format 240
12.3 Hierarchical Tensor Representations 243
12.4 Hierarchical Tensors as Differentiable Manifolds 248

12.5 Numerical Methods 251
12.6 Tensorisation 256
References 259

13 Nonlinear Eigenproblems in Data Analysis: Balanced Graph Cuts and the RatioDCA-Prox 263
Leonardo Jost, Simon Setzer, and Matthias Hein
13.1 Introduction 263
13.2 Exact Relaxation of Balanced Graph Cuts 264
13.3 Minimization of Ratios of Non-negative Differences of Convex Functions via the RatioDCA-Prox 265
13.4 The RatioDCA-Prox for Ratios of Lovasz Extensions: Application to Balanced Graph Cuts 273
13.5 Experiments 277
References 278

14 Adaptive Approximation Algorithms for Sparse Data Representation 281
Mijail Guillemard, Dennis Heinen, Armin Iske, Sara Krause-Solberg, and Gerlind Plonka
14.1 Introduction 281
14.2 The Easy Path Wavelet Transform 282
14.3 Dimensionality Reduction on High-Dimensional Signal Data 292
14.4 Audio Signal Separation and Signal Detection 295
References 301

15 Error Bound for Hybrid Models of Two-Scaled Stochastic Reaction Systems 303
Tobias Jahnke and Vikram Sunkara
15.1 Introduction 303
15.2 The Chemical Master Equation of Two-scale Reaction Systems 304
15.3 Model Reduction Based on Conditional Expectations 307
15.4 Error Analysis for the Hybrid Model 309
References 318

16 Valuation of Structured Financial Products by Adaptive Multiwavelet Methods in High Dimensions 321
Rüdiger Kiesel, Andreas Rupp, and Karsten Urban
16.1 Introduction 321
16.2 Variational Formulation 324
16.3 Multiwavelets 326
16.4 Discretization 328
16.5 The Hierarchical Tucker Format (HTF) 330
16.6 Numerical Experiments 335
16.7 The Kronecker Product 342
References 343

17 Computational Methods for the Fourier Analysis of Sparse High-Dimensional Functions 347
Lutz Kämmerer, Stefan Kunis, Ines Melzer, Daniel Potts, and Toni Volkmer
17.1 Introduction 347
17.2 Evaluation of Multivariate Trigonometric Polynomials 348
17.3 Reconstruction Using Multivariate Trigonometric Polynomials ... 352
References 360

18 Sparsity and Compressed Sensing in Inverse Problems 365
Evelyn Herrholz, Dirk Lorenz, Gerd Teschke, and Dennis Trede
18.1 Introduction 365
18.2 Exact Recovery for Ill-Posed Problems 367
18.3 Compressive Sensing Principles for Ill-Posed Problems 372
References 377

19 Low-Rank Dynamics 381
Christian Lubich
19.1 Introduction 381
19.2 Projecting onto the Tangent Space: The Dirac–Frenkel Time-Dependent Variational Approximation Principle 382
19.3 Dynamical Low-Rank Approximation of Matrices 384
19.4 A Projector-Splitting Integrator for Dynamical Low-Rank Approximation 387
19.5 Dynamical Low-Rank Approximation of Tensors 389
19.6 The MCTDH Method for Quantum Dynamics 392
19.7 Low-Rank Differential Equations for Structured Matrix Nearness Problems 394
References 395

20 Computation of Expectations by Markov Chain Monte Carlo Methods 397
Erich Novak and Daniel Rudolf
20.1 Introduction 397
20.2 Approximation of Expectations by MCMC 398
20.3 Application of the Error Bound and Limitations of MCMC 404
20.4 Open Problems and Related Comments 410
References 410

21 Regularity, Complexity, and Approximability of Electronic Wavefunctions 413
Harry Yserentant
21.1 Introduction 413
21.2 The Variational Form of the Equation 415
21.3 The Mixed Regularity of the Wavefunctions 416
21.4 The Transcorrelated Formulation and the Regularity Proof 418

21.5 The Radial-Angular Decomposition 420
21.6 Sparse Grids, Hyperbolic Cross Spaces, and Antisymmetry 421
21.7 Eigenfunction and Wavelet Expansions 427
References 428

Index 429

List of Contributors

Martin Altmayer University of Mannheim, Germany

Jonas Ballani RWTH Aachen, Germany

Denis Belomestny University of Duisburg-Essen, Germany

Christian Bender University of Saarland, Saarbrücken, Germany

Nabi Godarzvand Chegini University of Amsterdam, The Netherlands

Petru A. Cioica Philipps-University of Marburg, Germany

Stephan Dahlke Philipps-University of Marburg, Germany

Wolfgang Dahmen RWTH Aachen, Germany

Steffen Dereich WWU Münster, Germany

Fabian Dickmann University of Duisburg-Essen, Germany

Joscha Diehl Technical University of Berlin, Germany

Nicolas Döhring Technical University of Kaiserslautern, Germany

Patrick Dülk University of Bremen, Germany

Oliver G. Ernst Technical University of Chemnitz, Germany

Ulrich Friedrich Philipps-University of Marburg, Germany

Peter K. Friz Weierstrass Institute, Technical University of Berlin, Germany

Jochen Garcke University of Bonn, Germany

Lars Grasedyck RWTH Aachen, Germany

Michael Griebel University of Bonn, Germany

Mijail Guillemard Technical University of Berlin, Germany

Wolfgang Hackbusch MPI Mathematik in den Naturwiss., Leipzig, Germany

Jan Hamaekers University of Bonn, Germany

Frederik Heber University of Bonn, Germany

Matthias Hein University of Saarland, Saarbrücken, Germany

Dennis Heinen University of Göttingen, Germany

Evelyn Herrholz Neubrandenburg University of Applied Sciences, Neubrandenburg, Germany

Chunyan Huang School of Applied Mathematics, Central University of Finance and Economics, Beijing, P.R.China

Armin Iske University of Hamburg, Germany

Tobias Jahnke Karlsruhe Institute of Technology, Karlsruhe, Germany

Leonardo Jost University of Saarland, Saarbrücken, Germany

Lutz Kämmerer Technical University of Chemnitz, Germany

Kamil S. Kazimierski University of Graz, Austria

Rüdiger Kiesel University of Duisburg-Essen, Germany

Stefan Kinzel Philipps-University of Marburg, Germany

Irene Klompmaker Technical University of Berlin, Germany

Melanie Kluge RWTH Aachen, Germany

Sara Krause-Solberg University of Hamburg, Germany

Stefan Kunis University of Osnabrück, Germany

Gitta Kutyniok Technical University of Berlin, Germany

Sangmeng Li WWU Münster, Germany

Wang-Q Lim Technical University of Berlin, Germany

Felix Lindner Technical University of Kaiserslautern, Germany

Dirk Lorenz Technical University of Braunschweig, Germany

Christian Lubich University of Tübingen, Germany

Peter Maass University of Bremen, Germany

Hilmar Mai Weierstrass Institute, Berlin, Germany

Ines Melzer University of Osnabrück, Germany

Thomas Müller-Gronbach University of Passau, Germany

Andreas Neuenkirch University of Mannheim, Germany

Erich Novak Friedrich Schiller University of Jena, Germany

Harald Oberhauser University of Oxford, UK

Gerlind Plonka-Hoch University of Göttingen, Germany

Daniel Potts Technical University of Chemnitz, Germany

Thorsten Raasch University of Mainz, Germany

Rudolf Ressel DLR Oberpfaffenhofen, EOC, Wessling, Germany

Sebastian Riedel Technical University of Berlin, Germany

Klaus Ritter Technical University of Kaiserslautern, Germany

Daniel Rudolf Friedrich Schiller University of Jena, Germany

Andreas Rupp University of Ulm, Germany

René L. Schilling Technical University of Dresden, Germany

Reinhold Schneider Technical University of Berlin, Germany

Christoph Schwab ETH Zürich, Switzerland

Nikolaus Schweizer University of Saarland, Saarbrücken, Germany

Simon Setzer University of Saarland, Saarbrücken, Germany

Björn Sprungk Technical University of Chemnitz, Germany

Wilhelm Stannat Technical University of Berlin, Germany

Hans-Jörg Starkloff University of Applied Sciences Zwickau, Germany

Rob Stevenson University of Amsterdam, The Netherlands

Vikram Sunkara Karlsruhe Institute of Technology, Karlsruhe, Germany

Gerd Teschke Neubrandenburg University of Applied Sciences, Neubrandenburg, Germany

Dennis Trede University of Bremen, Germany

Karsten Urban University of Ulm, Germany

Toni Volkmer Technical University of Chemnitz, Germany

Gerrit Welper Texas A & M University, College Station, TX, USA

Larisa Yaroslavtseva University of Passau, Germany

Harry Yserentant Technical University of Berlin, Germany

Chapter 1
Solving Stochastic Dynamic Programs by Convex Optimization and Simulation

Denis Belomestny, Christian Bender, Fabian Dickmann, and Nikolaus Schweizer

Abstract In this chapter we review some recent progress on Monte Carlo methods for a class of stochastic dynamic programming equations, which accommodates optimal stopping problems and time discretization schemes for backward stochastic differential equations with convex generators. We first provide a primal maximization problem and a dual minimization problem, based on which confidence intervals for the value of the dynamic program can be constructed by Monte Carlo simulation. For the computation of the lower confidence bounds we apply martingale basis functions within a least-squares Monte Carlo implementation. For the upper confidence bounds we suggest a multilevel simulation within a nested Monte Carlo approach and, alternatively, a generic sieve optimization approach with a variance penalty term.

1.1 Introduction

In this chapter we review some recent progress on Monte Carlo methods for dynamic programming equations of the form

$$Y_j^* = F_j(E_j[\beta_{j+1} Y_{j+1}^*]),\ j = 0, \dots, J-1, \quad Y_J^* = F_J(0) \tag{1.1}$$

on a complete filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_j)_{j=0,\dots,J}, P)$ in discrete time. In this equation an adapted $\mathbb{R}^{D+1}$-valued process β and the adapted random field $F : \{0, \dots, J\} \times \Omega \times \mathbb{R}^{D+1} \to \mathbb{R}$ are given. Moreover, $E_j[\cdot]$ denotes the conditional expectation given $\mathscr{F}_j$. Assumptions on β and F will be specified later on.

D. Belomestny • F. Dickmann
Universität Duisburg-Essen, Thea-Leymann-Str. 9, 45127 Essen, Germany
e-mail: denis.belomestny@uni-due.de; fabian.dickmann@uni-due.de

C. Bender (✉) • N. Schweizer
Universität des Saarlandes, Postfach 151150, 66041 Saarbrücken, Germany
e-mail: bender@math.uni-sb.de; schweizer@math.uni-sb.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_1

Several time discretization schemes for backward stochastic differential equations (BSDEs) with or without reflection and for fully nonlinear second order parabolic PDEs lead to dynamic programs of the form (1.1), see [10,11,15,20,28]. In financial engineering, equations of the form (1.1) appear (after a time discretization is performed) in many nonlinear option pricing problems. These include:

- *Bermudan option pricing*: Here $\beta \equiv 1$ and $F_j(y) = \max\{S_j, y\}$, where the adapted process S_j denotes the discounted payoff of the Bermudan option, when called at the jth exercise time. Then, Y_0^* is the price of the Bermudan option (in discounted units), see e.g. [25].
- *Credit value adjustment*: Here $\beta \equiv 1$, $F_j(y) = (1 - r\Delta)y - (1 - R)\lambda\Delta(y)_+$ for $j < J$, where $r \geq 0$ is the risk-free interest rate, $\lambda > 0$ is the default intensity of the counterparty, $R \in [0, 1)$ is the recovery rate in case of default, and $(\cdot)_+$ denotes the positive part. The random variable $F_J(0)$ represents the payoff of the option at maturity T, if there is no default prior to maturity, and the interval $[0, T]$ is divided into J equidistant subintervals of length Δ. Then, Y_j^* is the price of the option at time $j\Delta$ including credit value adjustment (in a reduced form approach), provided that default did not occur prior to $j\Delta$. See e.g. [13,14] for BSDE approaches to pricing under credit risk.
- *Funding costs:* We now assume that funding costs are incorporated in the valuation mechanism, when at time $j\Delta$ the hedging costs for the delta hedge in the risky stocks $X_j^1, \ldots, X_j^D$ exceeds the price of the option with payoff $F_J(0)$ (at maturity T). In this case $F_j(y_0, \ldots, y_D) = (1 - r\Delta)y_0 - R(\sum_{d=1}^D y_d - y_0)_+\Delta$ for $j < J$, where r is the interest rate, at which money can be lent, and $(R + r)$ is the rate, at which money can be borrowed. This is a classical example of nonlinear option pricing by BSDEs, for which we refer to the survey paper [19]. The variable y_0 represents the price of the option and the variables y_d, $d = 1, \ldots, D$, describe the amount of money required for the delta hedge in the dth stock. Correspondingly one chooses $\beta_j^0 = 1$ and β_j^d as (a suitable approximation of) $X_{j-1}^d(X_j^d - E_{j-1}[X_j^d])/E_{j-1}[(X_j^d - E_{j-1}[X_j^d])^2]$.

The main difficulty when solving equations of the form (1.1) numerically is that, going backwards in time, in each time step a conditional expectation must be approximated which depends on the numerical approximation of Y^* one time step ahead. Therefore one needs to apply an approximate operator for the conditional expectation which can be nested without exploding costs. In particular, when the generator F depends on ω through a high-dimensional Markovian process, Monte Carlo methods are usually applied to estimate the conditional expectations. In this respect, the least-squares Monte Carlo method, which was suggested for Bermudan option pricing by [21,26] and for BSDEs by [20], is certainly among the most popular choices. For the Bermudan option pricing problem this approximate dynamic programming approach (i.e. solving the dynamic program with the conditional expectation replaced by an approximate operator) is often complemented with the primal-dual methodology of [1]. In a nutshell, the solution of the approximate dynamic program is taken as an input in order to construct confidence intervals for

the price Y_0^* of the option. This approach crucially relies on the dual representation of [18, 23] for Bermudan option pricing.

In Sect. 1.2 we first provide a review of this primal-dual approach for Bermudan option pricing. Following the lines of [7] we then generalize the theory behind this approach to dynamic programs of the form (1.1) under the assumptions that the driver F is convex and that a discrete comparison principle holds. The remaining sections are devoted to making this general primal-dual approach practical by designing and analyzing algorithms, which improve on the existing literature in various aspects. In Sect. 1.3 we suggest to run the least-squares Monte Carlo method for the approximate dynamic program with a set of basis functions which satisfy a martingale property. While this corresponds to the 'regression later' approach of [17] for the Bermudan option problem, it was recently observed by [8] that the use of martingale basis functions can significantly reduce the propagation of the projection error over time and the variance in the context of time discretization schemes for BSDEs.

Given the corresponding approximate solution of the dynamic program (1.1), the construction of a lower confidence bound for Y_0^* is usually a straightforward application of the primal-dual methodology. Contrarily, the construction of the upper bound requires a martingale as input, which should be close to the Doob martingale of βY^*. In the context of Bermudan option pricing, Andersen and Broadie [1] suggested a method to approximate this martingale starting from the solution of the approximate dynamic program and applying one layer of nested simulation in order to compute the Doob decomposition numerically. Based on [5] we present in Sect. 1.4.1 a multilevel variant of this algorithm, where varying numbers of paths are applied for the two layers of simulations at different levels. This multilevel variant can be shown to reduce the complexity of the Andersen-Broadie algorithm from ε^{-4} (generic nested Monte Carlo) to $\varepsilon^{-2}\log^2(\varepsilon)$, which up to the logarithmic factor is the same complexity as a plain non-nested Monte Carlo implementation. As an alternative to the Andersen-Broadie type algorithms we also present a completely generic approach to the approximation of the Doob martingale of βY^* via sieve optimization combined with a variance penalty term in Sect. 1.4.2. Convergence of this algorithm was analyzed in [3] for the Bermudan option pricing problem as the number of martingales in the sieve and the number of simulated samples converges to infinity. Finally, we illustrate the proposed algorithms by numerical experiments in the context of nonlinear expectations under model uncertainty and of option pricing under credit value adjustment.

1.2 The Primal-Dual Approach to Convex Dynamic Programs

In this section we first recall how the primal-dual approach works for the Bermudan option pricing problem. Then we present a generalization to dynamic programs of the form (1.1) with convex generator.

As stated in the introduction, the Bermudan option pricing problem leads to a dynamic program of the form

$$Y_j^* = \max\{S_j, E_j[Y_{j+1}^*]\}, \quad Y_J^* = S_J \tag{1.2}$$

for some adapted and integrable process S with $S_J \geq 0$. The starting point of the primal-dual approach is the well-known observation that this dynamic program is the one associated to the *optimal stopping problem* (primal problem), i.e.

$$Y_0^* = \sup_{\tau \in \mathscr{S}} E[S_\tau], \tag{1.3}$$

where $\mathscr{S}$ is the set of stopping times with values greater than or equal to j, and the (smallest) optimal stopping time τ^* can be expressed as

$$\tau^* = \inf\{i \geq 0;\ S_i \geq E_i[Y_{i+1}^*]\}.$$

Hence, for any stopping time τ, $Y_0^{low} := E[S_\tau]$ yields a lower bound for the Bermudan option price Y_0^*. In practice, a 'close-to-optimal' stopping time τ is often constructed as follows: One first rephrases the dynamic program in terms of the continuation value $Z_j^* := E_j[Y_{j+1}^*]$ as

$$Z_j^* = E_j[\max\{S_{j+1}, Z_{j+1}^*\}], \quad Z_J^* = 0.$$

Then, one solves this dynamic program numerically, replacing the conditional expectation by some approximate operator, which leads to an approximation Z of Z^*. Finally, based on Z one constructs the lower bound Y_0^{low} via the stopping time $\tau = \inf\{i \geq 0;\ S_i \geq Z_i\}$. The primal lower bound is then complemented by a dual upper bound. Indeed, Rogers [23] and Haugh and Kogan [18] showed independently that Y_0^* can be expressed via the dual minimization problem

$$Y_0^* = \inf_{M \in \mathscr{M}_1} E[\max_{j=0,\ldots,J}(S_j - M_j)], \tag{1.4}$$

where $\mathscr{M}_{D+1}$ denotes the set of $\mathbb{R}^{D+1}$-valued martingales with $M_0 = 0$, and that the Doob martingale of Y^* is optimal. Hence, the construction of a tight upper bound requires the numerical approximation of the Doob decomposition of Y^*. The nested Monte Carlo algorithm by [1] is popular to perform such numerical Doob decompositions, but in Sect. 1.4 we present algorithms that can produce tight upper bounds at the cost of a non-nested Monte Carlo implementation.

Following the approach of [7], which is detailed there for the case of discrete time reflected BSDEs, we now generalize the construction of a primal maximization problem and a dual minimization problem to dynamic programs of the form (1.1). The following assumptions are in force:

(R) $(\beta_j)_j = (\beta_{0,j}, \dots, \beta_{D,j})_j$ is a bounded, adapted $D+1$-dimensional process with $\beta_{0,j} \equiv 1$ for all j. The adapted random field $F : \{0, \dots, J\} \times \Omega \times \mathbb{R}^{D+1} \to \mathbb{R}$ is Lipschitz continuous in $z \in \mathbb{R}^{D+1}$ uniformly in (j, ω) and satisfies $E[|F_j(0)|^2] < \infty$ for every $j = 0, \dots, J$.

(Comp) For every j and any two $\mathscr{F}_{j+1}$-measurable, integrable real-valued random variables $Y, \tilde{Y}$ such that $Y \geq \tilde{Y}$ a.s., it holds that

$$F_j(E_j[\beta_{j+1} Y]) \geq F_j(E_j[\beta_{j+1} \tilde{Y}]).$$

(Conv) The map $z \mapsto F_j(\omega, z)$ is convex for every j and almost every ω.

We briefly comment on the first two assumptions. The regularity condition (R) makes sure that the dynamic program (1.1) recursively defines square-integrable random variables Y_j^*, $j = J, \dots, 0$. Condition (Comp) entails a *comparison principle* for the dynamic program (1.1). Indeed, if Y is a *subsolution* of (1.1), i.e.

$$Y_j \leq F_j(E_j[\beta_{j+1} Y_{j+1}]), \ j = 0, \dots, J-1, \quad Y_J \leq F_J(0),$$

then one can easily show by backward induction that, thanks to (Comp),

$$Y_j \leq Y_j^*, \quad j = 0, \dots, J. \tag{1.5}$$

Of course, the analogous statement holds for supersolutions.

Primal lower bounds. The construction of the primal maximization problem relies on a linearization of F in terms of its convex conjugate and is analogous to Proposition 3.4 in [19] for BSDEs in continuous time. Recall that the convex conjugate $F_j^\#$ of F_j is defined by

$$F_j^\#(\rho) = \sup_{z \in \mathbb{R}^{D+1}} \rho^\top z - F_j(z) \tag{1.6}$$

and lives on the (bounded uniformly in ω) domain $D_{F^\#}^{j,\omega} \subseteq \mathbb{R}^{D+1}$ where the supremum in (1.6) is finite. We denote by $\mathscr{A}$ the set of adapted, $\mathbb{R}^{D+1}$-valued processes ρ such that ρ_j takes values in $D_{F^\#}^{j,\omega}$ and satisfies $E[F_j^\#(\rho_j)] < \infty$ for $j = 0, \dots, J-1$. For a fixed $\rho \in \mathscr{A}$, we define recursively the typically non-adapted process $\theta^{low} := \theta^{low}(\rho)$ via $\theta_J^{low} := F_J(0)$ and

$$\theta_j^{low} := \rho_j^\top \beta_{j+1} \theta_{j+1}^{low} - F_j^\#(\rho_j) = F_J(0) \prod_{k=j}^{J-1} \rho_k^\top \beta_{k+1} - \sum_{i=j}^{J-1} F_i^\#(\rho_i) \prod_{k=j}^{i-1} \rho_k^\top \beta_{k+1}. \tag{1.7}$$

Then, the adapted process defined by $Y_j^{low} := Y_j^{low}(\rho) := E_j[\theta_j^{low}]$ satisfies

$$Y_j^{low} = \rho_j^\top E_j[\beta_{j+1} Y_{j+1}^{low}] - F_j^\#(\rho_j) \le \sup_{\rho \in D_{F^\#}^{j,\omega}} (\rho^\top E_j[\beta_{j+1} Y_{j+1}^{low}] - F_j^\#(\rho))$$

$$= F_j(E_j[\beta_{j+1} Y_{j+1}^{low}]), \quad j = 0, \dots, J-1. \tag{1.8}$$

where the final step uses that $F_j = F_j^{\#\#}$ by convexity. As $Y_J^{low} = F_J(0) = Y_J^*$, we observe that Y^{low} is a subsolution, and, hence, (1.5) yields $Y_j^{low}(\rho) \le Y_j^*$ for every $j = 0, \dots, J$. Finally, by the Lipschitz assumption there exists an adapted process ρ^* such that

$$\rho_j^{*\top} E_j[\beta_{j+1} Y_{j+1}^*] - F_j^\#(\rho_j^*) = F_j(E_j[\beta_{j+1} Y_{j+1}^*]). \tag{1.9}$$

One can now show by induction that $Y_j^* = Y_j^{low}(\rho^*)$ for every $j = 0, \dots, J$.

We can summarize these considerations in the following theorem.

Theorem 1.1 (Primal problem). *Under assumptions (R), (Comp), and (Conv), Y_0^* can be represented as value of the maximization problem*

$$Y_0^* = \sup_{\rho \in \mathscr{A}} E[\theta_0^{low}(\rho)] = \sup_{\rho \in \mathscr{A}} E\left[F_J(0) \prod_{k=0}^{J-1} \rho_k^\top \beta_{k+1} - \sum_{i=0}^{J-1} F_i^\#(\rho_i) \prod_{k=0}^{i-1} \rho_k^\top \beta_{k+1} \right].$$

Moreover, any process $\rho^* \in \mathscr{A}$, which satisfies (1.9), is optimal.

Dual upper bounds. For the construction of the dual minimization problem we apply a pathwise dynamic programming approach, i.e. the conditional expectations are dropped in (1.1), but some martingale increments are added to the equation instead. To this end we first fix an $\mathbb{R}^{D+1}$-valued martingale, i.e. an integrable and adapted process M with $E_j[M_{j+1} - M_j] = 0$ and $M_0 = 0$. Define recursively the typically non-adapted process $\theta^{up} := \theta^{up}(M)$ via $\theta_J^{up} := F_J(0)$ and

$$\theta_j^{up} = F_j(\beta_{j+1} \theta_{j+1}^{up} - (M_{j+1} - M_j)).$$

Taking conditional expectations and applying Jensen's inequality shows that the adapted process $Y_j^{up} = E_j[\theta_j^{up}]$ satisfies

$$Y_j^{up} \ge F_j(E_j[\beta_{j+1} \theta_{j+1}^{up}]) = F_j(E_j[\beta_{j+1} Y_{j+1}^{up}]), \quad j = 0, \dots, J-1. \tag{1.10}$$

As $Y_J^* = F_J(0) = Y_J^{up}$, Y^{up} is a supersolution of (1.1), and hence the comparison principle implies that $Y_j^{up} \ge Y_j^*$ for all j. Finally, choosing M^* as the the Doob martingale of βY^*, i.e., $M_j^* - M_{j-1}^* = \beta_j Y_j^* - E_{j-1}[\beta_j Y_j^*]$ for all j, one can check inductively that $\theta^{up}(M^*)$ is adapted and that $\theta^{up}(M^*) = Y^*$. We, thus, arrive at the following result.

Theorem 1.2 (Dual problem). *Under assumptions (R), (Comp), and (Conv),* Y_0^* *can be represented as value of the minimization problem*

$$Y_0^* = \inf_{M \in \mathscr{M}_{D+1}} E[\theta_0^{up}(M)].$$

Moreover, the Doob martingale of βY^* *is optimal even in the sense of pathwise control, i.e.* $\theta_0^{up}(M^*) = Y_0^*$

Remark 1.1. (i) As explained in Remark 3.5 of [7], the above minimization problem can be re-interpreted as the dual problem to the maximization problem in Theorem 1.1 in the sense of information relaxation. For the general theory of information relaxation duals for discrete time stochastic control problems we refer to [12].

(ii) The results in [7] also cover constructions of minimization and maximization problems with value given by Y_0^* for implicit dynamic programs of the form

$$Y_j^* = F_j(Y_j,\ E_j[\beta_{j+1} Y_{j+1}^*]),\ j = 0, \ldots, J-1, \quad Y_J^* = F_J(0),$$

even without imposing the convexity assumption on F.

(iii) The primal-dual methodology can also be applied for problems with a multi-dimensional value process Y^* such as multiple stopping problems, see [6, 24].

Examples. (i) We first revisit the Bermudan option problem, which is governed by the dynamic programming equation (1.2). As $D = 0$, $\beta \equiv 1$ and $F_j(z) = \max\{S_j, z\}$, the standing assumptions are satisfied. One easily computes $F_j^{\#}(\rho) = (\rho - 1)S_j$ with domain $D_{F^{\#}}^{j,\omega} = [0, 1]$. The primal problem of Theorem 1.1 then reads

$$Y_0^* = \sup_{\rho} E\left[S_J \prod_{k=0}^{J-1} \rho_k + \sum_{i=0}^{J-1} S_i (1 - \rho_i) \prod_{k=0}^{i-1} \rho_k, \right]$$

where ρ runs over the set of adapted process with values in $[0, 1]$. By the optimality condition (1.9), it obviously suffices to take the supremum over the set of adapted processes ρ with values in $\{0, 1\}$. The primal problem is then seen to be a reformulation of the optimal stopping problem (1.3), if one maps ρ on the stopping time $\inf\{i \geq 0;\ \rho_i = 0\}$. Concerning the dual minimization problem, one can check inductively that in this case $\theta_j^{up}(M) = \max_{i \in \{j,\ldots,J\}}(S_i - (M_i - M_j))$. Hence, the dual minimization problem in Theorem 1.2 collapses to the dual formulation in (1.4) due to [18, 23].

(ii) The second example is concerned with an Euler type time discretization scheme for *backward stochastic differential equations* (BSDEs) driven by a D-dimensional Brownian motion W. For a BSDE of the form

$$d\mathscr{Y}_t = -f(t, \mathscr{Y}_t, \mathscr{Z}_t)dt + \mathscr{Z}_t^\top dW_t, \quad \mathscr{Y}_T = h$$

we consider Y^* as discretization over the time grid $\{t_0, \ldots, t_J\}$, where:

$$Y_j^* = E_j[Y_{j+1}^*] + (t_{j+1} - t_j) f\left(t_j, E_j[Y_{j+1}], E_j\left[Y_{j+1}^* \frac{W_{t_{j+1}} - W_{t_j}}{t_{j+1} - t_j}\right]\right) \tag{1.11}$$

with terminal condition $Y_J^* = h$. The generator f is an adapted, square-integrable, convex and (uniformly in (t, ω)) Lipschitz continuous random field and h is a square-integrable $\mathscr{F}_J$-measurable random variable. This is a slight variant of the schemes studied by [11, 28] and coincides with the one suggested by [15] in the more general context of second order BSDEs. As filtration in discrete time we can choose the one generated by the Brownian motion up to the jth point in the time grid. By defining $\beta_1, \ldots, \beta_D$ as suitably normalized and truncated increments of the Brownian motion, this recursion is of the form $F_j(z) = z_0 + (t_{j+1} - t_j) f(t_j, z)$. Assumptions (R) and (Conv) are then certainly fulfilled. The truncation of β depends on the time grid and the Lipschitz constants of f in an appropriate way and is necessary to ensure that (Comp) is satisfied, see [7] for details.

1.3 Construction of Lower Bounds via Martingale Basis Functions

This section reviews the popular least-squares Monte Carlo approach for the approximate solution of a dynamic program of the form (1.1) via empirical regression on a set of basis functions, see e.g. [20,21,26]. A special emphasis will be on the particular situation where the basis functions form a set of martingales. This case was studied by [17] for optimal stopping problems and by [8] for the BSDE case.

In view of the optimality condition (1.9) for the primal maximization problem we first rewrite the dynamic program in terms of $Z_j^* := E_j[\beta_{j+1} Y_{j+1}^*]$ as

$$Z_j^* = E_j[\beta_{j+1} F_{j+1}(Z_{j+1}^*)], \quad Z_J^* = 0, \tag{1.12}$$

and note that the solution of the dynamic program (1.1) can be recovered from Z^* as $Y_j^* = F_j(Z_j^*)$. The basic idea of the least-squares Monte Carlo approach is to replace the conditional expectations in (1.12) by an orthogonal projection on a linear subspace of $L^2(\mathscr{F}_j)$, which is spanned by a set of basis functions. The orthogonal projection is then calculated numerically via Monte Carlo simulation by replacing the expectations in the definition of the orthogonal projection by empirical means. More precisely, denote by $\eta_{d,j}$ a row vector of Λ $\mathscr{F}_j$-measurable random

variables for every time index j and every $d = 0, \ldots, D+1$. We then define an approximation Z_j of Z_j^* by

$$Z_{d,j} = \eta_{d,j}\alpha_{d,j}, \quad d = 0, \ldots, D,$$

where the coefficients $\alpha_{d,j}$ are computed as follows: Assume we have N independent copies ('regression paths') of

$$\left\{(F_j^{(n)}, \beta_j^{(n)}, \eta_j^{(n)}),\ j = 0, \ldots, J,\ n = 1, \ldots, N\right\}$$

at hand. We now define $\alpha_{d,J} = 0$ for every $d = 0, \ldots, D$ and

$$\alpha_{d,j} = \arg\min_{\alpha \in \mathbb{R}^\Lambda} \frac{1}{N} \sum_{n=1}^{N} \left| \beta_{d,j+1}^{(n)} F_{j+1}^{(n)}(\eta_{0,j+1}^{(n)}\alpha_{0,j+1}, \ldots, \eta_{D,j+1}^{(n)}\alpha_{D,j+1}) - \eta_{d,j}^{(n)}\alpha \right|^2. \tag{1.13}$$

Given these coefficients we can compute on the one hand an approximation of Y^* by $Y_j = F_j(\eta_{0,j}\alpha_{0,j}, \ldots, \eta_{D,j}\alpha_{D,j})$, and on the other hand we can (approximatively) solve for the optimality criterion (1.9) with Z^* replaced by Z in order to get an approximation ρ of the optimizer ρ^* of the primal problem, i.e. ρ_j satisfies

$$(\eta_{0,j}\alpha_{0,j}, \ldots, \eta_{D,j}\alpha_{D,j})\rho_j - F_j^\#(\rho_j) \approx F_j((\eta_{0,j}\alpha_{0,j}, \ldots, \eta_{D,j}\alpha_{D,j})).$$

Then,

$$E\left[F_J(0) \prod_{k=0}^{J-1} \rho_k^\top \beta_{k+1} - \sum_{i=0}^{J-1} F_i^\#(\rho_i) \prod_{k=0}^{i-1} \rho_k^\top \beta_{k+1} \right]$$

is a lower bound for Y_0^* which is expected to be good, if the basis functions are well-chosen and the number of simulated sample paths is sufficiently large. For a detailed analysis of the projection error due to the choice of the basis and of the simulation error for least-squares Monte Carlo algorithms we refer to [27] and [2] for the Bermudan option pricing problems and to [20] for the BSDE case. Lower confidence bounds for Y_0^* can finally be calculated by replacing the expectation by a sample mean over a new set of independent samples ('outer paths') of $\{F, \beta, \eta\}$ (which are independent of the regression paths). We note that the complexity of this type of algorithm can be reduced by a multilevel approach, which balances the cost between the effort for approximating the conditional expectations and the number of outer paths at different levels, see [4] for the Bermudan option problem. We do not dwell on the details here, but present a similar idea for the computation of upper bounds in Sect. 1.4.1.

In order to illustrate the above least-squares Monte Carlo scheme, let us denote the simulation based projection on the dth set of basis functions at time j by $\mathscr{P}_{d,j}$.

Then the algorithm can be written (informally) as

$$Z_j = \mathscr{P}_{d,j}\left(\beta_{j+1}F_{j+1}(Z_{j+1})\right),$$

i.e. the conditional expectations of the dynamic program are replaced by the empirical projections. We now modify this algorithm by adding an additional projection. Precisely we replace the above Z_j by

$$\tilde{Z}_j = \mathscr{P}_{d,j}\left(\beta_{j+1}\mathscr{P}_{0,j+1}(F_{j+1}(\tilde{Z}_{j+1}))\right).$$

A-priori this does not look like a good idea, because each additional empirical projection is expected to increase the numerical error. However, this scheme can be simplified, if the basis satisfies the following additional martingale property:

(MB) The basis functions $\eta_{0,j}$ form a system of martingales, i.e. $E_j[\eta_{0,j+1}] = \eta_{0,j}$ for $j = 0,\dots,J-1$ and, for $d = 1,\dots,D$, the basis functions are defined via $\eta_{d,j} := E_i[\beta_{d,j+1}\eta_{0,j+1}]$ (which entails that these conditional expectations are available in closed form).

Under this martingale basis assumption one chooses one set of basis functions $\eta_{0,J}$ at terminal time only, and all the other basis functions are computed from this set. The main advantage of assumption (MB) is that conditional expectations of linear combinations of the basis functions (even if multiplied by the β-weights) are at hand in closed form. Hence, the outer empirical projections in the definition of $\tilde{Z}$ need not be performed, but should rather be replaced by the true conditional expectations. These considerations lead to the the *martingale basis algorithm*

$$\tilde{Z}_j^{(MB)} = E_j\left[\beta_{j+1}\mathscr{P}_{0,j+1}(F_{j+1}(\tilde{Z}_{j+1}^{(MB)}))\right].$$

More precisely, one modifies the construction of the coefficients $\alpha_{d,i}$ compared to the standard least-square Monte Carlo scheme as follows. Define $\alpha_{d,i} = \alpha_i$ for all $d = 0,\dots,D$, where $\alpha_J = 0$ and

$$\alpha_j = \arg\min_{\alpha\in\mathbb{R}^A} \frac{1}{N}\sum_{n=1}^{N}\left|F_{j+1}^{(n)}(\eta_{0,j+1}^{(n)}\alpha_{j+1},\dots,\eta_{D,j+1}^{(n)}\alpha_{j+1}) - \eta_{0,j+1}^{(n)}\alpha\right|^2 \tag{1.14}$$

Once the coefficients are computed, one constructs the approximations of Y^*, Z^*, ρ^* and the lower bound for Y_0^* in exactly the same way as described above. An obvious advantage of this martingale basis algorithm for $D \geq 1$ is, that only one empirical regression is performed at each time step, while the original least-squares Monte Carlo algorithm requires $(D+1)$ empirical regressions per time step.

In the setting of discrete time approximations of BSDEs one has $F_j(z) = z_0 + (t_{j+1} - t_j)f(t_j, z)$. Hence, (with a slight abuse of notation), the martingale basis

algorithm can be further simplified to

$$Z_j^{(MB)} = E_j\left[\beta_{j+1}\left(Z_{0,j+1}^{(MB)} + (t_{j+2} - t_{j+1})\mathscr{P}_{0,j+1}(f(t_{j+1}, Z_{j+1}^{(MB)}))\right)\right], j \leq J-2,$$

because $Z_{0,j+1}^{(MB)}$ is already a linear combination of the basis functions and, thus, need not be projected on the basis. In this BSDE setting, the projection error, i.e. the error which stems from the basis choice, was analyzed in [8]. Very roughly speaking, they show that with martingale basis functions the total projection error is an average of the projection error at the different time indices j, while it is known that, for general function bases, the projection errors sum up over time, see e.g. [20]. Concerning the simulation error we will demonstrate the strong variance reduction effect of the martingale basis algorithm in the numerical examples.

1.4 Construction of Upper Bounds: Multilevel Monte Carlo and Sieve Optimization

In this section we present two algorithms for approximating the Doob martingale of βY^*, which in view of the dual representation in Theorem 1.2 give rise to the computation of tight upper bounds on Y_0^*. The first one is a generalization of the multilevel approach of [5] to the setting of Sect. 1.2, while the second one relies on a generic sieve optimization approach [3].

1.4.1 Multilevel Dual Approach

In Sect. 1.2 we observed that the Doob martingale M^* of βY^* solves the dual minimization problem. The most common approach for deriving tight upper bounds on Y_0^* from this observation is to approximate M^* by the Doob martingale M associated to some approximation Y of Y^*. M is given by

$$M_j = \sum_{i=1}^{j} \left(\beta_i Y_i - E_{i-1}\left[\beta_i Y_i\right]\right), \quad j = 0, \ldots, J. \tag{1.15}$$

Unless the conditional expectations on the right hand side of (1.15) can be calculated explicitly, further approximations are necessary to obtain numerically tractable upper bounds on Y_0^* however. Andersen and Broadie [1] suggested to estimate these by one layer of nested Monte Carlo. This leads to an estimate

$$M_j^K = \sum_{i=1}^{j} \left(\beta_i Y_i - \frac{1}{K} \sum_{\nu=1}^{K} \xi_i^{(\nu)} \right), \quad K \in \mathbb{N},$$

where, conditionally on $\mathscr{F}_J$, all $\mathbb{R}^{D+1}$-valued random variables $\xi_j^{(\nu)}$, $\nu = 1, \ldots, K$, $j = 1, \ldots, J$, are independent and fulfill $Law(\xi_j^{(\nu)} | \mathscr{F}_J) = Law(\beta_j Y_j | \mathscr{F}_{j-1})$ and thus

$$E\left[\xi_j^{(\nu)} | \mathscr{F}_J \right] = E_{j-1}\left[\xi_j^{(\nu)} \right] = E_{j-1}\left[\beta_j Y_j \right].$$

The martingale M^K is explicit enough to allow for a straightforward Monte Carlo estimator relying on N so-called outer paths and K inner samples at each outer path and time point: Fix natural numbers N and K, and consider N independent copies

$$\left\{ (F_j^{(n)}, \beta_j^{(n)}, M^{K,(n)}),\ j = 0, \ldots, J,\ n = 1, \ldots, N \right\}$$

of the process (F, β, M^K). Denote by $\theta_0^{up,(n)}(M^{K,(n)})$ the n^{th} copy of $\theta_0^{up}(M^K)$, i.e.

$$\theta_j^{up,(n)}(M^{K,(n)}) = F_j^{(n)} \left(\beta_{j+1}^{(n)} \theta_{j+1}^{up,(n)}(M^{K,(n)}) - (M_{j+1}^{K,(n)} - M_j^{K,(n)}) \right),\ j = 0, \ldots, J-1,$$

with $\theta_J^{up,(n)}(M^{K,(n)}) = F_J^{(n)}(0)$, and consider the estimator

$$Y_0^{N,K} := \frac{1}{N} \sum_{n=1}^{N} \theta_0^{up,(n)}(M^{K,(n)}). \tag{1.16}$$

In the following we study the simulation error, i.e., the difference between $Y_0^{N,K}$ and $Y_0^{up} = E[\theta_0^{up}(M)]$. Denote by $|\cdot|$ the Euclidean norm in $\mathbb{R}^{D+1}$. We have

$$\begin{aligned} E\left[|M_j^K - M_j|^2 \right] &= E\left[\left| \sum_{i=1}^{j} \left(E_{i-1}\left[\beta_i Y_i \right] - \frac{1}{K} \sum_{l=1}^{K} \xi_i^{(l)} \right) \right|^2 \right] \\ &= \frac{1}{K} \sum_{i=1}^{j} E\left[\left| E_{i-1}\left[\beta_i Y_i \right] - \xi_i^{(1)} \right|^2 \right] \leq \frac{1}{K} \sum_{i=1}^{J} E[|\beta_i Y_i|^2] = O(1/K), \end{aligned}$$

provided that $E[|\beta_i Y_i|^2] < \infty$ for $i = 0, \ldots, J$. By the Lipschitz continuity of the random field F and the boundedness of β under (R), this implies

$$E\left[|\theta_0^{up}(M^K) - \theta_0^{up}(M)|^2 \right] \leq CK^{-1}, \tag{1.17}$$

for some constant C, and hence

$$E\left[(Y_0^{N,K} - Y_0^{up})^2\right] \leq N^{-1}\mathrm{Var}(\theta_0^{up}(M^K)) + CK^{-1} =: N^{-1}v_K + CK^{-1}.$$

Therefore, in order to ensure $\sqrt{E\left[(Y_0^{N,K} - Y_0^{up})^2\right]} \leq \varepsilon$, we may take

$$K = \left\lceil \frac{2C}{\varepsilon^2} \right\rceil, \quad \text{and} \quad N = \left\lceil \frac{2v_K}{\varepsilon^2} \right\rceil$$

with $\lceil x \rceil$ denoting the first integer which is larger than or equal to x. If v_K is non-increasing, then, given an accuracy ε, the complexity of the MC estimate $Y_0^{N,K}$ is, up to a constant,

$$\mathscr{C}^{N,K}(\varepsilon) := NK = \frac{2v_K}{\varepsilon^2}\frac{2C}{\varepsilon^2} \lesssim \frac{v_{\left\lceil \frac{2C}{\varepsilon^2} \right\rceil}}{\varepsilon^4}.$$

The question is whether one can reduce the complexity of the Andersen-Broadie approach. For the optimal stopping problem (1.2), [5] introduced and studied a new multilevel approach, which makes substantial reductions possible. The idea of this multilevel dual approach is inspired by the pathbreaking work of [16] on the multilevel approach to discretization of SDEs.

Now we describe the main idea of the approach of [5] in the more general setting of the convex dynamic program (1.1). Let $L \in \mathbb{N}$ and $\mathbf{K} = (K_0, \ldots, K_L)$ with $1 \leq K_0 < K_1 < \ldots < K_L$. Recall that $Y_0^{up}(M^{K_l}) = E[\theta_0^{up}(M^{K_l})]$ and observe that

$$\begin{aligned} Y_0^{up}(M^{K_L}) &= Y_0^{up}(M^{K_0}) + \sum_{l=1}^{L}[Y_0^{up}(M^{K_l}) - Y_0^{up}(M^{K_{l-1}})] \\ &= E[\theta_0^{up}(M^{K_0})] + \sum_{l=1}^{L} E[\theta_0^{up}(M^{K_l}) - \theta_0^{up}(M^{K_{l-1}})]. \end{aligned}$$

The multilevel algorithm estimates each expectation in the above sum by Monte Carlo in such a way that the variance of the resulting estimate is small. This is achieved by using the same trajectories within one level to compute martingale estimates M^{K_l} and $M^{K_{l-1}}$. Fix a sequence $\mathbf{N} = (N_0, \ldots, N_L) \in \mathbb{N}^L$ with $1 \leq N_L < \ldots < N_0$ and simulate the initial set of N_0 trajectories

$$\left\{\left(F_j^{(n)}, \beta_j^{(n)}, M_j^{K_0,(n)}\right),\ n = 1, \ldots, N_0,\ j = 0, \ldots, J\right\}$$

of (F, β, M^{K_0}). Further for each level $l = 1, \ldots, L$, we generate independently a set of N_l trajectories

$$\left\{\left(F_j^{(n)}, \beta_j^{(n)}, M_j^{K_{l-1},(n)}, M_j^{K_l,(n)}\right),\ n = 1, \ldots, N_l,\ j = 0, \ldots, J\right\}$$

of $(F, \beta, M^{K_{l-1}}, M^{K_l})$. We suppress the dependence on l of $F_j^{(n)}$, $\beta_j^{(n)}$ etc. in the notation. Finally, we construct a multilevel estimate for $Y_0^{up}(M)$:

$$Y_0^{\mathbf{N},\mathbf{K}} := \frac{1}{N_0}\sum_{n=1}^{N_0} \theta^{up,(n)}(M^{K_0,(n)}) + \sum_{l=1}^{L}\frac{1}{N_l}\sum_{n=1}^{N_l}\left[\theta_0^{up,(n)}(M^{K_l,(n)}) - \theta_0^{up,(n)}(M^{K_{l-1},(n)})\right]$$

where $\theta_0^{up,(n)}(M^{K_{l-1},(n)})$ and $\theta_0^{up,(n)}(M^{K_l,(n)})$ denote the nth copies of $\theta_0^{up}(M^{K_{l-1}})$ and $\theta_0^{up}(M^{K_l})$ at the lth level.

The next theorem shows that the complexity of $Y_0^{\mathbf{N},\mathbf{K}}$ is reduced almost to the one of a non-nested Monte Carlo estimation.

Theorem 1.3. *Suppose (R), (Comp), and (Conv). Fix $\varepsilon > 0$ and let $K_l = K_0\kappa^l$, $l = 0, \ldots, L$, for some $K_0 \in \mathbb{N}$ and $\kappa > 1$. Then there exist constants C_1 and C_2 (independent of ε) such that the choice*

$$L = \left\lceil C_1 \log(\kappa)^{-1} \log\left((\sqrt{K_0}\varepsilon)^{-1}\right)\right\rceil, \quad N_l = \left\lceil C_2\varepsilon^{-2}(L+1)K_0^{-1}\kappa^{-l}\right\rceil$$

yields

$$\sqrt{E[(Y_0^{\mathbf{N},\mathbf{K}} - Y_0^{up}(M))^2]} \le \varepsilon,$$

while the computational complexity of the estimator $Y^{\mathbf{N},\mathbf{K}}$ is of order

$$\mathscr{C}^{\mathbf{N},\mathbf{K}}(\varepsilon) = \sum_{l=0}^{L} N_l K_l \lesssim \varepsilon^{-2}\log^2\varepsilon.$$

For optimal stopping, this is a special case of Theorem 5.1 in [5]. Their proof instantly generalizes to the present setting, relying on the L_2-bound (1.17) (and the L_1-bound it entails by Cauchy-Schwarz) in place of inequality (3.4) in [5].

1.4.2 Sieve Optimization Approach

Most of the literature studies dual martingales which are derived from approximations of Y^*. In contrast, the sieve optimization approach introduced and analyzed in [3] constructs dual martingales and thus upper bounds directly without first constructing an approximation of Y^*. The guiding idea is that the optimal martingale

M^* is variance-minimizing in the sense that $\theta_0^{up}(M^*) = Y_0^*$ is deterministic and that $\theta_0^{up}(M)$ has a positive variance for any suboptimal martingale M. A second, complimentary advantage of using martingales M for which $\theta_0^{up}(M)$ has a small variance is that Monte Carlo estimators

$$Y_0^N = \frac{1}{N}\sum_{n=1}^{N} \theta_0^{up,(n)}(M)$$

also have small variance and can be evaluated with relatively few samples N. Here, the random variables $\theta_0^{up,(n)}(M)$ are N independent copies of $\theta_0^{up}(M)$ which are derived from independent copies

$$(F_j^{(1)}, \beta_j^{(1)}, M_j^{(1)}), \ldots, (F_j^{(N)}, \beta_j^{(N)}, M_j^{(N)}), \quad j = 0, \ldots, J,$$

of the process (F, β, M). Consider the penalized optimization problem

$$\inf_{M \in \mathscr{M}_{D+1}} \left\{ E\left[\theta_0^{up}(M)\right] + \lambda \sqrt{\mathrm{Var}\left[\theta_0^{up}(M)\right]} \right\} \tag{1.18}$$

Since $\mathrm{Var}\left[\theta_0^{up}(M^*)\right] = 0$ and M^* minimizes $E[\theta_0^{up}(M^*)]$ within the set $\mathscr{M}_{D+1}$ of adapted $\mathbb{R}^{D+1}$-valued martingales with $M_0 = 0$, M^* solves (1.18) for any $\lambda > 0$.

Now let $\mathscr{M}_\Psi$ be a family of adapted martingales with $M_0 = 0$ and consider the empirical version

$$M_N := \arg \inf_{M \in \mathscr{M}_\Psi} \left(\frac{1}{N}\sum_{n=1}^{N} \theta_0^{up,(n)}(M) + \lambda \sqrt{V_N(M)} \right), \quad \lambda > 0, \tag{1.19}$$

of (1.18), where

$$V_N(M) := \frac{1}{N(N-1)} \sum_{1 \le n < m \le N} (\theta_0^{up,(n)}(M) - \theta_0^{up,(m)}(M))^2.$$

How large are the variance $\mathrm{Var}[\theta_0^{up}(M_N)]$ and the bias $E[\theta_0^{up}(M_N)] - Y_0^*$? Let (Ψ, ρ) be a metric space and let $\mathscr{M}_\Psi = \{M(\psi) : \psi \in \Psi\}$ be a parameterized family of adapted continuous $\mathbb{R}^{D+1}$-valued martingales. Denote by $\Psi^* \subset \Psi$ a set of functions ψ in Ψ satisfying

$$Y_0^* = \theta_0^{up,(n)}(M(\psi)) \quad \text{a.s..}$$

We assume that the set Ψ^* is not empty. If Ψ is infinite-dimensional, minimizing

$$Q_N(\psi) := \frac{1}{N}\sum_{n=1}^{N} \theta_0^{up,(n)}(M(\psi)) + \lambda \sqrt{V_N(M(\psi))}$$

over $\psi \in \Psi$ may not be well-defined; or even if a minimizer exists, it is generally difficult to compute, and may have undesirable large sample properties such as inconsistency and/or a very slow rate of convergence. These difficulties arise because the problem of optimization over an infinite-dimensional noncompact space may no longer be well-posed. The method of sieves provides one general approach to resolve the difficulties associated with minimizing $Q_N(\psi)$ over an infinite-dimensional space Ψ by minimizing $Q_N(\psi)$ over a sequence of approximating spaces Ψ_ν, called sieves, which are less complex but are dense in Ψ. Popular sieves are typically compact, nondecreasing $\Psi_\nu \subseteq \Psi_{\nu+1} \subseteq \ldots \subseteq \Psi$ and are such that for any $\psi \in \Psi$ there exists an element $\pi_\nu \psi$ in Ψ_ν satisfying $\rho(\psi, \pi_\nu \psi) \to 0$ as $\nu \to \infty$, where the notation π_ν can be regarded as a projection mapping from Ψ to Ψ_ν.

Now suppose that we are in a Markovian setting: We assume that $F_j(\cdot)$ only depends on ω through the value at time t_j of an M-dimensional Markov process (X_t), $0 = t_0 < t_1 < \ldots < t_J = T$. With a slight abuse of notation we write $F_j(\cdot) = F_j(X_{t_j}, \cdot)$ and assume that $F_j(x, \cdot)$ is Hölder in x. Moreover we assume that β_{j+1} is independent of $\mathscr{F}_j$ for all j, and that X_t solves the following system of SDEs:

$$dX_t = \mu(t, X_t)\, dt + \sigma(t, X_t)\, dW_t, \quad X_0 = x. \tag{1.20}$$

The coefficient functions $\mu : [0, T] \times \mathbb{R}^M \to \mathbb{R}^M$ and $\sigma : [0, T] \times \mathbb{R}^M \to \mathbb{R}^{M \times M}$ are supposed to be Lipschitz in space and $1/2$-Hölder continuous in time and $W = (W^1, \ldots, W^M)^\top$ is an M-dimensional Brownian motion.

It is well-known that under the assumption that a $\mathbb{R}^{D+1}$-valued martingale M_t is square integrable and is adapted to the filtration generated by W_t, there is a square integrable $\mathbb{R}^{(D+1)\times M}$-valued process $H_t = (H_t^{d,m}), d = 0, \ldots D, m = 1, \ldots, M$ satisfying

$$M_t = \int_0^t H_s\, dW_s. \tag{1.21}$$

It is not hard to see that in our Markovian setting, we have $H_s = \psi(s, X_s)$ for some matrix-valued function $\psi(s, x) = (\psi^{d,m}(s, x), d = 0, \ldots D, m = 1, \ldots, M)$ satisfying

$$\int_0^T E[|\psi(s, X_s)|^2]\, ds < \infty.$$

As a result,

$$M_t = M_t(\psi) = \int_0^t \psi(s, X_s)\, dW_s.$$

Thus, the set of adapted square-integrable martingales can be "parameterized" by the set $L_{2,P}([0,T]\times\mathbb{R}^M)$ of square-integrable $\mathbb{R}^{(D+1)\times M}$-valued functions ψ on $[0,T]\times\mathbb{R}^M$ that satisfy $\|\psi\|_{2,P}^2 := \int_0^T E[|\psi(s,X_s)|^2]\,ds < \infty$.

Consider linear sieves of the form:

$$\Psi_\Lambda := \{\alpha_1\phi_1 + \ldots + \alpha_\Lambda\phi_\Lambda : \alpha_1,\ldots,\alpha_\Lambda \in \mathbb{R}\}, \tag{1.22}$$

where $\phi_1,\ldots,\phi_\Lambda$ are some given matrix-valued functions with components from the space of bounded continuous functions $C_b([0,T]\times\mathbb{R}^M)$ and $\Lambda \in \mathbb{N}$. Next define a class of adapted square-integrable martingales via

$$\mathscr{M}_\Lambda := \{M_\cdot(\psi) : \psi \in \Psi_\Lambda\}$$

and set

$$M_N := \arg\inf_{M\in\mathscr{M}_{\Lambda_N}} \left(\frac{1}{N}\sum_{n=1}^N \theta^{up,(n)}(M) + (1+\lambda_N)\sqrt{V_N(M)}\right), \tag{1.23}$$

where $\Lambda_N \to \infty$ and $\lambda_N \to 0$ as $N\to\infty$. For the optimal stopping problem, it was shown in [3] that under a proper choice of λ_N and Λ_N

$$\max\left\{E[\theta_0^{up}(M_N)] - Y_0^*,\ \sqrt{\mathrm{Var}(\theta_0^{up}(M_N))}\right\} = O_P\left(\delta_N + \Lambda_N^{M+1}\log(\Lambda_N)/\sqrt{N}\right),$$

where $\delta_N = \inf_{\psi\in\Psi_{\Lambda_N},\,\psi^*\in\Psi^*}\rho(\psi,\psi^*)$. In particular, both the variance and the bias of $\theta_0^{up}(M_N)$ converge to 0 as $N\to\infty$.

1.5 Numerical Experiments

In the numerical examples we consider a BSDE of the form

$$d\mathscr{Y}_t = -\left(\left(a\sum_{d=1}^5 \mathscr{Z}_{d,t} - b\mathscr{Y}_t\right)_+ - r\mathscr{Y}_t\right)dt + \mathscr{Z}_t^\top dW_t, \quad \mathscr{Y}_T = h\left(\max_{d=1,\ldots,5} X_{d,T}\right) \tag{1.24}$$

for constants $a,b,r\in\mathbb{R}$ and a Lipschitz continuous function $h:\mathbb{R}\to\mathbb{R}$. Here W is a five-dimensional Brownian motion and $X_{d,t}$, $d=1,\ldots,5$, are independent, identically distributed geometric Brownian motions with drift $\mu\in\mathbb{R}$, volatility $\sigma>0$, and initial value $x_0>0$, i.e.

$$X_{d,t} = x_0\exp\left\{\sigma W_{d,t} + (\mu-\sigma^2/2)t\right\}, \quad d=1,\ldots,D, \quad t\in[0,T].$$

This setting covers the funding risk example, explained in the Introduction, for an option with payoff function h on the maximum of five Black-Scholes stocks with $a = R/\sigma$, $b = R$, and $\mu = r$. Moreover, the example on pricing under credit value adjustment for an option with payoff function $-h$ on the maximum of five Black-Scholes stocks can be accommodated by setting $a = 0$, $b = \lambda(R - 1)$, $\mu = r$ and noting that the option price is given by $-\mathscr{Y}_0$. Finally we can recover a *g-expectation* [22] related to drift uncertainty of the form

$$\mathscr{Y}_0 = \sup_b E\left[h\left(\max_{d=1,\dots,5} x_0 \exp\left\{\sigma W_{d,T} + \int_0^T (b(s) - \sigma^2/2)ds\right\}\right)\right], \tag{1.25}$$

where the sup runs over the set of adapted processes b such that $\mu \le b(t) \le \mu + R$ by letting $b = r = 0$ and $a = R/\sigma$.

Given a time grid $0 = t_0 \le t_1 < \cdots < t_J = T$, a natural time discretization for $\mathscr{Y}_{t_j}$ is given by

$$\begin{aligned} Y_j^* &= E_j[Y_{j+1}^*] - \left(\left(a\sum_{d=1}^5 E_j\left[\frac{\Delta W_{d,j}}{\Delta_j} Y_{j+1}^*\right] - bE_j[Y_{j+1}^*]\right)_+ - rE_j[Y_{j+1}^*]\right)\Delta_j \\ Y_J^* &= h\left(\max_{d=1,\dots,5} X_{d,T}\right), \end{aligned}$$

where $\Delta_j = t_{j+1} - t_j$, and the Brownian increments $\Delta W_{d,j} = W_{d,t_{j+1}} - W_{d,t_j}$ can be truncated appropriately, whenever necessary for theoretical reasons. It can be deduced from the results in [11] and [28] that the time discretization error of this scheme converges at a rate of 1/2 in the mesh size of the time grid. Precisely, there is a constant C such that

$$\max_j E[|\mathscr{Y}_{t_j} - Y_j^*|^2] + \sum_j E\left[\int_{t_j}^{t_{j+1}} \left|\mathscr{Z}_s - E_j\left[\frac{\Delta W_j}{\Delta_j} Y_{j+1}^*\right]\right|^2 ds\right] \le C(\max_j \Delta_j).$$

Example 1.1 (g-expectation). The first example is in the context of g-expectation, as described in (1.25), with parameters $x_0 = 100$, $\mu = 0.01$, $\sigma = 0.2$, $R = 0.05$, and $T = 1/4$. For the terminal condition we choose the payoff function of a call spread option $h(x) = (x - 95)_+ - 2(x - 115)_+$. We first compare the performance of the martingale basis algorithm with the standard least-squares Monte Carlo algorithm of Lemor, Gobet, and Warin [20]. For both algorithms we choose the same function basis, namely the martingale basis computed from $\eta_{0,J} = (1, h(\max_{d=1,\dots,5} X_{d,T}))$. The corresponding approximations will be denoted by

$$(Y_j^{(MB,N)}, Z_{d,j}^{(MB,N)}), \quad \text{resp. } (Y_j^{(LGW,N)}, Z_{d,j}^{(LGW,N)}), \quad d = 1, \dots, D,$$

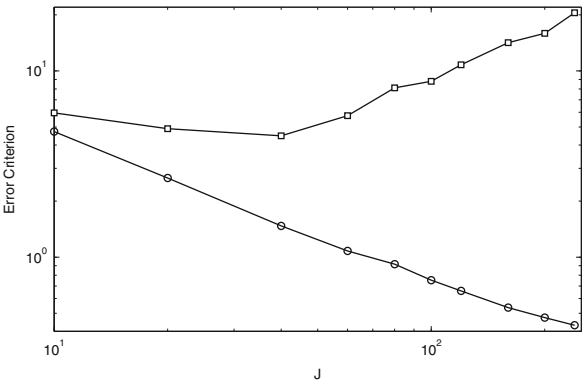


Fig. 1.1 Error criterion for the martingale basis algorithm with 100 regression paths (*circles*) and for the Lemor-Gobet-Warin algorithm with 10.000 regression paths (*boxes*) as the number of time steps J increases from 10 to 240

for the martingale basis algorithm and for the Lemor-Gobet-Warin algorithm, respectively. Here, the superscript N makes the dependence on the number of regression paths explicit. The time grid is a partition of the interval $[0, T]$ into J subintervals of equal length. For the comparison we apply the a-posteriori estimates of [9]. By Example 3.6 in [9], there are constants $c_1, c_2 > 0$ independent of the time grid such that for all square integrable processes $(Y_j, Z_{d,j})$, $d = 1, \ldots, D$, with $Y_J = \mathscr{Y}_T$

$$\frac{1}{c_1}\mathscr{E}(Y,Z) \le \max_j E[|\mathscr{Y}_{t_j} - Y_j|^2] + \sum_j E\left[\int_{t_j}^{t_{j+1}} \left|\mathscr{Z}_s - Z_j\right|^2 ds\right] \le c_2 \mathscr{E}(Y,Z). \tag{1.26}$$

Here the error criterion $\mathscr{E}(Y, Z)$ can be calculated from the discrete time approximation (Y, Z) as

$$\mathscr{E}(Y,Z) = \max_{1\le i\le J} E\left[\left|Y_i - Y_0 - \sum_{j=0}^{i-1}\left(\frac{R}{\sigma}\sum_{d=1}^{D} Z_{d,j}\right)_+ \Delta_j - \sum_{j=0}^{i-1} Z_j^\top \Delta W_j\right|^2\right]. \tag{1.27}$$

We note that (1.26) and (1.27) still hold true for the above least-squares Monte Carlo approximations, when the expectation is taken conditionally on the regression paths.

Figure 1.1 is a log-log-plot of the error criterion with the expectations replaced by sample means over 10.000 outer paths for the approximations

$$(Y_j^{(MB,100)}, Z_{d,j}^{(MB,100)}) \quad \text{and} \quad (Y_j^{(LGW,10.000)}, Z_{d,j}^{(LGW,10.000)})$$

as number of time steps J increases. For the martingale basis algorithm we observe an almost linear decay with a slope of about -0.8 for up to 240 time steps. By

Eq. (1.26) this indicates that the very cheap and fixed estimator for the conditional expectation consisting of 2 basis functions and 100 regression paths is sufficient to make the L^2-approximation error decrease at a rate of about 0.4 as the time discretization becomes as fine as about 10^{-3}. Contrarily, for the LGW algorithm the simulation error dominates and the L^2-error does not decrease, although the number of regression paths is increased by a factor of 100 compared to the martingale basis algorithm. We stress that this comparison is only meant to illustrate the strong variance reduction effect of the martingale basis algorithm. One can achieve, in principle, the same decay as for the martingale basis algorithm with the LGW algorithm, if the number of regression paths increases polynomially with the number of time steps, see [20] for the theoretical background and [9] for numerical examples. This example illustrates that exploiting martingale basis functions may be highly beneficial in the BSDE context, if a good set of such basis functions is at hand.

Table 1.1 shows lower and upper bounds based on the primal-dual approach in Theorems 1.1 and 1.2 as the time discretization becomes finer. The lower bounds are calculated as explained in Sect. 1.3 applying the martingale basis algorithm (with the same specification as above) to solve the dynamic program approximately. The upper bounds are computed with the Andersen-Broadie type algorithm described in Sect. 1.4.1, i.e. the 6-dimensional Doob martingale corresponding to the martingale basis approximation of the dynamic program is estimated by inner simulations. We apply 100 inner paths and 10.000 outer paths, and refer to [7] for more details on the general implementation including the use of control variates. Table 1.1 demonstrates that very tight 95 % confidence intervals can be constructed for the solution Y_0^* of the discretized BSDE for up to $J = 160$ time discretization steps, although the BSDE is five-dimensional and depends on the control part $\mathscr{Z}$ of the solution. Indeed, the relative width of the 95 % confidence intervals ranges from less than 0.4 % for $J = 40$ time steps to still less than 1 % for $J = 160$ time steps.

Example 1.2 (Credit value adjustment). In the second numerical example we test the sieve optimization algorithm for a pricing problem with credit value adjustment. As parameters we choose $x_0 = 100$, $\mu = 0.02$, $\sigma = 0.2$, $\lambda = 0.02$, $T = 2$ and recovery rates $R = 0.1, 0.5, 0.9$. The time grid is equidistant on $[0, 2]$ with $J = 50$ time steps. The payoff function $-h$ corresponds to a call spread option given by

$$-h(x) = 2(x - 115)_+ - (x - 95)_+.$$

Table 1.1 g-expectation for different time discretizations. Standard deviations are in brackets

J	40	80	120	160
Lower bound	13.936 (0.003)	13.935 (0.003)	13.941 (0.003)	13.942 (0.003)
Upper bound	13.976 (0.003)	14.001 (0.003)	14.033 (0.003)	14.061 (0.003)

The linear sieve in (1.22) with $\Lambda = 40$ is defined in terms of trigonometric functions. Precisely, let

$$\text{trig}^k(x) = \begin{cases} \cos(\frac{k}{2}x), & k \text{ even}, \\ \sin(\frac{k+1}{2}x), & k \text{ odd}. \end{cases}$$

Denote by $\iota_t(1) \in \{1, \ldots 5\}$ the index of the highest asset at time t, and by $\iota_t(2)$ the index of the second highest. For $k = 0, \ldots, 19$, choose

$$\phi_d^k(x,t) = \begin{cases} \text{trig}^k\left(\frac{X_t^{\iota_t(1)}}{100}\right) X_t^{\iota_t(1)}, & d = \iota_t(1) \\ 0, & \text{otherwise}. \end{cases}$$

For $k = 20, \ldots, 39$, define

$$\phi_d^k(x,t) = \begin{cases} \text{trig}^{k-20}\left(\frac{X_t^{\iota_t(1)} - X_t^{\iota_t(2)}}{10}\right) X_t^{\iota_t(1)}, & d = \iota_t(1) \\ -\text{trig}^{k-20}\left(\frac{X_t^{\iota_t(1)} - X_t^{\iota_t(2)}}{10}\right) X_t^{\iota_t(2)}, & d = \iota_t(2) \\ 0, & \text{otherwise}. \end{cases}$$

The optimization in (1.23) is performed with $N = 5.000$ sample paths in order to construct the martingale M_N. Given this martingale a new set of one million outer paths is applied to estimate the upper bound for Y_0^* (and hence a lower bound on the option price $-Y_0^*$). In order to evaluate the quality of these upper bounds, we compute lower bounds on Y_0^* (and hence upper bounds on the option price $-Y_0^*$) based on the martingale basis algorithm as described in the previous example.

Table 1.2 presents the sieve optimization lower bounds on the option price $-Y_0^*$ with credit value adjustment and the primal upper bounds for different recovery rates. The problem becomes more difficult for smaller recovery rates, as these lead to higher Lipschitz constants in the generator of the BSDE. Nonetheless, even for $R = 0.1$ the relative width of the corresponding 95 % confidence interval [8,570; 8,597] is just 0.3 %. Hence, the sieve optimization algorithm provides excellent price bounds in this example with a rather generic choice of the linear sieve and without the requirement of nested simulation, which makes it way faster than the Andersen-Broadie type algorithm.

Table 1.2 Price bounds for different recovery rates. Standard deviations are in brackets

Recovery	0.1	0.5	0.9
Lower bound	8.554 (0.008)	8.725 (0.008)	8.895 (0.008)
Upper bound	8.573 (0.012)	8.730 (0.012)	8.903 (0.012)

References

1. Andersen, L., Broadie, M.: Primal-dual simulation algorithm for pricing multidimensional American options. Manage. Sci. **50**(9), 1222–1234 (2004)
2. Belomestny, D.: Pricing Bermudan options by nonparametric regression: optimal rates of convergence for lower estimates. Finance Stoch. **15**(4), 655–683 (2011)
3. Belomestny, D.: Solving optimal stopping problems by empirical dual optimization. Ann. Appl. Prob. **23**(5), 1988–2019 (2013)
4. Belomestny, D., Dickmann, F., Nagapetyan, T.: Pricing American options via multi-level approximation methods. arXiv preprint 1303.1334 (2013)
5. Belomestny, D., Schoenmakers, J., Dickmann, F.: Multilevel dual approach for pricing American style derivatives. Finance Stoch. **17**, 717–742 (2013)
6. Bender, C., Schoenmakers, J., Zhang, J.: Dual representations for general multiple stopping problems. Math. Finance (2013). doi: 10.1111/mafi.12030
7. Bender, C., Schweizer, N., Zhuo., J.: A primal-dual algorithm for BSDEs. arXiv preprint 1310.3694 (2013)
8. Bender, C., Steiner, J.: Least-squares Monte Carlo for backward SDEs. In: Numerical Methods in Finance, pp. 257–289. Springer, Berlin (2012)
9. Bender, C., Steiner, J.: A posteriori estimates for backward SDEs. SIAM/ASA J. Uncertain. Quantif. **1**(1), 139–163 (2013)
10. Bouchard, B., Chassagneux, J.F.: Discrete-time approximation for continuously and discretely reflected BSDEs. Stoch. Process. Appl. **118**(12), 2269–2293 (2008)
11. Bouchard, B., Touzi, N.: Discrete-time approximation and Monte-Carlo simulation of backward stochastic differential equations. Stoch. Process. Appl. **111**(2), 175–206 (2004)
12. Brown, D.B., Smith, J.E., Sun, P.: Information relaxations and duality in stochastic dynamic programs. Oper. Res. **58**(4), 785–801 (2010)
13. Crépey, S.: Bilateral counterparty risk under funding constraints – Part II. CVA. Math. Finance (2012). doi: 10.1111/mafi.12005
14. Duffie, D., Schroder, M., Skiadas, C.: Recursive valuation of defaultable securities and the timing of resolution of uncertainty. Ann. Appl. Probab. **6**(4), 1075–1090 (1996)
15. Fahim, A., Touzi, N., Warin, X.: A probabilistic numerical method for fully nonlinear parabolic PDEs. Ann. Appl. Probab. **21**(4), 1322–1364 (2011)
16. Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res. **56**(3), 607–617 (2008)
17. Glasserman, P., Yu, B.: Simulation for American options: regression now or regression later? In: Niederreiter, H. (ed.) Monte Carlo and Quasi-Monte Carlo Methods 2002, pp. 213–226. Springer, Berlin (2004)
18. Haugh, M.B., Kogan, L.: Pricing American options: a duality approach. Oper. Res. **52**(2), 258–270 (2004)
19. Karoui, N.E., Peng, S., Quenez, M.C.: Backward stochastic differential equations in finance. Math. Finance **7**, 1–71 (1997)
20. Lemor, J., Gobet, E., Warin, X.: Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations. Bernoulli **12**(5), 889–916 (2006)
21. Longstaff, F.A., Schwartz, E.S.: Valuing American options by simulation: a simple least-squares approach. Rev. Financ. Stud. **14**, 113–147 (2001)
22. Peng, S.: Nonlinear expectations, nonlinear evaluations and risk measures. In: Stochastic Methods in Finance. Lecture Notes in Mathematics, vol. 1856, pp. 165–253. Springer, Berlin (2004)
23. Rogers, L.: Monte Carlo valuation of American options. Math. Finance **12**(3), 271–286 (2002)
24. Schoenmakers, J.: A pure martingale dual for multiple stopping. Finance Stoch. **16**(2), 319–334 (2012)
25. Schweizer, M.: On Bermudan options. In: Advances in Finance and Stochastics. Essays in Honour of Dieter Sondermann, pp. 257–270. Springer, Berlin (2002)

26. Tsitsiklis, J.N., Roy, B.V.: Regression methods for pricing complex American-style options. IEEE Trans. Neur. Netw. **12**(4), 694–703 (2001)
27. Zanger, D.Z.: Quantitative error estimates for a least-squares Monte Carlo algorithm for American option pricing. Finance Stoch. **17**(3), 503–534 (2013)
28. Zhang, J.: A numerical scheme for BSDEs. Ann. Appl. Probab. **14**(1), 459–488 (2004)

Chapter 2
Efficient Resolution of Anisotropic Structures

Wolfgang Dahmen, Chunyan Huang, Gitta Kutyniok, Wang-Q Lim, Christoph Schwab, and Gerrit Welper

Abstract We highlight some results obtained in the DFG-SPP project "Adaptive Anisotropic Discretization Concepts". We focus on new developments concerning the *sparse representation* of possibly *high-dimensional* functions exhibiting strong *anisotropic* features and low regularity in isotropic Sobolev or Besov scales. Specifically, we focus on the solution of *transport equations* which exhibit *propagation of singularities* where, additionally, high-dimensionality enters when the convection field, and hence the solutions, depend on parameters varying over some compact set. Important constituents of our approach are directionally adaptive discretization concepts motivated by compactly supported shearlet systems, and well-conditioned stable variational formulations that support trial spaces with anisotropic refinements with arbitrary directionalities. We prove that they provide tight error-residual relations which are used to contrive rigorously founded adaptive refinement schemes which converge in L_2. Moreover, in the context of parameter dependent problems we discuss two approaches serving different purposes and working under different regularity assumptions. For "frequent query problems", making essential use of the novel well-conditioned variational formulations, a new *Reduced Basis Method* is outlined which exhibits a certain *rate-optimal*

W. Dahmen (✉)
RWTH Aachen, Templergraben 55, 52056 Aachen, Germany
e-mail: dahmen@igpm.rwth-aachen.de

C. Huang
School of Applied Mathematics, Central University of Finance and Economics,
39 South College Road, Haidian District Beijing, 100081 P.R. China
e-mail: hcy@cufe.edu.cn

G. Kutyniok • W.-Q. Lim
Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: kutyniok@math.tu-berlin.de; lim@math.tu-berlin.de

C. Schwab
ETH Zürich, HG G 57.1, Rämistr. 101, 8092 Zürich, Switzerland
e-mail: christoph.schwab@sam.math.ethz.ch

G. Welper
Texas A & M University, Mailstop 3368, College Station, TX 77843-3368, USA
e-mail: welper@math.tamu.edu

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_2

performance for indefinite, unsymmetric or singularly perturbed problems. For the radiative transfer problem with scattering a *sparse tensor* method is presented which mitigates or even overcomes the curse of dimensionality under suitable (so far still isotropic) regularity assumptions. Numerical examples for both methods illustrate the theoretical findings.

2.1 Introduction

The more complex a data site or mathematical model is the more adapted a corresponding mathematical representation needs to be in order to capture its information content at acceptable cost in terms of storage and computational complexity. In principle, this is true for mathematical objects described explicitly by large sets of possibly noisy or corrupted data but also for those given only implicitly as the solution of an operator equation. The latter scenario is perhaps even more challenging because direct observations are not possible. By "adapted representation" we mean a representation of the unknown function that exploits possibly global features of this function so as to require, for a prescribed target accuracy, only relatively few parameters to determine a corresponding approximation. Such global features could take a variety of forms such as (i) a high degree of regularity except at isolated singularities *located on lower dimensional manifolds*, or (ii) a particular *sparsity* possibly with respect to a *dictionary* which may even depend on the problem at hand. In fact, corresponding scenarios are not strictly disjoint. In either case reconstruction or approximation methods are necessarily nonlinear. For instance, as for (i), 1D *best N-term* wavelet approximations offer a powerful method based on selecting only possible few coefficients in an exact representation with respect to a given *universal background* dictionary, e.g. a wavelet basis. When dealing with more than one spatial variable the situation quickly becomes more complicated and for spatial dimensions much larger than three, classical numerical tools designed for the low dimensional regime become practically useless. This is commonly referred to as *curse of dimensionality*. Unfortunately, there seems to be no universal strategy of dealing with the curse of dimensionality, i.e., that works in all possible cases.

One global structural feature which is encountered in many multivariate scenarios is *anisotropy*: images, as functions of two variables, exhibit edges and discontinuities along curves. Higher dimensional biological images have sharp interfaces separating more homogeneous regions. Likewise highly anisotropic phenomena such as shear- or boundary layers are encountered in solutions to transport dominated initial-boundary value problems.

One major focus of the DFG-SPP project "Adaptive Anisotropic Discretization Concepts" has been to efficiently recover and economically encode *anisotropic* structures represented by explicitly given data or determined as solutions of operator equations which are prone to give rise to such structures. Regarding this latter case, which we will focus on in this article, *parametric transport problems* (as well as

close relatives) have served as guiding model problems for the following reasons: (i) their solutions could exhibit shear or boundary layers and hence discontinuities across lower dimensional manifolds calling for suitable *anisotropic discretizations*; (ii) how to contrive suitable *variational formulations*, which in particular accommodate such anisotropic discretizations is much less clear than in the elliptic case; (iii) *parametric versions* give rise to *high-dimensional* problems.

Concerning (i), *directional representation systems* like *curvelets* and *shearlets* outperform classical *isotropic* wavelet bases when approximating so called "cartoon images", see [23] and [9, 31–35]. For recent applications to imaging data, in particular, *inpainting* as well as in combination with *geometric separation concepts* the reader is referred to [24,29]. In the present context of solving operator equations we outline in Sect. 2.2 trial spaces which accommodate directional adaptivity. They are motivated by recent constructions of compactly supported *piecewise polynomial shearlet systems* (see e.g. [30]) because they are close to classical multiresolution structures and similar in nature to classical discretization systems. Since cartoons exhibit structural similarities with the solution to transport problems we state best N-term error bounds for cartoon functions that will later serve as benchmarks for an adaptive solver. For related anisotropic simplicial discretizations and their analysis see e.g. [10, 12, 15].

As for (ii), our approach differs from previous works on anisotropic discretizations derived from "curvature information" on the current approximation and hence not based on a rigorous error control (see e.g. [22] and the references therein), in that we derive first in Sect. 2.3 *well conditioned variational formulations* for general unsymmetric or indefinite and singularly perturbed problems, see [14,16] for details on *convection-diffusion* and *transport problems*. The underlying basic principles are of independent interest by themselves and seem to have appeared first in [3]. They are also closely related to ongoing developments running under the flag of *Discontinuous Petrov Galerkin (DPG) Methods*, see e.g. [19, 20]. The approach is motivated by two crucial corner stones. On the one hand, one can essentially *choose* the norm for the (infinite dimensional) trial space X by which one would like to measure accuracy while adapting the norm for the (infinite dimensional) test space Y so as to ensure that (ideally) the operator induced by this variational formulation is even an isometry from X to Y' (the normed dual of Y). Numerical feasibility of (nearly optimal) Petrov Galerkin discretizations based on such formulations, even beyond a DPG framework, hinges on an appropriate *saddle point formulation* which turns out to be actually crucial in connection with *model reduction* [18]. On the one hand, this allows one to accommodate, for instance, L_2-frames. On the other hand, the resulting tight error-residual relation is the basis of computable a-posteriori error estimators [14, 16] and, ultimately, to rigorously founded adaptive anisotropic refinement strategies.

These variational formulations apply in much more generality but in order to address issue (iii) we exemplify them for the simple linear transport equation (stationary or instationary) whose *parametric* version leads to high-dimensional problems and forms a core constituent of kinetic models such as *radiative transport*. There the transport direction – the parameter – varies over a unit sphere so that

solutions are functions of the spatial variables (and, possibly, of time) and of the transport direction.

We briefly highlight two ways of treating such parametric problems under slightly different objectives. Both strategies aim at approximating the solution $u(x, \boldsymbol{s})$, $x \in \Omega \subset \mathbb{R}^d$, $\boldsymbol{s} \in S^{d-1}$, in the form

$$u(x, \boldsymbol{s}) \approx \sum_{j=1}^{n} c_j(\boldsymbol{s}) u_j(x). \tag{2.1}$$

In Sect. 2.4 the u_j are constructed *offline* in a greedy manner from *snapshots* of the solution manifold, thus forming a *solution dependent* dictionary. According to the paradigm of the *Reduced Basis Method* (RBM) the parameter dependent coefficients $c_j(\boldsymbol{s})$ are not given explicitly but can be efficiently computed in an *online* fashion, e.g. in the context of design or (online) optimization. This approach works the better the smoother the dependence of the solution on the parameters is so that the Kolmogorov n-widths decay rapidly with increasing n. Making essential use of the well conditioned variational formulations from Sect. 2.3, it can be shown that the resulting RBM has stability constants as close to one as one wishes yielding for the first time an RBM for transport and convection-diffusion problems with this property exhibiting the same rates as the Kolmogorov widths [18].

In Sect. 2.5 of this report, and in [26], we present algorithms which construct *explicitly* separable approximations of the form (2.1) for the parametric transport problem of radiative transfer. We also mention that separable approximations such as (2.1) arise in a host of other applications; for example, in parametric representations of PDEs with random field input data with the aid of sparse tensor product interpolation methods; we refer to [11, 13] and to the references therein. Adaptive near-minimal rank tensor solvers for problems in high dimensional phase space are established and analyzed in [2].

2.2 Anisotropic Approximations

Let $D = (0,1)^2$ and let $\mathrm{curv}(\partial\Omega)$ denote the curvature of $\partial\Omega \cap D$. The class of *cartoon-like functions* on $D = (0,1)^2$,

$$\begin{aligned} \mathscr{C}(\zeta, L, M, D) := \{ f_1 \chi_\Omega + f_2 \chi_{D\setminus\Omega} : \ \Omega \subset D,\ |\partial\Omega \cap D| \le L, \partial\Omega \cap D \in C^2, \\ \mathrm{curv}(\partial\Omega) \le \zeta, \|f_i^{(l)}\|_{L_\infty(D)} \le M,\ l \le 2,\ i = 1,2\}, \end{aligned} \tag{2.2}$$

(where the parameters ζ, L are not mutually independent) has become a well accepted benchmark for sparse approximation in imaging [23]. Compactly supported shearlet systems for $L^2(\mathbb{R}^2)$ have been introduced in [30, 33] to provide

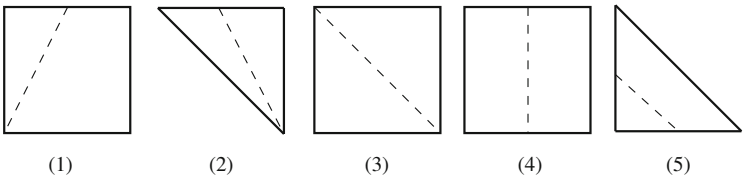


Fig. 2.1 Illustration of the partition rules

(near-) optimal sparse approximations for such classes. We observe that such cartoons also exhibit similar features as solutions to transport problems.

Unfortunately, even compactly supported shearlets do not comply well with quadrature and boundary adaptation tasks faced in variational methods for PDEs. We are therefore interested in generating locally refinable anisotropic partitions for which corresponding piecewise polynomial approximations realize the favorable near-optimal approximation rates for cartoon functions achieved by shearlet systems. Unfortunately, as shown in [36, Chapter 9.3], simple triangular bisections connecting the midpoint of an edge to the opposite vertex is not sufficient for warranting such rates, see [10, 15] for related work. In fact, a key feature would be to realize a "parabolic scaling law" similar to the shearlet setting. By this we mean a sufficient rapid directional resolution by anisotropic cells whose width scales like the square of the diameter. To achieve this we consider partitions comprised of triangles *and* quadrilaterals pointed out to us in Cohen and Mirebeau (2013, private communication). We sketch the main ideas and refer to [17] for details.

Starting from some initial partition consisting of triangles and quadrilaterals, *refined partitions* are obtained by splitting a given cell Q of a current partition according to one of the following rules:

(i) Connect a vertex with the midpoint of an edge not containing the vertex.
(ii) Connect two vertices.
(iii) Connect the midpoints of two edges which, when Q is a quadrilateral, do not share any vertex.

The types of bisections are indicated in Fig. 2.1: (1), (2) are examples of (i), (3) illustrates (ii), and (4), (5) are examples for (iii). One easily checks that these refinement rules produce only triangles and quadrilaterals. Moreover, a quadrilateral can be bisected in eight possible ways whereas a triangle can be split in six possible ways. Assigning to each split type a number in $I_Q = \{1, \dots, 8\}$ when Q is a quadrilateral and a number in $I_Q = \{9, \dots, 14\}$ when Q is a triangle, we denote by

$$R_{\iota_Q}(Q) = \{Q_1, Q_2\} \quad \text{for some } \iota_Q \in I_Q, \tag{2.3}$$

the refinement operator which replaces the cell Q by its two children Q_1, Q_2 generated, according to the choice ι_Q, by the above split rules (i)–(iii).

For any partition $\mathscr{G}$ of D, let $\mathbb{P}_1(\mathscr{G}) = \{v \in L_2(D) : v|Q \in \mathbb{P}_1, Q \in \mathscr{G}\}$ be the space of piecewise affine functions on $\mathscr{G}$ and denote by $\mathfrak{G}$ the set of all finite partitions that can be created by successive applications of R_{ι_Q} to define then

$$\Sigma_N := \bigcup \{\mathbb{P}_1(\mathscr{G}) : \mathscr{G} \in \mathfrak{G}, \#(\mathscr{G}) \leq N\}.$$

The next result from [17] shows that approximations by elements of Σ_N realize (and even slightly improve on) the known rates obtained for shearlet systems for the class of cartoon-like functions [33].

Theorem 2.1 ([17]). *Let $f \in \mathscr{C}(\zeta, L, M, D)$ with $D = (0,1)^2$ and assume that the discontinuity curve $\Gamma = \partial\Omega \cap D$ is the graph of a C^2-function. Then one has*

$$\inf_{\varphi \in \Sigma_N} \|f - \varphi\|_{L_2(D)} \leq C(\zeta, L) M \, N^{-1} \log N,$$

where $C(\zeta, L)$ is an absolute constant depending only on ζ, L.

The proof of Theorem 2.1 is based on constructing a specific sequence $\mathscr{C}_j$ of admissible partitions from $\mathfrak{G}$ where the refinement decisions represented by R_{ι_Q} use full knowledge of the approximated function f. A similar sequence of partitions is employed in Sect. 2.3.4.2 where $\iota_Q \in I_Q$, however, results from an a posteriori criterion described later below. We close this section by a few remarks on the structure of the $\mathscr{C}_j$. Given $\mathscr{C}_{j-1}$, we first generate

$$\tilde{\mathscr{C}}_j = \{Q' \in \tilde{R}(Q) : Q \in \mathscr{C}_{j-1}\}, \tag{2.4}$$

where $\tilde{R}$ is either R_{ι_Q} or the identity. To avoid unnecessary refinements we define then $\mathscr{C}_j$ by replacing any pair of triangles $Q, Q' \in \tilde{\mathscr{C}}_j$ whose union forms a parallelogram P by P itself. This reduces the number of triangles in favor of parallelograms.

2.3 Well-Conditioned Stable Variational Formulations

In this section we highlight some new conceptual developments from [14, 16, 18] which are, in particular, relevant for the high dimensional parametric problems addressed later below.

2.3.1 The General Principles

Anisotropic structures are already exhibited by solutions of elliptic boundary value problems on polyhedral domains in 3D. However, related singularities are known a priori and can be dealt with by anisotropic *preset* mesh refinements. Anisotropic

structures of solutions to transport dominated problems can be less predictable so that a quest for *adaptive anisotropic* discretization principles gains more weight. Recall that every known rigorously founded adaptation strategy hinges in one way or the other on being able to relate a current error of an approximate solution to the corresponding *residual* in a suitable norm. While classical variational formulations of elliptic problems grant exactly such an error-residual relation, this is unclear for transport dominated problems. The first fundamental issue is therefore to *find* also for such problems suitable variational formulations yielding a well conditioned error-residual relation.

2.3.1.1 Abstract Petrov-Galerkin Formulation

Suppose that for a pair of Hilbert spaces X, Y (with scalar products $(\cdot,\cdot)_X, (\cdot,\cdot)_Y$ and norms $\|\cdot\|_X, \|\cdot\|_Y$), and a given bilinear form $b(\cdot,\cdot) : X \times Y \to \mathbb{R}$, the problem

$$b(u,v) = f(v), \quad v \in Y, \tag{2.5}$$

has for any $f \in Y'$ (the normed dual of Y) a unique solution $u \in X$. It is well-known that this is equivalent to the existence of constants $0 < c_b \leq C_b < \infty$ such that

$$\sup_{w\in X}\sup_{v\in Y}\frac{b(w,v)}{\|w\|_X\|v\|_Y} \leq C_b, \quad \inf_{w\in X}\sup_{v\in Y}\frac{b(v,w)}{\|w\|_X\|v\|_Y} \geq c_b, \tag{2.6}$$

and that, for each $v \in Y$, there exists a $w \in X$ such that $b(w,v) \neq 0$. This means that the operator $B : X \to Y'$, defined by $(Bu)(v) := b(u,v)$, $u \in X, v \in Y$, is an isomorphism with condition number $\kappa_{X,Y}(B) := \|B\|_{\mathscr{L}(X,Y')}\|B^{-1}\|_{\mathscr{L}(Y',X)} \leq C_b/c_b$. For instance, when (2.5) represents a *convection dominated convection-diffusion problem* with the classical choice $X = Y = H^1_0(\Omega)$, the quotient C_b/c_b becomes very large. Since

$$\|B\|^{-1}_{\mathscr{L}(X,Y')}\|Bv - f\|_{Y'} \leq \|u - v\|_X \leq \|B^{-1}\|_{\mathscr{L}(Y',X)}\|Bv - f\|_{Y'}, \tag{2.7}$$

the error $\|u - v\|_X$ can then *not* be tightly estimated by the residual $\|Bv - f\|_{Y'}$.

2.3.1.2 Renormation

On an abstract level the following principle has surfaced in a number of different contexts such as *least squares methods* (see e.g. [5, 8]) and the so-called, more recently emerged *Discontinuous Petrov Galerkin (DPG) methods*, see e.g. [3, 16, 19, 20] and the references therein. The idea is to fix a norm, $\|\cdot\|_Y$, say, and modify

the norm for X so that the corresponding operator even becomes an *isometry*. More precisely, define

$$\|u\|_{\hat{X}} := \sup_{v\in Y} \frac{b(u,v)}{\|v\|_Y} = \|Bu\|_{Y'} = \|R_Y^{-1}Bu\|_Y, \tag{2.8}$$

where $R_Y : Y \to Y'$ is the Riesz map defined by $(v,z)_Y = (R_Y v)(z)$. The following fact is readily verified, see e.g. [16, 40].

Remark 2.1. One has $\kappa_{\hat{X},Y}(B) = 1$, i.e., (2.6) holds with $c_b = C_b = 1$ when $\|\cdot\|_X$ is replaced by $\|\cdot\|_{\hat{X}}$.

Alternatively, fixing X and redefining $\|\cdot\|_Y$ by $\|v\|_{\hat{Y}} := \|B^*v\|_{X'}$, one has $\kappa_{X,\hat{Y}}(B) = 1$, see [16]. Both possibilities lead to the error residual relations

$$\|u-w\|_X = \|f - Bw\|_{\hat{Y}'}, \quad \|u-w\|_{\hat{X}} = \|f-Bw\|_{Y'}, \quad u,w \in X. \tag{2.9}$$

2.3.2 Transport Equations

Several variants of these principles are applied and analyzed in detail in [14] for convection-diffusion equations. We concentrate in what follows on the limit case for vanishing viscosity, namely pure transport equations. For simplicity we consider the domain $D = (0,1)^d, d = 1,2,3$, with $\Gamma := \partial D$, denoting as usual by $\boldsymbol{n} = \boldsymbol{n}(x)$ the unit outward normal at $x \in \Gamma$ (excluding the four corners, of course). Moreover, we consider velocity fields $\boldsymbol{b}(x)$, $x \in D$, which for simplicity will always be assumed to be differentiable, i.e., $\boldsymbol{b}(x) \in C^1(\overline{D})^d$. Likewise $c(x) \in C^0(\overline{D})$ will serve as the reaction term in the *first order transport equation*

$$\boldsymbol{b}\cdot\nabla u + cu = f_\circ \text{ in } D\,, \quad u = g \text{ on } \Gamma_-\,, \tag{2.10}$$

where $\Gamma_\pm := \{x \in \partial D : \pm\boldsymbol{b}(x)\cdot\boldsymbol{n}(x) > 0\}$ denotes the *inflow, outflow boundary*, respectively. Furthermore, to simplify the exposition we shall always assume that $2c - \nabla\cdot\boldsymbol{b} \geq c_0 > 0$ in D holds.

A priori there does not seem to be any "natural" variational formulation. Nevertheless, the above principle can be invoked as follows. Following e.g. [16], one can show that the associated bilinear form with derivatives on the test functions

$$b(w,v) := \int_D w(-\boldsymbol{b}\cdot\nabla v + v(c - \nabla\cdot\boldsymbol{b}))\,dx, \tag{2.11}$$

is trivially bounded on $L_2(D)\times W_0(-\boldsymbol{b},D)$, where

$$W_0(\mp\boldsymbol{b},D) := \mathrm{clos}_{\|\cdot\|_{W(\boldsymbol{b},D)}}\{v \in C^1(D)\cap C(\overline{D}),\ v\,|_{\Gamma_\pm} \equiv 0\}, \tag{2.12}$$

and

$$\|v\|_{W(\boldsymbol{b},D)} := \left(\|v\|^2_{L_2(D)} + \int_D |\boldsymbol{b}\cdot\nabla v|^2\, dx \right)^{1/2}. \tag{2.13}$$

Moreover, the trace $\gamma_-(v)$ exists for $v \in W_0(\boldsymbol{b}, D)$ in $L_2(\Gamma_-, |\boldsymbol{b}\cdot\boldsymbol{n}|)$, endowed with the norm $\|g\|^2_{L_2(\Gamma_\pm,|\boldsymbol{b}\cdot\boldsymbol{n}|)} = \int_{\Gamma_\pm} |g|^2 |\boldsymbol{b}\cdot\boldsymbol{n}| ds$ so that

$$f(v) := (f_\circ, v) + \int_{\Gamma_-} g\gamma_-(v)|\boldsymbol{b}\cdot\boldsymbol{n}| ds \tag{2.14}$$

belongs to $(W_0(\boldsymbol{b}, D))'$ and the variational problem

$$b(u, v) = f(v), \quad v \in W_0(-\boldsymbol{b}, D) \tag{2.15}$$

possesses a unique solution in $L_2(D)$ which, when regular enough, coincides with the classical solution of (2.10), see [16, Theorem 2.2].

Moreover, since $X = L_2(D) = X'$, the quantity $\|v\|_Y := \|B^* v\|_{L_2(D)}$ is an equivalent norm on $W_0(-\boldsymbol{b}, D)$, see [16], and Remark 2.1 applies, i.e.,

$$\|B\|_{\mathscr{L}(L_2(D),(W_0(\boldsymbol{b},D))')} = \|B^*\|_{\mathscr{L}(W_0(\boldsymbol{b},D),L_2(D))} = 1, \tag{2.16}$$

see [16, Proposition 4.1]. One could also reverse the roles of test and trial space (with the inflow boundary conditions being then essential ones) but the present formulation imposes least regularity on the solution which will be essential in the next section. Note that whenever a PDE is written as a first order system, X can always be arranged as an L_2-space.

Our particular interest concerns the *parametric case*, i.e., the constant convection field $\boldsymbol{s}$ in

$$\begin{gathered} \boldsymbol{s}\cdot\nabla u(x,\boldsymbol{s}) + \kappa(x)u(x,\boldsymbol{s}) = f_\circ(x), \quad x \in D \subset \mathbb{R}^d,\ d = 2, 3, \\ u(x,\boldsymbol{s}) = g(x,\boldsymbol{s}),\ x \in \Gamma_-(\boldsymbol{s}), \end{gathered} \tag{2.17}$$

may *vary* over a set of directions $\mathscr{S}$ so that now the solution u depends also on the transport direction $\boldsymbol{s}$. In (2.17) and the following we assume that $\operatorname{ess\,inf}_{x\in D} \kappa(x) \geq 0$. Thus, for instance, when $\mathscr{S} = S^2$, the unit 2−sphere, u is considered as a function of five variables, namely $d = 3$ spatial variables and parameters from a two-dimensional set $\mathscr{S}$. This is the simplest example of a kinetic equation forming a core constituent in *radiative transfer* models. The in- and outflow boundaries now depend on $\boldsymbol{s}$:

$$\Gamma_\pm(\boldsymbol{s}) := \{x \in \partial D : \mp\boldsymbol{s}\cdot\mathbf{n}(x) < 0\}, \qquad \boldsymbol{s} \in \mathscr{S}\,. \tag{2.18}$$

Along similar lines one can determine u as a function of x and $\boldsymbol{s}$ in $X = L_2(D\times\mathscr{S})$ as the solution of a variational problem with test space $Y := \mathrm{clos}_{\|\cdot\|_{W(D\times\mathscr{S})}}\{v \in C(\mathscr{S}, C^1(D)) : v|_{\Gamma_\pm} \equiv 0\}$ with $\|v\|^2_{W(D\times\mathscr{S})} := \|v\|^2_{L_2(D\times\mathscr{S})} + \int_{\mathscr{S}\times D} |\boldsymbol{s}\cdot\nabla v|^2 dx d\boldsymbol{s}$. Again this formulation requires minimum regularity. Since later we shall discuss yet another formulation, imposing stronger regularity conditions, we refer to [16] for details.

2.3.3 δ-Proximality and Mixed Formulations

It is initially not clear how to exploit (2.9) numerically since the perfect inf-sup stability on the infinite dimensional level is *not* automatically inherited by a given pair $X_h \subset X, Y_h \subset Y$ of equal dimension. However, given $X_h \subset \hat{X}$, one can identify the "ideal" test space $Y(X_h) = R_Y^{-1}B(X_h)$ which may be termed ideal because

$$\sup_{w\in X_h}\sup_{v\in Y(X_h)} \frac{b(w,v)}{\|w\|_X\|v\|_Y} = \inf_{w\in X_h}\sup_{v\in Y(X_h)} \frac{b(v,w)}{\|w\|_X\|v\|_Y} = 1, \tag{2.19}$$

see [16]. In particular, this means that the solution $u_h \in X_h$ of the corresponding *Petrov-Galerkin* scheme

$$b(u_h, v) = f(v), \quad v \in Y(X_h), \tag{2.20}$$

realizes the best $\hat{X}$-approximation to the solution u of (2.5), i.e.,

$$\|u - u_h\|_{\hat{X}} = \inf_{w\in X_h} \|u - w\|_{\hat{X}}. \tag{2.21}$$

Of course, unless Y is an L_2 space, the ideal test space $Y(X_h)$ is, in general, not computable exactly. To retain stability it is natural to look for a numerically computable test space Y_h that is sufficiently close to $Y(X_h)$.

One can pursue several different strategies to obtain numerically feasible test spaces Y_h. When (2.5) is a discontinuous Galerkin formulation one can choose Y as a product space over the given partition, again with norms induced by the graph norm for the adjoint B^* so that the approximate inversion of the Riesz map R_Y can be *localized* [19, 20]. An alternative, suggested in [14, 16], is based on noting that by (2.8) the ideal Petrov Galerkin solution u_h from (2.20) is a *minimum residual* solution in Y', i.e., $u_h = \mathrm{argmin}_{w\in X_h}\|f - Bw\|_{Y'}$ whose normal equations read $(f - Bu_h, Bw)_{Y'} = 0$, $w \in X_h$. Since the inner product $(\cdot,\cdot)_{Y'}$ is numerically hard to access, one can write $(f - Bu_h, Bw)_{Y'} = \langle R_Y^{-1}(f - Bu_h), Bw\rangle$, where the dual pairing $\langle\cdot,\cdot\rangle$ is now induced by the standard L_2-inner product. Introducing as an auxiliary variable the "lifted residual"

$$y = R_Y^{-1}(f - Bu_h), \tag{2.22}$$

or equivalently $(R_Y y)(v) = \langle R_Y y, v\rangle = (y, v)_Y = \langle f - Bu_h, v\rangle$, $v \in Y$, one can show that (2.20) is equivalent to the saddle point problem

$$\begin{aligned} \langle R_Y y, v\rangle + b(u_h, v) &= \langle f, v\rangle, \ v \in Y, \\ b(w, y) &= 0, \quad w \in X_h, \end{aligned} \tag{2.23}$$

which involves only standard L_2-inner products, see [16, 18].

Remark 2.2. When working with $X, \hat{Y}$ instead of $\hat{X}, Y$, one has $R_Y = BR_X^{-1}B^*$ and hence, when $X = L_2(D)$ as in (2.11), one has $R_Y = BB^*$.

Since the test space Y is still infinite dimensional, a numerical realization would require finding a (possibly small) subspace $V \subset Y$ such that the analogous saddle point problem with Y replaced by V is still inf-sup stable. The relevant condition on V can be described by the notion of δ-*proximality* introduced in [16], see also [14]. We recall the formulation from [18]: $V \subset Y$ *is* δ-*proximal for* $X_h \subset \hat{X}$ *if*, for some $\delta \in (0, 1)$, with $P_{Y,V}$ denoting the Y-orthogonal projection from Y to V,

$$\|(I - P_{Y,V})R_Y^{-1}Bw\|_Y \le \delta \|R_Y^{-1}Bw\|_Y, \quad w \in X_h\ . \tag{2.24}$$

For a discussion of how to monitor or realize δ-proximality we refer to [14, 16], see also Sect. 2.3.5.

Theorem 2.2 ([14, 16, 18]). *Assume that for given $X_h \times V \subset X \times Y$ the test space V is δ-proximal for X_h, i.e. (2.24) is satisfied. Then, the solution $(u_{X_h,V}, y_{X_h,V}) \in X_h \times V$ of the saddle point problem*

$$\begin{aligned} \langle R_Y y_{X_h,V}, v\rangle + b(u_{X_h,V}, v) &= \langle f, v\rangle, \ v \in V, \\ b(w, y_{X_h,V}) &= 0, \quad w \in X_h, \end{aligned} \tag{2.25}$$

satisfies

$$\|u - u_{X_h,V}\|_{\hat{X}} \le \frac{1}{1-\delta} \inf_{w \in X_h} \|u - w\|_{\hat{X}}. \tag{2.26}$$

and

$$\|u - u_{X_h,V}\|_{\hat{X}} + \|y - y_{X_h,V}\|_Y \le \frac{2}{1-\delta} \inf_{w \in X_h} \|u - w\|_{\hat{X}}. \tag{2.27}$$

Moreover, one has

$$\inf_{w \in X_h} \sup_{v \in V} \frac{b(w, v)}{\|v\|_Y \|q\|_{\hat{X}}} \ge \sqrt{1 - \delta^2}. \tag{2.28}$$

Finally, (2.25) *is equivalent to the Petrov-Galerkin scheme*

$$b(u_{X_h,V}, v) = f(v), \quad v \in Y_h := P_{Y,V}(R_Y^{-1} B(X_h)) = P_{Y,V}(Y(X_h)). \tag{2.29}$$

The central message is that the Petrov-Galerkin scheme (2.29) can be realized *without* computing a basis for the test space Y_h, which for *each* basis function could require solving a problem of the size $\dim V$, by solving instead the saddle point problem (2.25). Moreover, the stability of both problems is governed by the δ-proximality of V. As a by-product, in view of (2.22), the solution component $y_{X_h,V}$ approximates the exact lifted residual $R_Y^{-1}(f - Bu_{X_h,V})$ and, as pointed out below, can be used for an a posteriori error control.

The problem (2.25), in turn, can be solved with the aid of an *Uzawa iteration* whose efficiency relies again on δ-proximality. For $k = 0, \ldots$, solve

$$\begin{aligned} \langle R_Y y^k, v\rangle &= \langle f - Bu^k, v\rangle, \quad v \in V, \\ (u^{k+1}, w)_{\hat{X}} &= (u^k, w)_{\hat{X}} + \langle B^* y^k, w\rangle, \quad w \in X_h. \end{aligned} \tag{2.30}$$

Thus, each iteration requires solving a symmetric positive definite Galerkin problem in V for the approximate lifted residual.

Theorem 2.3 ([16, Theorem 4.3]). *Assume that* (2.24) *is satisfied. Then the iterates generated by the scheme* (2.30) *converge to* $u_{X_h,V}$ *and*

$$\|u_{X_h,V} - u^{k+1}\|_{\hat{X}} \le \delta \|u_{X_h,V} - u^k\|_{\hat{X}}, \quad k = 0, 1, 2, \ldots. \tag{2.31}$$

2.3.4 Adaptive Petrov-Galerkin Solvers on Anisotropic Approximation Spaces

The benefit of the above saddle point formulation is not only that it saves us the explicit calculation of the test basis functions but that it provides also an *error estimator* based on the *lifted residual* $y_h = y_h(u_{X_h,V}, f)$ defined by the first row of (2.25).

2.3.4.1 Abstract δ-Proximinal Iteration

In fact, it is shown in [16] that when $V_h \subset Y$ is even δ-proximal for $X_h + B^{-1}F_h$, with some finite dimensional subspace $F_h \subset Y'$, one has

$$(1 - \delta)\|f_h - Bw\|_{Y'} \le \|y_h(w, f_h)\|_Y \le \|f_h - Bw\|_{Y'}, \quad w \in X_h, \tag{2.32}$$

where $f_h \in F_h$ is an approximation of $f \in Y'$. The space F_h controls which components of f are accounted for in the error estimator. The term $f - f_h$ is a *data oscillation* error as encountered in adaptive finite element methods. It follows that the current error of the Petrov-Galerkin approximation $u_{X_h,V}$ is controlled from below and above by the quantity $\|y_h\|_Y$. This can be used to formulate the *adaptive* Algorithm 1 that can be proven to give rise to a *fixed error reduction* per step. Its precise formulation can be found in [16, § 4.2]. It is shown in [16, Proposition 4.7] that each refinement step in Algorithm 1 below reduces the error by a fixed fraction. Hence it terminates after finitely many steps and outputs an approximate solution $\bar{u}$ satisfying $\|u - \bar{u}\|_{\hat{X}} \leq \varepsilon$.

Algorithm 1 Adaptive algorithm

1: Set target accuracy ε, initial guess $\bar{u} = 0$, initial error bound $e = \|f\|_{Y'}$, parameters $\rho, \eta, \alpha_1, \alpha_2 \in (0,1)$, initial trial and δ-proximal test spaces X_h, V_h;
2: **while** $e > \varepsilon$ **do** solve (2.25) within accuracy $\alpha_1 \rho$ (e.g. by an Uzawa iteration with initial guess $\bar{u}$) to obtain an approximate solution pair $(\hat{y}, \hat{u}) \in V_h \times X_h$;
3: enlarge X_h to $X_{h,+}$ in such a way that

$$\inf_{g \in X_{h,+}} \|B^* \hat{y} - g\|_{\hat{X}'} \leq \eta \|B^* \hat{y}\|_{\hat{X}'} \quad \text{and set} \quad r := \operatorname{argmin}_{g \in X_{h,+}} \|B^* \hat{y} - g\|_{\hat{X}'}; \tag{2.33}$$

4: compute $X_{h'} \supset X_h$, $F_{h'} \supset F_h$, $f_h \in BX_{h'} + F_{h'}$ such that $\|f - f_h\|_{Y'} \leq \alpha_2 \rho e$;
5: set $X_h + X_{h,+} + X_{h'} \to X_h$, $\rho e \to e$, and choose a δ-proximal subspace V_h for X_h;
6: set $\hat{u} + r_X \to \bar{u}$.
7: **end while**

2.3.4.2 Application to Transport Equations

We adhere to the setting described in Sect. 2.3.2, i.e., $X = \hat{X} = L_2(D)$, $\hat{Y} = Y = W_0(-\boldsymbol{b}, D)$, and $R_Y = BB^*$.

The trial spaces that we now denote by X_j to emphasize the nested construction below, are spanned by discontinuous piecewise linear functions on a mesh composed of cells from collections $\mathscr{C}_j$, i.e.,

$$X_j = \mathbb{P}_1(\mathscr{C}_j), \quad j \geq 0, \tag{2.34}$$

where the collections $\mathscr{C}_j$ are derived from collections $\tilde{\mathscr{C}}_j$ of the type (2.4) as described in Sect. 2.2.

Given X_j of the form (2.34), the test spaces V_j are defined by

$$V_j := \mathbb{P}_2(\mathscr{G}_j) \cap C(D) \quad \text{with} \quad \mathscr{G}_j := \{R^{iso}(Q) : Q \in \mathscr{C}_j\}, \tag{2.35}$$

where $R^{iso}(Q) = \{Q \cap P_i : i = 1, \ldots, 4\}$ is defined as follows. Let P be a parallelogram containing Q and sharing at least three vertices with Q. (There exist at most two such parallelograms and we choose one of them). Then the parallelograms P_i result from a dyadic refinement of P. As pointed out later, the

test spaces V_j constructed in this way, appear to be sufficiently large to ensure δ-proximality for X_j for δ significantly smaller than one uniformly with respect to j.

Since the test spaces V_j are determined by the trial spaces X_j the crucial step is to generate X_{j+1} by enlarging X_j based on an a posteriori criterion that "senses" directional information. This, in turn, is tantamount to a possibly anisotropic adaptive refinement of $\mathscr{C}_j$ leading to the updated spaces for the next iteration sweep of the form (2.30). The idea is to use a greedy strategy based on the largest "fluctuation coefficients". To describe this we denote for each $\iota_Q \in I_Q$ by $\Psi_{R_{\iota_Q}(Q)}$ an orthonormal Alpert-type wavelet basis for the difference space $\mathbb{P}_1(R_{\iota_Q}(Q)) \ominus \mathbb{P}_1(Q)$, see [1]. We then set

$$\Psi_j = \{\psi_\gamma \in \Psi_{R(Q)} : Q \in \mathscr{C}_{j-1}\}, \tag{2.36}$$

where $\Psi_{R(Q)} = \bigcup_{\iota_Q \in I_Q} \Psi_{R_{\iota_Q}(Q)}$. Initializing $\mathscr{C}_0$ as a uniform partition (on a low level), we define for some fixed $\theta \in (0,1)$

$$T_j = \theta \cdot \max_{\psi_\gamma \in \Psi_j} |\langle B^* r_j^K, \psi_\gamma \rangle|$$

for $j > 0$, where Ψ_j is the two level basis defined in (2.36) and $r_j^K = y^K$ is the lifted residual from the first row of the Uzawa iteration. Then, for each $Q \in \mathscr{C}_{j-1}$, we define its refinement $\tilde{R}(Q)$ (see the remarks following (2.4)) by

$$\tilde{R}(Q) := \begin{cases} \{Q\}, & \text{if} \quad \max_{\psi_\gamma \in \Psi_{R(Q)}} |\langle B^* r_j^K, \psi_\gamma \rangle| \le T_j, \\ R_{\hat{\iota}_Q}(Q), & \text{otherwise,} \end{cases}$$

where $\hat{\iota}_Q$ is chosen to maximize $\max_{\psi_\gamma \in \Psi_{R_{\iota_Q}(Q)}} |\langle B^* r_j^K, \psi_\gamma \rangle|$ among all $\iota_Q \in I_Q$. One can then check whether this enrichment yields a sufficiently accurate L_2-approximation of $B^* r_j^K$ (step 3 of Algorithm 1). In this case, we adopt $\mathscr{C}_j$. Otherwise, the procedure is repeated for a smaller threshold θ.

2.3.5 Numerical Results

We provide some numerical experiments to illustrate the performance of the previously introduced anisotropic adaptive scheme for first order linear transport equations and refer to [17] for further tests. We monitor δ-proximality by computing

$$\frac{\inf_{\phi \in V_j} \|u_j - u_j^K - B^* \phi\|_{L_2([0,1]^2)}}{\|u_j - u_j^K\|_{L_2([0,1]^2)}}, \tag{2.37}$$

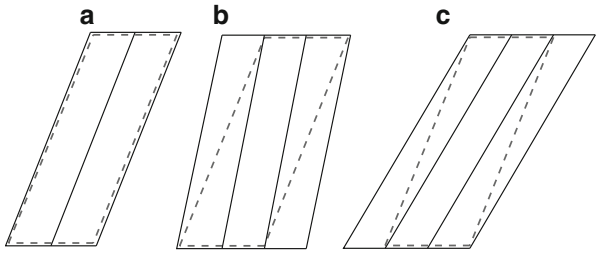


Fig. 2.2 Possible directional adjustments are illustrated for a parallelogram P (*dashed line*). (**a**) Rule (iii) of Sect. 2.2 yields two parallelograms with the same "direction". (**b**), (**c**) Applying rule (i) twice, changes the anisotropic direction slightly. The three refined parallelograms depicted in (**b**), (**c**) illustrate the results of a possible merging of adjacent triangles

where $u_j = \operatorname{argmin}_{v_j \in X_j} \|u - v_j\|_{L_2(D)}$. This is only a lower bound of the δ-proximality constant δ for one particular choice of w in (2.24) which coincides with the choice of w in the proof in [16]. In the following experiment, the number K of Uzawa iterations is for simplicity set to $K = 10$. One could as well employ an early termination of the inner iteration based on a posteriori control of the lifted residuals r_j^k.

We consider the transport equation (2.10) with zero boundary condition $g = 0$, convection field $\boldsymbol{b} = (x_2, 1)^T$, and right hand side $f = \chi_{\{x_1 > x_2^2/2\}} + 1/2 \cdot \chi_{\{x_1 \le x_2^2/2\}}$ so that the solution exhibits a discontinuity along the curvilinear shear layer given by $x_1 = \frac{1}{2}x_2^2$.

In this numerical example we actually explore ways of reducing the relatively large number of possible splits corresponding to the operators $R_{\iota_Q}, \iota_Q \in I_Q$, while still realizing the parabolic scaling law. In fact, we confined the cells to intersections of parallelograms P and their intersections with the domain D, much in the spirit of shearlet systems, employing anisotropic refinements as illustrated in Fig. 2.2 as well as the isotropic refinement R^{iso}. Permitting occasional overlaps of parallelograms, one can even avoid any interior triangles, apparently without degrading the accuracy of the adaptive approximation. The general refinement scheme described in Sect. 2.2 covers the presently proposed one as a special case, except, of course, for the possible overlap of cells.

Figure 2.3a, b show the adaptive grids associated with the trial space X_5 and the test space V_5. The refinement in the neighborhood of the discontinuity curve reflects a highly anisotropic structure. Figure 2.3c illustrates the approximation given by 306 basis elements. We emphasize that the solution is very smooth in the vicinity of the discontinuity curve and oscillations across the jump are almost completely absent and in fact much less pronounced than observed for isotropic discretizations. Figure 2.3d indicates the optimal rate realized by our scheme, see Theorem 2.1. The estimated values of the proximality parameter δ, displayed in Table 2.1, indicate the numerical stability of the scheme.

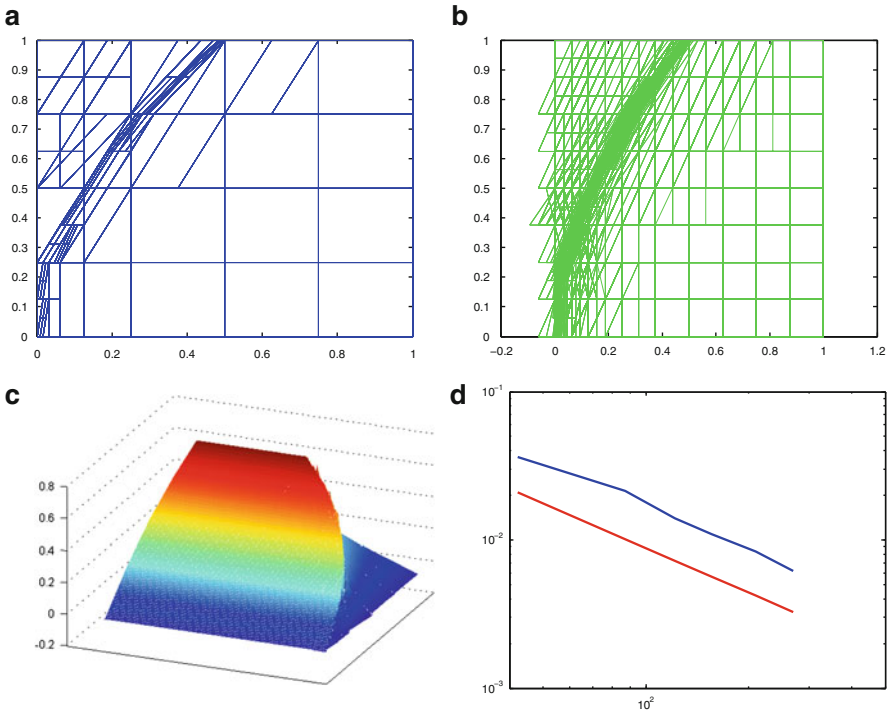


Fig. 2.3 (**a**) Adaptive grid for the trial space X_5. (**b**) Adaptive grid for the test space V_5. (**c**) Approximate solution (306 basis elements). (**d**) $L^2(D)$ errors (vertical axis) for N degrees of freedom (horizontal axis) achieved by the adaptive scheme (*blue*) in comparison with the optimal rate N^{-1} (*red*), predicted by Theorem 2.1. This is to be compared with the rate $N^{-1/2}$ realized by adaptive *isotropic* refinements [16]

Table 2.1 Numerical estimates (2.37) for the proximality constant δ and for the L_2 approximation error

n	Estimated δ	$\|u_j^K - u\|_{L_2([0,1]^2)}$
48	0.298138	0.036472
99	0.442948	0.021484
138	0.352767	0.013948
177	0.322156	0.010937
237	0.316545	0.008348
306	0.307965	0.006152

In the remainder of the paper we discuss parametric equations whose solutions are functions of spatial variables and additional parameters. Particular attention will here be paid to the *radiative transfer problems*, where the dimension of the physical domain is 2 or 3.

2.4 Reduced Basis Methods

2.4.1 Basic Concepts and Rate Optimality

Model reduction is often necessary when solutions to *parametric families* of PDEs are frequently queried for different parameter values e.g. in an online design or optimization process. The linear transport equation (2.17) is a simple example of such a *parameter dependent* PDE. Since (a) propagation of singularities is present and (b) the parameters determine the propagation direction $\boldsymbol{s}$ it turns out to already pose serious difficulties for standard model reduction techniques.

We emphasize that, rather than considering a *single* variational formulation for functions of spatial variables and parameters, as will be done later in Sect. 2.5, we take up the parametric nature of the problem by considering a *parametric family of variational formulations*. That is, for each fixed $\boldsymbol{s}$ the problem is an ordinary linear transport problem for which we can employ the corresponding variational formulation from Sect. 2.3.2, where now the respective spaces may depend on the parameters. In this section we summarize some of the results from [18] which are based in an essential way on the concepts discussed in the previous section.

In general, consider a family

$$b_\mu(u,v) = f(v), \quad u \in X_\mu,\, v \in Y_\mu,\ \mu \in \mathscr{P},\ b_\mu(u,v) = \sum_{k=1}^{M} \Theta_k(\mu) b_k(u,v) \tag{2.38}$$

of well-posed problems, where $\mathscr{P} \subset \mathbb{R}^P$ is a compact set of parameters μ, and the parameter dependence is assumed to be *affine* with smooth functions Θ_k. The solutions $u(\cdot;\mu) = u(\mu)$ then become functions of the spatial variables and of the parameters $\mu \in \mathscr{P}$.

As before we can view (2.38) as a parametric family of operator equations $B_\mu u = f$, where $B_\mu : X_\mu \to Y_\mu'$ is again given by $(B_\mu u)(v) = b_\mu(u,v)$. Each particular solution $u(\mu)$ is a *point* on the *solution manifold*

$$\mathscr{M} := \{B_\mu^{-1} f : \mu \in \mathscr{P}\}. \tag{2.39}$$

Rather than viewing $u(\mu)$ as a point in a very high-dimensional (in fact infinite dimensional) space, and calling a standard solver for each evaluation in a frequent query problem, the *Reduced Basis Method* (RBM) tries to exploit the fact that each $u(\mu)$ belongs to a much smaller dimensional manifold $\mathscr{M}$. Assuming that all the spaces X_μ are equivalent to a reference Hilbert space X with norm $\|\cdot\|_X$, the key objective of the RBM is to construct a possibly small dimensional linear space $X_n \subset X$ such that for a given target accuracy $\varepsilon > 0$

$$\sup_{\mu\in\mathscr{P}} \inf_{w\in X_n} \|u(\mu) - w\|_X := \operatorname{maxdist}_X(\mathscr{M}, X_n) \le \varepsilon. \tag{2.40}$$

Once X_n has been found, bounded linear functionals of the exact solution $u(\mu)$ can be approximated within accuracy ε by the functional applied to an approximation from X_n which, when n is small, can hopefully be determined at very low cost. The computational work in an RBM is therefore divided into an *offline* and an *online* stage. Finding X_n is the core offline task which is allowed to be computationally (very) expensive. More generally, solving problems in the "large" space X is part of the offline stage. Of course, solving a problem in X is already idealized. In practice X is replaced by a possibly very large trial space, typically a finite element space, which is referred to as the *truth* space and should be chosen large enough to guarantee the desired target accuracy, ideally certified by a posteriori bounds.

The computation of a (near-)best approximation $u_n(\mu) \in X_n$ is then to be *online feasible*. More precisely, one seeks to obtain a representation

$$u_n(\mu) = \sum_{j=1}^{n} c_j(\mu)\phi_j, \tag{2.41}$$

where the ϕ_j form a basis for X_n and where for each query $\mu \in \mathcal{P}$ the expansion coefficients $c_j(\mu)$ can be computed by solving only problems of the size n, see e.g. [39] for principles of practical realizations. Of course, such a concept pays off when the dimension $n = n(\varepsilon)$, needed to realize (2.40), grows very slowly when ε decreases. This means that the elements of $\mathcal{M}$ have sparse representations with respect to certain *problem dependent* dictionaries.

The by now most prominent strategy for constructing "good" spaces X_n can be sketched as follows. Evaluating for a given X_n the quantity $\mathrm{maxdist}_X(\mathcal{M}, X_n)$ is infeasible because this would require to determine for *each* $\mu \in \mathcal{P}$ (or for *each* μ in a large training set $\mathcal{P}_h \subset \mathcal{P}$ which for simplicity we also denote by $\mathcal{P}$) the solution $u(\mu)$ which even for the offline stage is way too expensive. Therefore, one chooses a *surrogate* $R_n(\mu)$ such that

$$\inf_{w \in X_n} \|u(\mu) - w\|_X \le R_n(\mu, X_n), \quad \mu \in \mathcal{P}, \tag{2.42}$$

where the evaluation of $R_n(\mu, X_n)$ is fast and an optimization of $R_n(\mu, X_n)$ can therefore be performed in the offline stage. This leads to the *greedy algorithm* in Algorithm 2. A natural question is to ask how the spaces X_n constructed in such a greedy fashion compare with "best spaces" in the sense of the *Kolmogorov n-widths*

$$d_n(\mathcal{M})_X := \inf_{\dim W_n = n} \sup_{w \in \mathcal{M}} \inf_{z \in W_n} \|w - z\|_X \,. \tag{2.44}$$

The n-widths are expected to decay the faster the more regular the dependence of $u(\mu)$ is on μ. In this case an RBM has a chance to perform well.

Clearly, one always has $d_n(\mathcal{M})_X \le \mathrm{maxdist}_X(\mathcal{M}, X_n)$. Unfortunately, the best constant C_n for which $\mathrm{maxdist}_X(\mathcal{M}, X_n) \le C_n d_n(\mathcal{M})_X$ is $C_n = 2^n$, see [4, 6]. Nevertheless, when comparing *rates* rather than individual values, one arrives at

Algorithm 2 Greedy algorithm

1: **function GA**
2: Set $X_0 := \{0\}$, $n = 0$,
3: **while** $\operatorname{argmax}_{\mu\in\mathscr{P}} R(\mu, X_n) \geq \varepsilon$ **do**
4:

$$\begin{aligned} \mu_{n+1} &:= \operatorname{argmax}_{\mu\in\mathscr{P}} R(\mu, X_n), \\ u_{n+1} &:= u(\mu_{n+1}), \\ X_{n+1} &:= \operatorname{span}\{X_n, \{u(\mu_{n+1})\}\} = \operatorname{span}\{u_1, \ldots, u_{n+1}\} \end{aligned} \tag{2.43}$$

5: **end while**
6: **end function**

more positive results [4, 21]. The following consequence of these results asserts optimal performance of the greedy algorithm provided that the surrogate sandwiches the error of best approximation.

Theorem 2.4 ([18, Theorem 1.3]). *Assume that there exists a constant* $0 < c_R \leq 1$ *such that one has for all n*

$$c_R R_n(\mu, X_n) \leq \inf_{w\in X_n} \|u(\mu) - w\|_X \leq R_n(\mu, X_n), \quad \mu \in \mathscr{P}. \tag{2.45}$$

Then, the spaces X_n *produced by Algorithm 2 satisfy*

$$d_n(\mathscr{M})_X \leq Cn^{-\alpha} \quad \Longrightarrow \quad \operatorname{maxdist}_X(\mathscr{M}, X_n) \leq \bar{C}n^{-\alpha}, \tag{2.46}$$

where $\bar{C}$ *depends only on* C, α, *and* $\kappa(R_n) := 1/c_R$, *the* condition *of the surrogate.*

We call the RBM *rate-optimal* whenever (2.46) holds for any $\alpha > 0$. Hence, finding rate-optimal RBMs amounts to finding *feasible well-conditioned surrogates*.

2.4.2 A Double Greedy Method

Feasible surrogates that do not require the explicit computation of truth solutions for each $\mu \in \mathscr{P}$ need to be based in one way or the other on *residuals*. When (2.38) is a family of uniformly X-elliptic problems so that B_μ are uniformly bounded isomorphisms from X onto X', residuals indeed lead to feasible surrogates whose condition depends on the ratio of the continuity and coercivity constant. This follows from the mapping property of B_μ, stability of the Galerkin method, and the best approximation property of the Galerkin projection, see [18].

When the problems (2.38) are indefinite or unsymmetric and singularly perturbed these mechanisms no longer work in this way, which explains why the conventional RBMs do not perform well for transport dominated problems in that they are far from rate-optimal.

As shown in [18], a remedy is offered by the above *renormation principle* providing well-conditioned variational formulations for (2.38). In principle, these allow one to relate errors (in a norm of choice) to residuals in a suitably adapted dual norm which are therefore candidates for surrogates. The problem is that, given a trial space X_n, in particular a space generated in the context of an RBM, it is not clear how to obtain a sufficiently good test space such that the corresponding Petrov-Galerkin projection is comparable to the best approximation. The new scheme developed in [18] is of the following form:

(I) Initialization: take $X_1 := \text{span}\{u(\mu_1)\}$, μ_1 randomly chosen, $Y_1 := \{0\}$;
(II) Given a pair of spaces $X_n, \tilde{V}_n$, the routine UPDATE-INF-SUP-δ enriches $\tilde{V}_n$ to a larger space V_n which is δ-proximal for X_n;
(III) Extend X_n to X_{n+1} by a greedy step according to Algorithm 2, set $\tilde{V}_{n+1} = V_n$, and go to (II) as long as a given target tolerance for an a posteriori threshold is not met.

The routine UPDATE-INF-SUP-δ works roughly as follows (see also [25] in the case of the Stokes system). First, we search for a parameter $\bar{\mu} \in \mathscr{P}$ and a function $\bar{w} \in X_n$ for which the inf-sup condition is worst, i.e.

$$\sup_{v\in\tilde{V}_n} \frac{b_{\bar{\mu}}(\bar{w}, v)}{\|v\|_{Y_{\bar{\mu}}}\|\bar{w}\|_{\hat{X}_{\bar{\mu}}}} = \inf_{\mu\in\mathscr{P}} \left(\inf_{w\in X_n} \sup_{v\in\tilde{V}_n} \frac{b_\mu(w, v)}{\|v\|_{Y_\mu}\|w\|_{\hat{X}_\mu}} \right). \tag{2.47}$$

If this worst case inf-sup constant does not exceed yet a desired uniform lower bound, $\tilde{V}_n$ does not contain an effective *supremizer*, i.e., a function realizing the supremum in (2.47), for $\bar{\mu}, \bar{w}$, yet. However, since the truth space satisfies a uniform inf-sup condition, due to the same variational formulation, there exists a good supremizer in the truth space which, is given by the Galerkin problem

$$\bar{v} = R^{-1}_{Y_{\bar{\mu}}} B_{\bar{\mu}} \bar{w} = \text{argmax}_{v\in Y_{\bar{\mu}}} \frac{b_{\bar{\mu}}(\bar{w}, v)}{\|v\|_{Y_{\bar{\mu}}}\|\bar{w}\|_{\hat{X}_{\bar{\mu}}}},$$

providing the enrichment $\tilde{V}_n \to \text{span}\{\tilde{V}_n, R^{-1}_{Y_\mu} B_\mu \bar{w}\}$.

The *interior greedy stabilization loop* (II) ensures that the input pair X_n, Y_n in step (III) is inf-sup stable with an inf-sup constant as close to one as one wishes, depending on the choice of $\delta < 1$. By Theorem 2.2, each solution $u_n(\mu)$ of the discretized system for $(X_h, V) = (X_n, V_n)$ satisfies the near-best approximation property (2.26) and (2.27). Hence $\|f - B_\mu u_n(\mu)\|_{Y'_\mu}$ is a well conditioned surrogate (with condition close to one). Therefore, the assumptions of Theorem 2.4 hold so that the *outer greedy* step (III) yields a rate-optimal update. In summary, under the precise assumptions detailed in [18], the above *double greedy scheme* is *rate-optimal*.

Before turning to numerical examples, a few comments on the interior greedy loop UPDATE-INF-SUP-δ are in order.

(a) Finding $\bar{\mu}$ in (2.47) requires for each μ-query to perform a singular value decomposition in the low dimensional reduced spaces so that this is offline feasible, see [18, Remark 4.2].
(b) When the test spaces Y_μ all agree with a reference Hilbert space Y as sets and with equivalent norms it is easy to see that the interior stabilization loop terminates after at most M steps where M is the number of parametric components in (2.38), see [18, Remark 4.9] and [25, 38]. If, on the other hand, the spaces Y_μ differ even as sets, as in the case of transport equations when the transport direction is the parameter, this is not clear beforehand. By showing that the inf-sup condition is equivalent to a δ-proximality condition one can show under mild assumptions though that the greedy interior loop still terminates after a number of steps which is independent of the truth dimension, [18, Remark 4.11].
(c) In this latter case the efficient evaluation of $\|f - B_\mu u(\mu)\|_{Y'_\mu}$ requires additional efforts, referred to as *iterative tightening*, see [18, Section 5.1].
(d) The renormation strategy saves an expensive computation of stability constants as in conventional RBMs since, by construction, through the choice of δ, the stability constants can be driven as close to one as one wishes.

The scheme has been applied in [18] to convection-diffusion and pure transport problems where the convection directions are parameter dependent. Hence the variational formulations are of the form (2.38). We briefly report some results for the transport problem since this is an extreme case in the following sense. The test spaces Y_μ do *not* agree as sets when one would like the X_μ to be equivalent for different parameters. Hence, one faces the obstructions mentioned in (b), (c) above. Moreover, for discontinuous right hand side and discontinuous boundary conditions the dependence of the solutions on the parameters has low regularity so that the n-widths do not decay as rapidly as in the convection-diffusion case. Nevertheless, the rate-optimality still shows a relatively fast convergence for the reduced spaces X_n shown below.

The first example concerns (2.17) (with $\mu = \boldsymbol{s}$ ranging over a quarter circle, $D = (0,1)^2$) for $f_\circ \equiv 1$, $g \equiv 0$. In the second example, we take $f_\circ(x_1, x_2) = 0.5$ for $x_1 < x_2$, $f_\circ(x_1, x_2) = 1$ for $x_1 \geq x_2$ (Tables 2.2, 2.3 and Fig. 2.4).

Table 2.2 Numerical results for Example 1, maximal truth error in L_2 0.000109832

Dimension				Maximal error between		
Trial	Test	δ	Maximal surr	rb truth	rb L_2	Surr/err
4	11	3.95e−01	8.44e−03	2.45e−02	2.45e−02	3.45e−01
10	33	4.32e−01	3.37e−03	5.74e−03	5.74e−03	5.87e−01
16	57	4.32e−01	1.50e−03	2.56e−03	2.56e−03	5.84e−01
20	74	4.16e−01	1.21e−03	2.10e−03	2.10e−03	5.77e−01
24	91	4.05e−01	7.27e−04	1.58e−03	1.58e−03	4.61e−01

Table 2.3 Numerical results for Example 2 after a single cycle of iterative tightening. Maximal truth error in L_2 0.0154814

Dimension				Maximal error between		
Trial	Test	δ	Maximal surr	rb truth	rb L_2	Surr/err
First reduced basis creation						
20	81	3.73e−01	2.71e−02	5.46e−02	5.62e−02	4.82e−01
Second reduced basis creation						
10	87	3.51e−01	6.45e−02	7.40e−02	7.53e−02	8.57e−01

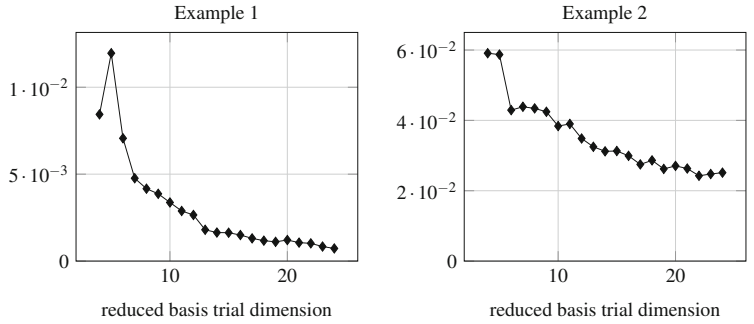


Fig. 2.4 Surrogates of the reduced basis approximation for Examples 1 and 2

2.5 Sparse Tensor Approximation for Radiative Transfer

We now extend the parametric transport problem (2.17) to the *radiative transport problem* (RTP) (see, e.g., [37]) which consists in finding the *radiative intensity* $u : D \times \mathscr{S} \to \mathbb{R}$, defined on the Cartesian product of a bounded physical domain $D \subset \mathbb{R}^d$, where $d = 2, 3$, and the unit $d_{\mathbb{S}}$-sphere as the parameter domain: $\mathscr{P} = \mathscr{S}$ with $d_{\mathbb{S}} = 1, 2$. Given an *absorption coefficient* $\kappa \geq 0$, a *scattering coefficient* $\sigma \geq 0$, and a *scattering kernel* or *scattering phase function* $\Phi > 0$, which is normalized to $\int_{\mathscr{S}} \Phi(\boldsymbol{s}, \boldsymbol{s}')\mathrm{d}\boldsymbol{s}' = 1$ for each direction $\boldsymbol{s}$, one defines the transport operator $\mathrm{T}u := (\boldsymbol{s} \cdot \nabla_x + \kappa)u$, and the scattering operator $\mathrm{Q}u := \sigma \mathrm{Q}_1 u = \sigma(u - \int_{\mathscr{S}} \Phi(\boldsymbol{s}, \boldsymbol{s}')u(\boldsymbol{x}, \boldsymbol{s}')\mathrm{d}\boldsymbol{s}')$. The radiative intensity is then given by

$$(\mathrm{T} + \mathrm{Q})u = f, \quad u|_{\partial\Omega_-} = g, \tag{2.48}$$

where $f := \kappa I_b$, $\partial\Omega_- := \{(\boldsymbol{x}, \boldsymbol{s}) \in \partial D \times \mathscr{S} : \boldsymbol{s} \cdot \boldsymbol{n}(\boldsymbol{x}) < 0\}$, and g denote the source term, the inflow-boundary, and the inflow-boundary values, respectively. As before, $\Gamma_-(\boldsymbol{s}) := \{\boldsymbol{x} \in \partial D : \boldsymbol{s} \cdot \boldsymbol{n}(\boldsymbol{x}) < 0\}$ (see (2.18)) stands for the physical inflow-boundary.

The partial differential equation (2.48) is known as *stationary monochromatic radiative transfer equation* (RTE) with scattering, and can be viewed as (nonlocal) extension of the parametric transport problem (2.17), where the major difference

to (2.17) is the scattering operator Q. Sources with support contained in D are modeled by the *blackbody intensity* $I_b \geq 0$, radiation from sources outside of the domain or from its enclosings is prescribed by the *boundary data* $g \geq 0$.

Deterministic numerical methods for the RTP which are commonly used in engineering comprise the *discrete ordinates (S_N-) method* and the *spherical harmonics (P_N-) method*.

In the *discrete ordinate method* (DOM), the angular domain is collocated by a finite number of fixed propagation directions in the angular parameter space; in this respect, the DOM resembles the greedy collocation in the parameter domain: each of the directions Eq. (2.48) results in a spatial PDE which is solved (possibly in parallel) by standard finite differences, finite elements, or finite volume methods.

In the *spherical harmonics method* (SHM), a spectral expansion with spatially variable coefficients is inserted as ansatz into the variational principle Eq. (2.48). By orthogonality relations, a coupled system of PDEs (whose type can change from hyperbolic to elliptic in the so-called diffuse radiation approximation) for the spatial coefficients is obtained, which is again solved by finite differences or finite elements.

The common deterministic methods S_N- and P_N-approximation exhibit the so-called "curse of dimensionality": the error with respect to the total numbers of degrees of freedom (DoF) M_D and $M_{\mathscr{S}}$ on the physical domain D and the parameter domain $\mathscr{S}$ scales with the dimension d and $d_{\mathbb{S}}$ as $O(M_D^{-s/d} + M_{\mathscr{S}}^{-t/d_{\mathbb{S}}})$ with positive constants s and t.

The so called *sparse grid approximation method* alleviates this curse of dimensionality for elliptic PDEs on cartesian product domains, see [7] and the references therein. Widmer et al. [41] has developed a sparse tensor method to overcome the curse of dimensionality for radiative transfer with a wavelet (isotropic) discretization of the angular domain. Under certain regularity assumptions on the absorption coefficient κ and the blackbody intensity I_b, their method achieves the typical benefits of sparse tensorization: a log-linear complexity in the number of degrees of freedom of a component domain with an essentially (up to a logarithmic factor) undeteriorated rate of convergence. However, scattering had not been addressed in that work.

In order to include scattering and to show that the concepts of sparse tensorization can also be applied to common solution methods, sparse tensor versions of the spherical harmonics approximation were developed extending the "direct sparse" approach by [41]. The presently developed version also accounts for scattering [27]. For this sparse spherical harmonics method, we proved that the benefits of sparse tensorization can indeed be harnessed.

As a second method a *sparse tensor product version of the DOM* based on the *sparse grid combination technique* was realized and analyzed in [26, 28]. Solutions to discretizations of varying discretization levels, for a number of collocated transport problems, and *with scattering discretized by combined Galerkin plus quadrature approximation in the transport collocation directions* are combined in this method to form a sparse tensor solution that we proved in [26, 28] breaks the curse of dimensionality as described above. These benefits hold as long as the exact solution of the RTE is sufficiently regular. An overview follows.

2.5.1 Sparse Discrete Ordinates Method (Sparse DOM)

We adopt a formulation where the inflow boundary conditions are enforced in a weak sense. To this end, we define the boundary form (see, e.g., [26])

$$\partial b(u,v) := (v, \boldsymbol{s}\cdot\boldsymbol{n}u)_{L^2(\partial\Omega_-)} = \int_{\mathscr{S}}\int_{\Gamma_-(\boldsymbol{s})} \boldsymbol{s}\cdot\boldsymbol{n}uv\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{s}\ . \tag{2.49}$$

Writing for $v : D\times\mathscr{S}\to\mathbb{R}$ briefly $\|v\| := \|v\|_{L_2(D\times\mathscr{S})}$, the norms

$$\|v\|_-^2 := -\partial b(v,v), \quad \|v\|_1{}^2 := \|v\|^2 + \|\boldsymbol{s}\cdot\nabla_x v\|^2 + \|\mathrm{Q}_1 v\|^2 + \|v\|_-^2$$

define the Hilbert space $\mathscr{V}_1 := \{v\in L_2(D\times\mathscr{S}) : \|v\|_1<\infty\}$. The SUPG-stabilized Galerkin variational formulation reads: find $u\in\mathscr{V}_1$ such that

$$(\mathrm{R}v, (\mathrm{T}+\mathrm{Q})u)_{L^2(D\times\mathscr{S})} - 2\partial b(u,v) = (\mathrm{R}v, f)_{L_2(D\times\mathscr{S})} - 2\partial b(g,v) \quad \forall v\in\mathscr{V}_1 \tag{2.50}$$

with SUPG stabilization $\mathrm{R}v := v + \eta\boldsymbol{s}\cdot\nabla_x v$, where $\eta\approx 2^{-L}$.

For the discretization of (2.50), we replace $\mathscr{V}_1$ by $V^{L,N} = V_D^L\otimes V_{\mathscr{S}}^N$. In the physical domain, standard P_1-FEM with a one-scale basis on a uniform mesh of width $h_{\max}\lesssim 2^{-L}$ is used, in the angular domain, piecewise constants on a quasiuniform mesh of width $h_{\max}\lesssim N^{-1}$. Fully discrete problems are obtained with a one-point quadrature in the angular domain. The resulting Galerkin formulation (2.50) can be shown to result in the same linear system of equations as the standard collocation discretization [26, Sec. 5.2]. The solution is constructed with the sparse grid combination technique (see [7]):

$$\hat{u}_{L,N} = \sum_{\ell_D=0}^{L}\left(u_{\ell_D,\ell_{\mathscr{S}}^{\max}(\ell_D)} - u_{\ell_D,\ell_{\mathscr{S}}^{\max}(\ell_D+1)}\right),$$

where $u_{\ell_D,\ell_{\mathscr{S}}}\in V^{\ell_D,\ell_{\mathscr{S}}}$ denotes the solution to a full tensor subproblem of physical resolution level ℓ_D and angular resolution level $\ell_{\mathscr{S}}$. The maximum angular index $\ell_{\mathscr{S}}^{\max} = 2^{\lfloor\log_2(N+1)\rfloor/L(L-l_D)}$ ensures that the angular resolution decreases when the physical resolution increases and vice versa.

While the full tensor solution $u_{L,N}$ requires $O(2^{dL}N^{d_{\mathbb{S}}})$ degrees of freedom, the sparse solution involves asymptotically at most $O((L+\log N)(2^{dL}+N^{d_{\mathbb{S}}}))$ degrees of freedom [26, Lemma 5.6]. At the same time, we have

$$\|u-u_{L,N}\|_1 \le C2^{-L}\|u\|_{H^{2,0}(D\times\mathscr{S})} + N^{-1}\|u\|_{H^{1,1}(D\times\mathscr{S})},$$

while for solutions in $H^{2,1}(D\times\mathscr{S})\subset(H^{2,0}(D\times\mathscr{S})\cap H^{1,1}(D\times\mathscr{S}))$

$$\|u - \hat{u}_{L,N}\|_1 \leq CL \max\{2^{-L}, N^{-1}\} \|u\|_{H^{2,1}(D\times\mathscr{S})} ,$$

where C is an absolute constant.

2.5.2 Numerical Experiment

To evaluate the solution numerically we monitor the incident radiation $G(\boldsymbol{x}) = \int_{\mathscr{S}} u(\boldsymbol{x}, \boldsymbol{s})\mathrm{d}\boldsymbol{s}$ and its relative error $err(G_{L,N})_X = \|G - G_{L,N}\|_X / \|G\|_X$, $X = L_2(D), H^1(D)$.

The setting for the experiment is $D = (0,1)^d$, $\mathscr{S} = \mathscr{S}^d_{\mathbb{S}}$. We solve the RTP (2.48) with isotropic scattering $\Phi(\boldsymbol{s}, \boldsymbol{s}') = 1/|\mathscr{S}|$ and with zero inflow boundary conditions $g = 0$. A blackbody radiation $I_b(\boldsymbol{x}, \boldsymbol{s})$ corresponding to the exact solution

$$u(\boldsymbol{x}, \boldsymbol{s}) = \frac{3}{16\pi}(1 + (\boldsymbol{s} \cdot \boldsymbol{s}')^2) \prod_{i=1}^{3} (-4x_i(x_i - 1)),$$

with fixed $\boldsymbol{s}' = (1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^\top$ is inserted in the right hand side functional in (2.50). The absorption coefficient is set to $\kappa = 1$, the scattering coefficient to $\sigma = 0.5$.

This $3+2$-dimensional problem was solved with a parallel C++ solver designed for the sparse tensor solution of large-scale radiative transfer problems. Figure 2.5 shows the superior efficiency of the sparse approach with respect to number of degrees of freedom vs. achieved error. The convergence rates indicate that the curse

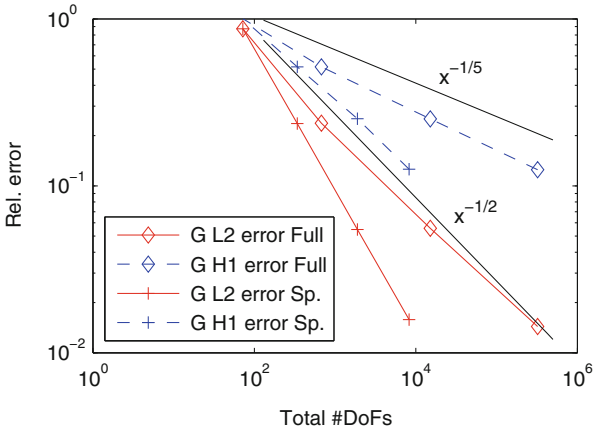


Fig. 2.5 Convergence in incident radiation with full and sparse DOM. Resolution for reference solution was $L_{\text{ref}} = 4$. Reference slopes provided as visual aids only

of dimensionality is mitigated by the sparse DOM. Further gains are expected once the present, nonadaptive sparse DOM is replaced by the greedy versions outlined in Sect. 2.3.1.1.

References

1. Alpert, B.K.: A class of bases in L^2 for the sparse representation of integral operators. SIAM J. Math. Anal. **24**, 246–262 (1993)
2. Bachmayer, M., Dahmen, W.: Adaptive near-optimal rank tensor approximation for high-dimensional operator equations. Found. Comput. Math. March, 2014, doi: 10.1007/s10208-013-9187-3, http://arxiv.org/submit/851475
3. Barrett, J.W., Morton, K.W.: Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems. Comput. Methods Appl. M. **45**, 97–12 (1984)
4. Binev, P., Cohen, A., Dahmen, W., DeVore, R., Petrova, G., Wojtaszczyk, P.: Convergence rates for greedy algorithms in reduced basis methods. SIAM J. Math. Anal. **43**, 1457–1472 (2011)
5. Bochev, P.B., Gunzburger, M.D.: Least-Squares Finite Element Methods. Applied Mathematical Sciences. Springer, New York (2009)
6. Buffa, A., Maday, Y., Patera, A.T., Prud'homme, C., Turinici, G.: A priori convergence of the greedy algorithm for the parameterized reduced basis. ESAIM: Math. Model. Numer. Anal. **46**, 595–603 (2012)
7. Bungartz, H.-J., Griebel, M.: Sparse grids. In: Iserles, A. (ed.) Acta Numerica, vol. 13, pp. 147–269. Cambridge University Press, Cambridge (UK) (2004)
8. Cai, Z., Manteuffel, T., McCormick, S., Ruge, J.: First-order system $\mathscr{L}\mathscr{L}^*$ $(FOSLL)^*$: scalar elliptic partial differential equtions. SIAM J. Numer. Anal. **39**, 1418–1445 (2001)
9. Candès, E.J., Donoho, D.L.: New tight frames of curvelets and optimal representations of objects with piecewise-C^2 singularities. Commun. Pure Appl. Math. **57**, 219–266 (2002)
10. Chen, L., Sun, P., Xu, J.: Optimal anisotropic simplicial meshes for minimizing interpolation errors in L^p norm. Math. Comput. **76**, 179–204 (2007)
11. Chkifa, A., Cohen, A., DeVore, R., Schwab, C.: Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs. ESAIM: Math. Model. Numer. Anal. **47**(1), 253–280 (2013). http://dx.doi.org/10.1051/m2an/2012027
12. Cohen, A., Mirebeau, J.-M.: Greedy bisection generates optimally adapted triangulations. Math. Comput. **81**, 811–837 (2012)
13. Cohen, A., DeVore, R., Schwab, C.: Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. Found. Comput. Math. **10**, 615–646 (2010)
14. Cohen, A., Dahmen, W., Welper, G.: Adaptivity and variational stabilization for convection-diffusion equations. ESAIM: Math. Model. Numer. Anal. **46**, 1247–1273 (2012)
15. Cohen, A., Dyn, N., Hecht, F., Mirebeau, J.-M.: Adaptive multiresolution analysis based on anisotropic triangulations. Math. Comput. **81**, 789–810 (2012)
16. Dahmen, W., Huang, C., Schwab, C., Welper, G.: Adaptive Petrov-Galerkin methods for first order transport equations. SIAM J. Numer. Anal. **50**, 2420–2445 (2012)
17. Dahmen, W., Kutyniok, G., Lim, W.-Q, Schwab, C., Welper, G.: Adaptive anisotropic discretizations for transport equations, preprint, 2014
18. Dahmen, W., Plesken, C., Welper, G.: Double greedy algorithms: reduced basis methods for transport dominated problems. ESAIM: Math. Model. Numer. Anal. **48**, 623–663 (2014). doi:10.1051/m2an/2013103, http://arxiv.org/abs/1302.5072
19. Demkowicz, L.F., Gopalakrishnan, J.: A class of discontinuous Petrov-Galerkin methods I: the transport equation. Comput. Methods Appl. Mech. Eng. **199**, 1558–1572 (2010)
20. Demkowicz, L., Gopalakrishnan, J.: A class of discontinuous Petrov-Galerkin methods. Part II: optimal test functions. Numer. Methods Part. Differ. Equ. **27**, 70–105 (2011)

21. DeVore, R., Petrova, G., Wojtaszczyk, P.: Greedy algorithms for reduced bases in Banach spaces. Constr. Approx. **37**, 455–466 (2013)
22. Dolejsi, V.: Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes. Comput. Vis. Sci. **1**, 165–178 (1998)
23. Donoho, D.L.: Sparse components of images and optimal atomic decompositions. Constr. Approx. **17**, 353–382 (2001)
24. Donoho, D.L., Kutyniok, G.: Microlocal analysis of the geometric separation problem. Commun. Pure Appl. Math. **66**, 1–47 (2013)
25. Gerner, A., Veroy-Grepl, K.: Certified reduced basis methods for parametrized saddle point problems. SIAM J. Sci. Comput. **35**, 2812–2836 (2012)
26. Grella, K.: Sparse tensor approximation for radiative transport. PhD thesis 21388, ETH Zurich (2013)
27. Grella, K., Schwab, C.: Sparse tensor spherical harmonics approximation in radiative transfer. J. Comput. Phys. **230**, 8452–8473 (2011)
28. Grella, K., Schwab, C.: Sparse discrete ordinates method in radiative transfer. Comput. Methods Appl. Math. **11**, 305–326 (2011)
29. King, E.J., Kutyniok, G., Zhuang, X.: Analysis of inpainting via clustered sparsity and microlocal analysis. J. Math. Imaging Vis. **48**, 205–234 (2014)
30. Kittipoom, P., Kutyniok, G., Lim, W.-Q: Construction of compactly supported shearlet frames. Constr. Approx. **35**, 21–72 (2012)
31. Kutyniok, G., Labate, D.: Resolution of the wavefront set using continuous shearlets. Trans. Am. Math. Soc. **361**, 2719–2754 (2009)
32. Kutyniok, G., Labate, D.: Shearlets: Multiscale Analysis for Multivariate Data. Birkhäuser, Boston (2012)
33. Kutyniok, G., Lim, W.-Q: Compactly supported shearlets are optimally sparse. J. Approx. Theory **163**, 1564–1589 (2011)
34. Kutyniok, G., Lemvig, J., Lim, W.-Q: Optimally sparse approximations of 3D functions by compactly supported shearlet frames. SIAM J. Math. Anal. **44**, 2962–3017 (2012)
35. Lim, W.-Q: The discrete shearlet transform: a new directional transform and compactly supported shearlet frames. IEEE Trans. Image Proc. **19**, 1166–1180 (2010)
36. Mirebeau, J.-M.: Adaptive and anisotropic finite element approximation: theory and algorithms, PhD thesis, Université Pierre et Marie Curie – Paris VI (2011). http://tel.archives-ouvertes.fr/tel-00544243
37. Modest, M.F.: Radiative Heat Transfer, 2nd edn. Elsevier, Amsterdam (2003)
38. Rozza, G., Veroy, K.: On the stability of reduced basis techniques for Stokes equations in parametrized domains. Comput. Methods Appl. Mech. **196**, 1244–1260 (2007)
39. Sen, S., Veroy, K., Huynh, D.B.P., Deparis, S., Nguyn, N.C., Patera, A.T.: "Natural norm" a-posteriori error estimators for reduced basis approximations. J. Comput. Phys. **217**, 37–62 (2006)
40. Welper, G.: Infinite dimensional stabilization of convection-dominated problems. PhD thesis, RWTH Aachen (2013). http://darwin.bth.rwth-aachen.de/opus3/volltexte/2013/4535/
41. Widmer, G., Hiptmair, R., Schwab, C.: Sparse adaptive finite elements for radiative transfer. J. Comput. Phys. **227**, 6071–6105 (2008)

Chapter 3
Regularity of the Parameter-to-State Map of a Parabolic Partial Differential Equation

Rudolf Ressel, Patrick Dülk, Stephan Dahlke, Kamil S. Kazimierski, and Peter Maass

Abstract In this paper, we present results that have been obtained in the DFG-SPP project "Adaptive Wavelet Frame Methods for Operator Equations: Sparse Grids, Vector-Valued Spaces and Applications to Nonlinear Inverse Problems". This project has been concerned with (nonlinear) elliptic and parabolic operator equations on nontrivial domains as well as with related inverse parameter identification problems. In this paper we study analytic properties of the underlying parameter-to-state map, which is motivated by a parabolic model for the embryonal development of drosophila melanogaster.

3.1 Introduction

The DFG-SPP project "Adaptive Wavelet Frame Methods for Operator Equations" has been concerned with (nonlinear) elliptic and parabolic operator equations on nontrivial domains as well as with related inverse parameter identification problems. In this paper we study analytic properties of the underlying parameter-to-state map. The complementary results on the development of optimally convergent adaptive wavelet schemes obtained in this project are presented in [5].

R. Ressel (✉)
DLR Oberpfaffenhofen, EOC, Münchner Str. 20, 82234 Wessling, Germany
e-mail: rudolf.ressel@gmx.de

P. Dülk • P. Maass
University of Bremen, Zentrum für Technomathematik, Bibliothekstr. 1, 28359 Bremen, Germany
e-mail: pduelk@math.uni-bremen.de; pmaass@math.uni-bremen.de

S. Dahlke
Philipps-University of Marburg, Hans Meerwein Str., Lahnberge, 35032 Marburg, Germany
e-mail: dahlke@mathematik.uni-marburg.de

K.S. Kazimierski
University of Graz, Heinrichstr. 36, 8010 Graz, Austria
e-mail: kazimier@uni-graz.at

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_3

The starting point of the present paper is the common Drosophila melanogaster organism at a particular embryonal stage, when the entire metabolism still takes place in one large (multinuclear) cell. More precisely, we are concerned with a presumed governing differential equation for the model system, where the gene-expression concentrations (state) appear as the solutions of a model equation and the parameters model the regulating interaction (based on models proposed by [12, 13]). To derive a differential equation resulting from these assumptions, we take some connected Lipschitz-bounded domain $U \subset \mathbb{R}^n$, $n = 2, 3$ as the physical domain of the metabolism. We denote by the scalar functions u_i, $i = 1, \ldots, d$ the gene product concentrations on U over the time interval $[0, T]$. The partial differential equation governing the biochemical evolution under consideration reads:

$$\begin{aligned} \frac{\partial u_i}{\partial t} - \nabla \cdot (D_i \nabla u_i) + \lambda_i \cdot u_i &= R_i \Phi\Big(\sum_{j=1}^{d} W_{ij} u_j\Big) && \text{in } U_T = U \times (0, T], \\ \frac{\partial u_i}{\partial \nu} &= 0 && \text{on } \partial U \times [0, T], \\ u(0) &= u_0 && \text{on } U \times \{0\}, \end{aligned} \tag{3.1}$$

where $i, j = 1, \ldots, d$ and $u_i, D_i, \lambda_i, R_i, W_{ij}$ are functions of space and time and the function $\Phi_i : \mathbb{R} \to \mathbb{R}$ is given via $\Phi(y) = 0.5\left(y/\sqrt{y^2 + 1} + 1\right)$.

The non-negative parameter functions D, λ, R are measurable, almost everywhere pointwise bounded functions in time and space. The terms in the PDE correspond to diffusion, chemical consumption, and synthesis, respectively. The function Φ is some smooth sigmoidal signal response function (cf. [12]). The matrix function W is likewise bounded, but may attain negative values. In fact, in the matrix-vector product (Wu) negative entries in W correspond to an inhibiting influence of one expressed gene on the synthesis of the other gene and positive ones represent an amplifying effect. The task is then to identify the parameters from measured concentrations of the $(u_i)_{i=1,\ldots,d}$ at certain times of the observation period, most importantly the genetic interaction encoded in W.

A well-known technique to solve the nonlinear inverse problem that arises from Eq. (3.1) is Tikhonov regularization with sparsity constraints [4, 8, 11]. This paper aims at laying down a functional analytic framework for the parameter-to-state operator, i.e. the operator that maps a parameter tuplet to the respective solution of the differential equation, which allows us to apply Tikhonov regularization. Especially, it is shown, that the parameter-to-state map has a Lipschitz continuous derivative. This is the key ingredient to apply numerical minimization schemes for Tikhonov-type functionals [4].

3.2 Function Spaces

Usually, one reformulates the model equation as a Cauchy problem with solutions in $L_2([0,T], H^1(U))$ and parameters in L_∞ (cf. [17]). We choose a slightly more general approach by utilizing recent regularity results of [9]. The benefit lies in obtaining greater leeway for the topology of the parameters (cf. [10, 15]).

3.2.1 The Parameter Space

We start with pointwise a.e. (almost everywhere) bounded sets of measurable functions, i.e. bounded subsets of L_∞-spaces (bounds are indicated in Eq. (3.2) right below). The topology we then work with is induced by L_p-norms. The global pointwise bounds are denoted by

$$0 < C_{\mathscr{P},1} \le D,\ \lambda \le C_{\mathscr{P},2}, \quad 0 \le R \le C_{\mathscr{P},2}, \quad \|W\|_\infty \le C_{\mathscr{P},2}. \tag{3.2}$$

The parameter space for D is defined as

$$\mathscr{P}_D = \{D \in L_{p_D}(U \times [0,T], \mathbb{R}^d) : 0 < C_{\mathscr{P},1} \le D \le C_{\mathscr{P},2}\}.$$

Accordingly, the spaces $\mathscr{P}_\lambda, \mathscr{P}_R, \mathscr{P}_W$ for λ, R, W are defined with the exponents of integrability $2 \le p_\lambda, p_R, p_W \le \infty$. These parameters will be specified later to meet further analytical goals, e.g., the differentiability of differential operators appearing in the PDE. We denote the space of admissible parameter tuplets by $\mathscr{P} = \mathscr{P}_D \times \mathscr{P}_\lambda \times \mathscr{P}_R \times \mathscr{P}_W$ and endow this space with the norm $\|(D,\lambda,R,W)\| := \|D\|_{L_{p_D}(U_T,\mathbb{R}^d)} + \|\lambda\|_{L_{p_D}(U_T,\mathbb{R}^d)}\| + \|R\|_{L_{p_R}(U_T,\mathbb{R}^d)} + \|W\|_{L_{p_W}(U_T,\mathbb{R}^{d\times d})}$. Note, that in case one of the exponents of integrability is smaller than ∞, $\mathscr{P}$ does not include all the nearby L_p (or L_∞) elements but only those that satisfy the stated almost everywhere bounds. Therefore all the results presented later on should be understood with respect to the relative topology.

3.2.2 Solution Spaces

To establish the solvability theory for our model PDE (3.1) we need to choose spaces V_q and Y_q, such that these spaces "fit" the elliptic part $\nabla \cdot (D_i \nabla g_i)$ as well as the nonlinear part $R_i \Phi(W g_i)$ of the differential operator. As a reference for Sobolev and Besov spaces we refer to [6, Ch.3], [16, 2.1.2] or [18].

With $dim(U) = n$ and $1/q + 1/q' = 1$ we define:

$$V_q = H^1_q(U, \mathbb{R}^d) \quad \text{and} \quad Y_q = (H^1_{q'}(U, \mathbb{R}^d))', \qquad q \in (n, n+\varepsilon).$$

This gives us a Gelfand triple $I_{triple} : V_q \hookrightarrow L_q \hookrightarrow Y_q$. For notational convenience, we will not distinguish between elements $u \in V_q$ and their embeddings in Y_q.

We now define a suitable space for the solutions of (3.1):

$$\mathscr{W}_{s,q} = \{u \in L_s([0,T]; V_q) : u' \in L_s([0,T]; Y_q)\}$$

with norm $\|u\| = \|u\|_{L_s([0,T];V_q)} + \|u'\|_{L_s([0,T];Y_q)}$, $s \geq 2$, where u' denotes the distributional derivative with values in Y_q.

We choose $q \in (n, n+\varepsilon)$, where ε depends on results of maximal parabolic regularity, see Sect. 3.3.3. For simplicity we will always set $s = q$. This avoids a second set of exponents, however, all subsequent results remain correct for a general $s \geq 2$. For notational convenience, we define $\mathscr{V}_q = L_q([0,T], V_q)$. Using $s = q$ and omitting the subscript s from now on, we introduce the shorter notation

$$\mathscr{W} = \{u \in \mathscr{V}_q : u' \in (\mathscr{V}_{q'})'\} \quad \text{with norm } \|u\|_{\mathscr{W}} = \|u\|_{\mathscr{V}_q} + \|u'\|_{(\mathscr{V}_{q'})'}.$$

We will need the following embeddings, which are derived from well-known facts about solution spaces of vector-valued parabolic equations, see e.g. [14].

Theorem 3.1. *$\mathscr{W}$ is a Banach space with continuous embeddings $\mathscr{W} \hookrightarrow \mathscr{C}([0,T], G) \hookrightarrow \mathscr{C}([0,T], \mathscr{C}(\overline{U}, \mathbb{R}^d))$, where $G = B^m_{q,q}(U, \mathbb{R}^d)$, $m = 1 - 2/q$.*

Theorem 3.2. *The embedding $I_M : \mathscr{W} \hookrightarrow L_r([0,T]; L_r(U, \mathbb{R}^d))$ is continuous for any $r \in (q, \infty]$.*

From now on we choose $r > q$.

3.3 The Model PDE as an Evolution Equation

In this section we will analyze the differential operators defining (3.1) and derive an existence result for solutions of (3.1).

3.3.1 The Linear Differential Operator

Now we restrict the indices of the function spaces and choose $q \in (n, n+\varepsilon)$, $p_\lambda, p_R, p_W > 4q$, $p_D = \infty$ and $r \in (2q, \infty]$. Under these restraints we find

$$\frac{1}{p_R} + \frac{1}{p_W} + \frac{1}{r} < \frac{1}{q}, \qquad \frac{1}{p_\lambda} + \frac{1}{r} \leq \frac{1}{q}\,. \tag{3.3}$$

These indices have been chosen in order to be able to apply Hölder estimates (cf. the concept of multiplicator spaces in [1, pp.90]) in the subsequent sections. We now

define the differential operator $\frac{d}{dt} + \mathscr{A} : \mathscr{P} \times \mathscr{W} \to (\mathscr{V}_{q'})'$ by $\left(\frac{d}{dt} + \mathscr{A}\right)(\pi, u) = F$, where F acts like

$$F(v\rho) = -\int_0^T \Big(\int_U uv dx\Big)\rho' dt + \int_0^T \Big(\int_U D\nabla u \nabla v dx\Big)\rho dt + \int_0^T \Big(\int_U \lambda\, u\, v\, dx\Big)\rho dt,$$

with $\rho \in \mathscr{C}_0^\infty([0,T],\mathbb{R})$, $v \in V_{q'}$, $\pi = (D, \lambda, R, W) \in \mathscr{P}$. Note, that ρ captures the distributional time derivative of the test functions.

Theorem 3.3. *Let $\mathscr{W}$ and $\mathscr{P}$ be defined as above. Then $\frac{d}{dt} + \mathscr{A} : \mathscr{P} \times \mathscr{W} \to (\mathscr{V}_{q'})'$ is linear and continuous in each argument. Moreover $\frac{d}{dt} + \mathscr{A}$ is continuously differentiable in (π, u).*

Proof. The linearity of the operator is obvious. The continuity of $\frac{d}{dt}$ follows directly from the fact that $u' \in (\mathscr{V}_{q'})'$ by $u \in \mathscr{W}$. By the choice of the parameter p_λ and the inclusion $\mathscr{W} \subset L^r$, the map

$$\mathscr{P} \times \mathscr{W} \to (\mathscr{V}_{q'})',\ (\lambda, u) \mapsto \lambda u$$

is continuous and by its bilinearity also continuously differentiable. The map

$$\mathscr{P} \times \mathscr{W} \to (\mathscr{V}_{q'})',\ (D, u) \mapsto -div(D\nabla u)$$

is bilinear, (locally Lipschitz) continuous and therefore also continuously differentiable. □

3.3.2 Superposition Operators

For analyzing Hölder and Lipschitz continuity as well as differentiability of the nonlinear right-hand side of (3.1) we use results on superposition operators, see the standard references [1, 16]. Here we just summarize the relevant results, the reader is referred to [14] for their proofs. We start with a technical remark.

Remark 3.1. The function $\varphi(x) = 0.5(x/\sqrt{x^2+1} + 1)$ is globally Lipschitz continuous and so are all its derivatives. In particular, they decay as $\varphi^{(n)}(x) = o(|x|^n)$ for $|x| \to \infty$. Typical other examples for such sigmoidal functions are arctan and tanh.

Now we return to our quest of analyzing superposition operators. In all the generic statements and claims below, the space Z is a finite measure space.

Definition 3.1. A function $f : Z \times \mathbb{R} \to \mathbb{R}$, $(z, u) \mapsto f(z, u)$ which is measurable in z and continuous in u is called a *Caratheodory function*.

Remark 3.2. Let f be a Caratheodory function as above. Let $v : Z \to \mathbb{R}$ be a measurable function. Then $f(\cdot, v(\cdot))$ is a measurable function. The proof can be found in [1].

Definition 3.2. Let $1 \leq q, p \leq \infty, q \neq \infty$, and f be a Caratheodory function as defined above. Suppose that f satisfies a growth estimate (in case of $p \neq \infty$)

$$|f(z, u(z))|^q \leq C(h(z)^q + |u(z)|^p) \quad \text{a.e. on } Z,$$

or (in case of $p = \infty$)

$$|f(z, u(z))|^q \leq C(h(z)^q + \|u(z)\|_{L_\infty(Z)}) \quad \text{a.e. on } Z,$$

where $h \geq 0$ is from $L_q(Z)$. Then we define the operator $(\mathscr{B}(u))(z) := f(z, u(z))$.

Theorem 3.4. *The operator $\mathscr{B}$ defined above is a bounded, continuous map $L_p(Z) \to L_q(Z)$, $q \neq \infty$ which satisfies*

$$\|\mathscr{B}(u)\|_{L_q(Z)} \leq C(\|h\|_{L_q(Z)} + \|u\|^{\frac{p}{q}}_{L_p(Z)}).$$

The nonlinear operator $\mathscr{F} = R\Phi(Wu)$ in (3.1) is indeed a function of u and of the parameters R, W, i.e. $\mathscr{F} = \mathscr{F}(R, W, u)$. With a slight abuse of notation we will frequently write $\mathscr{F}(u)$, $\mathscr{F}(R)$, $\mathscr{F}(W)$ instead of $\mathscr{F}(R, W, u)$, whenever it is clear how the other arguments are fixed.

Remark 3.3. The compositions $u \mapsto \mathscr{F}(u), \mathscr{F}_i(u) = R_i \tilde{\mathscr{F}}_i((Wu)_i) = R_i \Phi((Wu)_i)$ and $W \mapsto \mathscr{F}(W), \mathscr{F}_i(W) = R_i \tilde{\mathscr{F}}_i((Wu)_i) = R_i \Phi((Wu)_i)$ are continuous.

A stronger type of continuity is stated under additional assumptions.

Theorem 3.5. *Let $\mathscr{B} : L_p(Z) \to L_q(Z)$ be a continuous superposition operator as in the preceding Theorem 3.4 generated by a Caratheodory function f, $1 \leq q \leq p < \infty$. f is assumed to be globally Lipschitz continuous in u, such that we have an a.e. pointwise Lipschitz estimate*

$$|f(z, u) - f(z, v)| \leq T(z)\|u - v\|_{\mathbb{R}^d},$$

where $T \geq 0$, $T(z) \in L_r(Z)$, $r = \frac{pq}{p-q}$. Then the operator $\mathscr{B}$ is globally Lipschitz continuous with Lipschitz constant $\|T\|_{L_r(Z)}$.

For later use we have to compute the strong (or Fréchet) derivative of $\mathscr{F}$ with respect to u and W. A first step towards this result is to define this operator formally and to prove its continuity properties. Define the exponents

$$\mathfrak{p} = (1/p_R + 1/p_W + 1/r)^{-1}, \quad \mathfrak{a} = \frac{\mathfrak{p}q}{\mathfrak{p} - q}$$

and the multiplication space $L_{\mathfrak{a}}(0,T,L_{\mathfrak{a}}(U,\mathbb{R}^d)) =: \mathbb{M}_{\mathfrak{p},q}$. Now we will utilize the liberty in the choice of r: For fixed exponents p_R and p_W we find

$$\begin{aligned} \mathfrak{p} = \mathfrak{p}(r) &\to \left(\frac{1}{p_R} + \frac{1}{p_W}\right)^{-1} && \text{for } r \to \infty, \\ \mathfrak{a} = \mathfrak{a}(r) &\to C_{\mathscr{P},1}(p_R, p_W) && \text{for } r \to \infty. \end{aligned}$$

Therefore, we assume r to be sufficiently large such that $\mathfrak{a} \leq r$. This will suit the analysis concerning the argument u. A similar reasoning implies $\mathfrak{a} < p_W$ for r large enough (see [14, pp.31] for details). We will use this space in the following technical lemma.

Lemma 3.1. *The superposition operators*

$$L_r([0,T],L_r(U,\mathbb{R}^d)) \to L_{\mathfrak{a}}([0,T],L_{\mathfrak{a}}(U,\mathbb{R}^d)), \quad u \mapsto \Phi'(Wu)$$

and

$$\mathscr{P}_W \to L_{\mathfrak{a}}([0,T],L_{\mathfrak{a}}(U,\mathbb{R}^d)), \quad W \mapsto \Phi'(Wu)$$

are Lipschitz continuous for $\mathfrak{a} \leq r$ ($\mathfrak{a} \leq p_W$).

From now on we will assume that $p_W > \mathfrak{a}$ and $r > \mathfrak{a}$. A statement similar to the above (yet without proof) can be found in [1, Rem. p. 105]. The next step is to define the spaces

$$\mathbb{M}_{r,q} := \mathscr{L}(L_r([0,T],L_r(U,\mathbb{R}^d)),L_q([0,T],L_q(U,\mathbb{R}^d)))$$

and

$$\mathbb{M}_{p_W,q} := \mathscr{L}(L_{p_W}([0,T],L_{p_W}(U,\mathbb{R}^{d\times d})),L_q([0,T],L_q(U,\mathbb{R}^d))).$$

Corollary 3.1. *The superposition operators*

$$L_r([0,T],L_r(U,\mathbb{R}^d)) \to \mathbb{M}_{r,q},\ u \mapsto (h \mapsto R\Phi'(Wu)Wh)$$

and

$$\mathscr{P}_W \to \mathbb{M}_{p_W,q},\ W \mapsto (\tilde{W} \mapsto R\Phi'(Wu)\tilde{W}u)$$

are Lipschitz continuous.

To conclude this section, we state a result about differentiability. Take the parameter function W, the signal response function Φ, and q, p_W, p_R as before. Then we obtain

Theorem 3.6. *The operator* $u \mapsto \mathscr{F}(u)$ *as defined above is continuously differentiable and its (Lipschitz continuous) derivative is given by*

$$\mathscr{F}_u : L_r([0,T], L_r(U, \mathbb{R}^d)) \to \mathbb{M}_{r,q},$$
$$\mathscr{F}_u(\tilde{u}) = R\Phi'(W\tilde{u})W.$$

An analogous result holds for the other argument W in our nonlinear superposition operator: u is continuous on $[0,T] \times \overline{U}$ by Theorem 3.1, so it is certainly an L_∞-function.

Theorem 3.7. *The superposition operator* $W \mapsto \mathscr{F}(Wu)$ *mapping between spaces* $\mathscr{P}_W \to L_q([0,T], L_q(U, \mathbb{R}^d))$ *as defined above is continuously differentiable and the (Lipschitz continuous) derivative is the respective one stated in Corollary 3.1. The derivative is given by*

$$\mathscr{P}_W \to \mathbb{M}_{p_W,q},\ W \mapsto (\tilde{W} \mapsto R\Phi'(Wu)\tilde{W}u).$$

The continuity of $\mathscr{F}(R)$ is trivial, since $R \mapsto \mathscr{F}(R)$ is just a linear (and by Eq. (3.3) continuous) map, which entails even continuous differentiability. So we have ultimately shown that the operator on the right-hand side is partially continuously differentiable, hence totally differentiable. The derivative is Lipschitz continuous in each entry.

3.3.3 Solution of the Model PDE

Our model PDE (3.1) can be rephrased as an evolution equation by using the definitions of the operators $\mathscr{A}$ and $\mathscr{F}$.

Definition 3.3. $u \in \mathscr{W}$ is a *weak solution* of (3.1), if

$$u' + \mathscr{A}(\pi, u) = \mathscr{F}(\pi, u) \text{ in } (\mathscr{V}_{q'})', \quad u(0) = u_0 \in G. \tag{3.4}$$

This definition is made under the constraints $p_D = \infty$ and $p_\lambda, p_R, p_W > 4q$ and $q \in (n, n+\varepsilon)$. In order to prove existence and uniqueness of a solution we will first fix the argument u in $\mathscr{F}$ to some arbitrary $w \in \mathscr{C}([0,T]; G)$ and obtain $\mathscr{F}(\pi, w(\pi)) = f(\pi) \in (\mathscr{V}_{q'})'$. The system now decouples for each component of u and we even obtain a linear equation. So we restate the definition of a solution for this simplified Cauchy problem as a function $u \in \mathscr{W}$ which satisfies:

$$u' + \mathscr{A}(\pi, u) = f \text{ in } (\mathscr{V}_{q'})', \quad u(0) = u_0 \in G. \tag{3.5}$$

Recent results on maximal parabolic regularity [2, 9] can be used to proof the following statement, for a proof see [14, Thm.8.6.3].

Theorem 3.8. *There exists an $\tilde{\varepsilon}$ sufficiently small and depending on $U, C_{\mathscr{P},1}$ and $C_{\mathscr{P},2}$ such that for any $f \in (\mathscr{V}_{q'})'$ and $q \in (n, n+\tilde{\varepsilon})$ there exists a unique solution of the linear problem (3.5). This solution depends continuously on f and the initial data u_0, i.e.*

$$\|u\|_{\mathscr{W}} \leq C(\|f\|_{(\mathscr{V}_{q'})'} + \|u_0\|_G),$$

where the constant C depends on $U, C_{\mathscr{P},1}$ and $C_{\mathscr{P},2}$.

We can slightly extend the scope of the theorem above by the common trick of exponential shifting and by reversing the time axis:

Remark 3.4. The preceding theorem also holds without the non-negativity constraint of λ as long as $\lambda \in L_\infty$ is pointwise bounded. Moreover, the time-reversed equation

$$-u' + \mathscr{A}(\pi, u) = f, \quad u(T) = u_T \in G$$

has a unique solution under the assumptions of the previous theorem.

After treating the Cauchy problem with simplified right-hand side, we will now address the original problem. This uniqueness result will mark the final point of this section and will constitute the major building block for the analysis of the parameter-to-state map in the next section, since it ensures that the parameter-to-state map is well defined. The admissible range of the parameters will be $\mathscr{P}$ as before. We rephrase our original model equation (3.1) as a Cauchy problem

$$u \in L_2([0,T];V) : \; u' + \mathscr{A}(\pi, u) = \mathscr{F}(\pi, u) \text{ in } L_2([0,T];V'), \; u(0) = u_0. \tag{3.6}$$

To establish existence and uniqueness for the nonlinear setting, we make use of Banach's fixed point theorem in $\mathscr{C}([0,T];L_2(U,\mathbb{R}^d))$. Exploiting the continuous embeddings $\mathscr{W} \hookrightarrow \mathscr{C}([0,T];G)$ and $G \hookrightarrow L_2(U,\mathbb{R}^d)$ one can follow the ideas of [7, pp. 500]. This leads to a solution in $\mathscr{C}([0,T];L_2(U,\mathbb{R}^d))$ for the nonlinear equation. Finally, one bootstraps back to the regularity of $\mathscr{W}$ by using Theorem 3.8. We summarize the key finding of this section on the solvability of our nonlinear model equation:

Theorem 3.9. *The nonlinear Cauchy problem as stated in (3.6) has a unique solution.*

3.4 The Parameter-to-State Map

In this section we will show the continuous differentiability of the parameter-to-state map, see Sect. 3.4.1. Then in Sect. 3.4.2 we verify its Lipschitz continuity and the uniform continuity of the derivative.

3.4.1 Continuity and Differentiability of the Parameter-to-State Map

In order to prove that the parameter-to-state map

$$\mathscr{D} : \mathscr{P} \to \mathscr{W}, (D, \lambda, R, W) \mapsto u$$

is continuous and differentiable we will make use of its implicit definition by application of the implicit function theorem. For this purpose, we introduce the operator

$$\begin{aligned} \mathscr{C} : \mathscr{P} \times \mathscr{W} &\to G \times (\mathscr{V}_{q'})' \\ \pi = ((D, \lambda, R, W), u) &\mapsto (u(0) - u_0, u' + \mathscr{A}(\pi, u) - \mathscr{F}(\pi, u)). \end{aligned}$$

The point evaluation at 0 in the first term is well-defined due to Theorem 3.1, the second component of $\mathscr{C}$ is well-defined by the assumptions on u, $\mathscr{A}$, and $\mathscr{F}$. For the next lemma we fix the first argument and show continuous differentiability.

Lemma 3.2. *The map $\mathscr{S} = \mathscr{C}(\pi_0, \cdot) : \mathscr{W} \to G \times (\mathscr{V}_{q'})'$ is continuously differentiable and the derivative at any $y \in \mathscr{W}$ is an isomorphism from $\mathscr{W}$ to $G \times (\mathscr{V}_{q'})'$.*

Proof. For the first component of the image of $\mathscr{S}$ we see that the evaluation map $u \mapsto u(0)$ is a linear map. It is furthermore continuous, since $\|u(0)\|_G \leq \|u\|_{C(0,T;G)} \leq C_{embed}\|u\|_{\mathscr{W}}$.

For the second component we consider each summand separately. By Theorem 3.3 we know that the operator $\frac{d}{dt} + \mathscr{A}$ is in C^1. The third summand $u \mapsto \mathscr{F}(\pi, u) = R \cdot \Phi(Wu)$ has been discussed in the conclusions of Sect. 3.3.2, where we proved differentiability of the superposition operator with respect to u and the Lipschitz continuity of the derivative. So we may deduce that $\mathscr{S}$ is a continuously differentiable map. For the derivative $\mathscr{S}'$ in the direction of some $h \in W$, we obtain

$$\mathscr{S}'(y)h = (h(0), h' + \mathscr{A}(\pi, h) - \partial_u \mathscr{F}(\pi, y)h).$$

Define $\tilde{\lambda} = \lambda - \partial_u \mathscr{F}(\pi, y)$ and denote the resulting linear differential operator by $\tilde{\mathscr{A}}$. By the (componentwise) boundedness of $\partial_u \mathscr{F}$ it is clear that $\tilde{\lambda}$ satisfies the requirements of Sect. 3.3.3, see Remark 3.4. Therefore we may apply the findings of that subsection to $\mathscr{S}'(u)$: For arbitrary $(v_0, f) \in G \times \mathscr{V}'$ there exists a solution h for $(h(0), h' + \mathscr{A}(\pi, h) - \partial_u \mathscr{F}(\pi, y)h) = (v_0, f)$. This ensures surjectivity of $\mathscr{S}'(y)$. The uniqueness of this solution ensures injectivity of $\mathscr{S}'(y)$. The continuous dependence of the solution on the right-hand side, see Theorem 3.8 and Remark 3.4, is equivalent to the continuity of the solution map $(\mathscr{S}'(y))^{-1} : (v_0, f) \mapsto h$. So we may conclude that for any $y \in \mathscr{W}$, $\mathscr{S}'(y)$ is an isomorphism from $\mathscr{W}$ to $G \times \mathscr{V}'$. □

Lemma 3.3. *The map* $\mathscr{P} = \mathscr{C}(\cdot, u) : \mathscr{P} \to G \times (\mathscr{V}_{q'})'$ *is continuously differentiable.*

Proof. Following the arguments of Sect. 3.3.1 we see that $(D, \lambda) \mapsto \mathscr{A}$ is continuously differentiable. By the findings of Sect. 3.3.2 we obtain the continuous differentiability of the operator $(R, W) \mapsto \mathscr{F}$. Hence, continuous differentiability of all partial derivatives is confirmed and we may conclude that $\mathscr{P}$ is totally continuously differentiable. □

We obtain the following explicit expressions of the partial derivatives

$$\mathscr{C}_u(\pi_0, u)(h) = (h(0), h' + \mathscr{A}(\pi_0, h) - R_0\Phi'(W_0u)W_0h),$$
$$\mathscr{C}_\pi(\pi_0, u)(\pi_1) = (0,\ (\lambda_1 u - \nabla \cdot (D_1 \nabla u) - R_1\Phi(W_0u) - R_0\Phi'(W_0u)W_1u)).$$

The last two lemmata enable us to state the following application of the implicit function theorem in the formulation of [14, Thms. 8.7.8.,8.7.9].

Theorem 3.10. *The parameter-to-state map* $\mathscr{D} : \mathscr{P} \to \mathscr{W}, (D, \lambda, R, W) \mapsto u$ *of (3.4) is continuously differentiable and the derivative is given by*

$$\mathscr{D}'(\pi_0)(\pi_1) = -(\mathscr{C}_u(\pi_0, u))^{-1} \circ \mathscr{C}_\pi(\pi_0, u)(\pi_1) =: v.$$

Hence, v is a solution of the differential equation: $v(0) = 0$,

$$v' + \mathscr{A}(\pi_0, v) - R_0\Phi'(W_0u)W_0v = -\mathscr{A}(\pi_1, u) + R_1\Phi(W_0u) + R_0\Phi'(W_0u)W_1u,$$

where $u = u(\pi_0) = \mathscr{D}(\pi_0)$.

Proof. The existence proof from Theorem 3.9 ensures the existence of a root of the map $\mathscr{C}$. The last two lemmata supply the other parts of the conditions in the implicit function theorem. □

The underlying implicit function theorem crucially hinges on the completeness of the space $\mathscr{W}$ and the continuity (differentiability) of the operator $\mathscr{A}$ to utilize Banach's fixed point theorem. This leaves no alternative to the definition $p_D = \infty$, at least not, when one relies on the implicit function theorem.

3.4.2 Properties of the Derivative of the Parameter-to-State Map

The last result, namely the explicit formula for the derivative of the parameter-to-state map at some π_0, enables us to investigate further interesting and useful properties of $\mathscr{D}'$. Our inspection will be divided into several lemmata, which then allow us to show the (local) Lipschitz continuity of the operators $\mathscr{D}$ and $\mathscr{D}'$ on bounded sets.

Lemma 3.4. *The map $\mathscr{D}'(\cdot) : \mathscr{P} \to \mathscr{L}(\mathscr{P}, \mathscr{W})$ is bounded on $\mathscr{P}$.*

Proof. Denote $u = \mathscr{D}(\pi_0)$. Let $\tilde{\lambda} = \lambda - R_0\Phi'(W_0u)W_0$ and $\tilde{\pi} = (D_0, \tilde{\lambda}, R_0, W_0)$, and $r(\pi_0)(\pi_1) = -\mathscr{A}(\pi_1, u) + R_1\Phi(W_0u) + R_0\Phi'(W_0u)W_1u$. Let π_1 be an arbitrary admissible displacement vector. By Remark 3.4, $\mathscr{D}'(\pi_0)(\pi_1)$ is the solution of the Cauchy problem $v(0) = 0,\ v' + \mathscr{A}(\tilde{\pi}, v) = r(\pi_0)(\pi_1)$. By the continuous dependence of the PDE solution on the right-hand side (see Sect. 3.3.3, Theorem 3.8), we have the linear stability estimate

$$\|v\|_{\mathscr{V}_q} \le \|v\|_{\mathscr{W}} \le C\,\|r(\pi_0)(\pi_1)\|_{(\mathscr{V}_{q'})'}.$$

An analogous estimate holds for the solution of our model PDE (Eq. (3.4)):

$$\begin{aligned}\|u\|_{\mathscr{W}} &\le C(\|R_0\Phi(W_0u)\|_{\mathscr{V}'} + \|u_0\|_G)\\ &\le C(\|R_0\|_{(\mathscr{V}_{q'})'} + \|u_0\|_G) \le C(\|R_0\|_{\mathscr{P}_R} + \|u_0\|_G) \le C(\|\pi_0\|_{\mathscr{P}} + \|u_0\|_G).\end{aligned}$$

The last estimate implies

$$\begin{aligned}\|[r(\pi_0)](\pi_1)\|_{(\mathscr{V}_{q'})'} &\le \|\mathscr{A}(\pi_1, u)\| + \|R_1\Phi(W_0u)\| + \|R_0\Phi'(W_0u)W_1u\|\\ &\le C\,\|u\|_{\mathscr{W}}\|\pi_1\|_{\mathscr{P}} + C_\Phi\|R_1\|_{\mathscr{P}} + \|R_0\|_{\mathscr{P}_R}\|W_1\|_{\mathscr{P}_W}\|u\|_{\mathscr{W}}\\ &\le (C(\|\pi_0\|_{\mathscr{P}} + \|u_0\|_G + 1) + C\,\|\pi_0\|_{\mathscr{P}}(\|\pi_0\|_{\mathscr{P}} + \|u_0\|_G))\|\pi_1\|_{\mathscr{P}},\end{aligned}$$

i.e., $r(\pi_0)$ is bounded. So

$$\frac{\|v\|_{\mathscr{W}}}{\|\pi_1\|_{\mathscr{P}}} \le \frac{\|[r(\pi_0)](\pi_1)\|_{(\mathscr{V}_{q'})'}}{\|\pi_1\|_{\mathscr{P}}} \le C((\|\pi_0\|_{\mathscr{P}} + \|u_0\|_G)(1 + \|\pi_0\|_{\mathscr{P}}) + 1).$$

This implies the boundedness of the operator $\mathscr{D}'$:$\|\mathscr{D}'(\pi_0)\|_{\mathscr{L}(\mathscr{P},\mathscr{W})} \le C$. □

Using the mean value theorem (compare [19, Satz III.5.4 b)]) on $\mathscr{D}$ we obtain

Lemma 3.5. *The map $\mathscr{D} : \mathscr{P} \to \mathscr{W}$ is Lipschitz continuous on convex, bounded sets.*

Proof. By the mean value theorem we have for some $\pi_0, \pi_2 \in \mathscr{P}$, and any $\pi_\theta = \theta\pi_0 + (1-\theta)\pi_2$, $\theta \in (0,1)$

$$\|\mathscr{D}(\pi_0) - \mathscr{D}(\pi_2)\|_{\mathscr{W}} \le \sup_{\theta\in(0,1)} (\|\mathscr{D}'(\pi_\theta)\|_{\mathscr{L}(\mathscr{P},\mathscr{W})})\|\pi_0 - \pi_2\|_{\mathscr{P}}.$$

The supremum in the estimate exists and is bounded uniformly by the preceding lemma. □

We would like to establish a similar result for $\mathscr{D}'$. However $\mathscr{D}$ involves the inversion of an operator and we will only obtain local Lipschitz continuity. We consider the superposition operators $\mathscr{C}_\pi$ and $\mathscr{C}_u$ separately.

Lemma 3.6. *The operator* $\mathscr{C}_u : \mathscr{P} \times \mathscr{W} \to \mathscr{L}(\mathscr{W}, G \times (\mathscr{V}_{q'})')$ *defined by* $(\pi_0, u) \mapsto (h \mapsto (h(0), h' + \mathscr{A}(\pi_0, h) - R_0\Phi'(W_0u)W_0h))$ *is Lipschitz continuous.*

Proof. The first component $h \mapsto h(0)$ is independent of the argument (π_0, u) and therefore Lipschitz continuous. Considering the second component we start with the summand $(\pi_0, u) \mapsto (h \mapsto h' + \mathscr{A}(\pi_0, h))$. This is a bounded linear operator and hence globally Lipschitz. For the last summand $(\pi_0, u) \mapsto (h \mapsto R_0\Phi'(W_0u)W_0h)$ we refer to Corollary 3.1. For $(\pi_0, u), (\pi_2, v)$, where $\pi_0, \pi_2 \in \mathscr{P}$ we get the estimate

$$\begin{aligned}&\|[R_2\Phi'(W_2v)W_2 - R_0\Phi'(W_0u)W_0]h\|_{(\mathscr{V}_{q'})'}\\ &\leq C(\|R_0 - R_2\|_{p_R} + (\|W_2 - W_0\|_{p_W} + \|u_2 - u_0\|_{\mathscr{W}}))\|h\|_{L^r(0,T,L^r(UT],\mathbb{R}^d))}.\end{aligned}$$

The constant C in the last estimate depends on the bound of the considered domain and on $\sup(\Phi')$. □

A helpful technical lemma is the following:

Lemma 3.7. *On the domain* $\mathscr{P} \times \mathscr{W}$ *the map* $\mathscr{C}_u^{-1}$ *is bounded in dependence of* $C_{\mathscr{P},1}, C_{\mathscr{P},2}$, *the domains* U *and* $[0, T]$.

Proof. This follows by maximal regularity results stated in [14, Lemma 4.2.4, Thm. 8.6.3], which are based on the results of [9] and [2]. □

Using the fact that operator inversion is locally Lipschitz continuous, we can deduce the following

Corollary 3.2. *The map* $(\mathscr{C}_u(\cdot,\cdot))^{-1} : \mathscr{P} \times \mathscr{W} \to \mathscr{L}(G \times (\mathscr{V}_{q'})', \mathscr{W})$ *is locally Lipschitz continuous with uniform Lipschitz constant* $(C_{3.7})^2C_{3.6}$, *where* $C_{3.6}$ *and* $C_{3.7}$ *denote the respective bounds of Lemmas 3.6 and 3.7.*

Proof. We take for (π_0, u) the bounded neighborhood

$$A = \{(\pi, v) : \|(\pi_0, u) - (\pi_2, v)\|_{\mathscr{P}\times\mathscr{W}} < C_{3.7}(2C_{3.6})^{-1}\}.$$

Then on A it holds

$$\|\mathscr{C}_u(\pi_0, u) - \mathscr{C}_u(\pi_2, v)\|_{L(\mathscr{W},(\mathscr{V}_{q'})')} \leq C_{3.6}\left[C_{3.7}(2C_{3.6})^{-1}\right] = (1/2)C_{3.7}.$$

For two invertible operators $x, y \in \mathscr{L}(X, Y)$ with $\|x - y\| \leq 1/2C$, it holds $\|x^{-1} - y^{-1}\| \leq C^2\|x - y\|$, see [14, Cor. 8.7.4] (based on [3, Ch. 50]). Hence, we may conclude

$$\begin{aligned}\|\mathscr{C}_u^{-1}(\pi_0, u) - \mathscr{C}_u^{-1}(\pi_2, v)\|_{L((\mathscr{V}_{q'})',\mathscr{W})} &\leq (C_{3.7})^2\|\mathscr{C}_u(\pi_0, u) - \mathscr{C}_u(\pi_2, v)\|_{L(\mathscr{W},(\mathscr{V}_{q'})')}\\ &\leq (C_{3.7})^2C_{3.6}\|(\pi_0, u) - (\pi_2, v)\|_{\mathscr{P}\times\mathscr{W}}.\end{aligned}$$

This establishes the claim. □

Lemma 3.8. *The operator*

$$\mathscr{C}_\pi : \mathscr{P} \times \mathscr{W} \to \mathscr{L}(\mathscr{P}, G \times (\mathscr{V}_{q'})') \text{ defined as}$$
$$\mathscr{C}_\pi(p,u)(\pi_1) = (0, \mathscr{A}(\pi_1,u) - R_1\Phi(W_0u) - R_0\Phi'(W_0u)W_1u)$$

is Lipschitz continuous on bounded sets with a Lipschitz constant depending on the diameter of the considered set and on Φ'.

Proof. The first component of $\mathscr{C}_\pi$, i.e. the point evaluation, is clear. For the second component take $u, v \in \mathscr{W}$ and $\pi_0, \pi_2 \in \mathscr{P}$, and the displacement vector π_1. Consider

$$\begin{aligned}&\|[\mathscr{A}(\pi_1,u) - R_1\Phi(W_0u)] - [\mathscr{A}(\pi_1,v) - R_1\Phi(W_0v)]\|_{(\mathscr{V}_{q'})'}\\ &\le \|\pi_1\|_{\mathscr{P}}\|u-v\|_{\mathscr{W}} + \|\pi_1\|C_\Phi(\|W_0 - W_2\|_{p_W} + \|u-v\|_{\mathscr{W}}).\end{aligned}$$

Similar to the proof of Lemma 3.6 we invoke again the partial Lipschitz continuity of partial derivatives of $\mathscr{F}$ (as in Corollary 3.1) and obtain

$$\begin{aligned}&\|R_0\Phi'(W_0u)W_1u - R_2\Phi'(W_2v)W_1v\|_{(\mathscr{V}_{q'})'}\\ &\le \|R_0\Phi'(W_0u)W_1u - R_0\Phi'(W_2u)W_1u\|_{(\mathscr{V}_{q'})'} + \|R_0\Phi'(W_2u)W_1u - R_0\Phi'(W_2u)W_1v\|_{(\mathscr{V}_{q'})'}\\ &+ \|R_0\Phi'(W_2u)W_1v - R_0\Phi'(W_2v)W_1v\|_{(\mathscr{V}_{q'})'} + \|R_0\Phi'(W_2v)W_1v - R_2\Phi'(W_2v)W_1v\|_{(\mathscr{V}_{q'})'}\\ &\le \|\pi_0\|_{\mathscr{P}}\|\pi_1\|_{\mathscr{P}}\left(L_{\Phi'}\left[\|W_0 - W_2\|_{\mathscr{P}}\|u\|_{L_r(U\times[0,T];\mathbb{R}^d)}\right]\|u\|_{\mathscr{W}} + \sup(\Phi')\|u-v\|_{\mathscr{W}}\right)\\ &+ \|\pi_1\|_{\mathscr{P}}\|v\|_{\mathscr{W}}\left(\|\pi_0\|_{\mathscr{P}}L_{\Phi'}\left[\|W_2\|_{\mathscr{P}}\|u-v\|_{L_r(U\times[0,T];\mathbb{R}^d)}\right] + \|\pi_0-\pi_2\|_{\mathscr{P}}\sup(\Phi')\right)\\ &\le C(\|u\|_{\mathscr{W}}, \|v\|_{\mathscr{W}}, \|\pi_0\|_{\mathscr{P}})\|\pi_1\|_{\mathscr{P}}\left[\|u-v\|_{\mathscr{W}} + \|\pi_0-\pi_2\|_{\mathscr{P}}\right].\end{aligned}$$

From all these estimates we obtain Lipschitz continuity of $\mathscr{C}_\pi$ on bounded sets. □

Lemma 3.8 and Corollary 3.2 lead to a local estimate:

Theorem 3.11. *$\mathscr{D}'$ is locally Lipschitz continuous with uniform Lipschitz constant.*

Proof. Take π_0, π_1, π_2 as in the above proofs. Then for $\|\pi_0 - \pi_2\|$ sufficiently small, by Lipschitz continuity $\|(\mathscr{C}_u)^{-1}(\pi_0, \mathscr{D}(\pi_0)) - (\mathscr{C}_u)^{-1}(\pi_2, \mathscr{D}(\pi_2))\|$ becomes small enough such that we may apply Corollary 3.2. Lemma 3.8 implies the estimate

$$\begin{aligned}&\|[\mathscr{D}'(\pi_0) - \mathscr{D}'(\pi_2)]\|_{L(\mathscr{P},\mathscr{W})}\\ &\le \|\left(\mathscr{C}_u(\pi_0,\mathscr{D}(\pi_0))\right)^{-1}\left(\mathscr{C}_\pi(\pi_0,\mathscr{D}(\pi_0))\right) - \left(\mathscr{C}_u(\pi_2,\mathscr{D}(\pi_2))\right)^{-1}\left(\mathscr{C}_\pi(\pi_0,\mathscr{D}(\pi_0))\right)\|\\ &+ \|\left(\mathscr{C}_u(\pi_2,\mathscr{D}(\pi_2))\right)^{-1}\left(\mathscr{C}_\pi(\pi_0,\mathscr{D}(\pi_0)) - \left(\mathscr{C}_u(\pi_2,\mathscr{D}(\pi_2))\right)^{-1}\left(\mathscr{C}_\pi(\pi_2,\mathscr{D}(\pi_2))\right)\|\\ &\le \|\left(\mathscr{C}_u(\pi_0,\mathscr{D}(\pi_0))\right)^{-1} - \left(\mathscr{C}_u(\pi_2,\mathscr{D}(\pi_2))\right)^{-1}\left(\mathscr{C}_\pi(\pi_0,\mathscr{D}(\pi_0))\right)\|\end{aligned}$$

$$+ \| \left(\mathscr{C}_u(\pi_2, \mathscr{D}(\pi_2))\right)^{-1} \left(\mathscr{C}_\pi(\pi_0, \mathscr{D}(\pi_0))\right) - \left(\mathscr{C}_\pi(\pi_2, \mathscr{D}(\pi_2))\right) \|$$
$$\leq C(\|\pi_0 - \pi_2\|_{\mathscr{P}} + \|\mathscr{D}(\pi_0) - \mathscr{D}(\pi_2)\|_{\mathscr{W}}) \|\mathscr{C}_\pi(\pi_0, \mathscr{D}(\pi_0))\|_{L(\mathscr{P}, (\mathscr{V}_{q'})')}$$
$$+ C(\|\pi_0 - \pi_2\|_{\mathscr{P}} + \|\mathscr{D}(\pi_0) - \mathscr{D}(\pi_2)\|_{\mathscr{W}}).$$

The Lipschitz continuity of $\mathscr{D}$ yields the claim. □

References

1. Appell, J., Zabrejko, P.: Nonlinear Superposition Operators. Cambridge University Press, Cambridge, UK (1990)
2. Arendt, W., Chill, R., Fornaro, S., Poupaud, C.: L_p-maximal regularity for non-autonomous evolution equations. J. Differ. Equ. **237**, 1–26 (2007)
3. Berberian, S., Lectures in Functional Analysis and Operator Theory. Springer, New York/Heidelberg/Berlin (1974)
4. Bredies, K., Bonesky, T., Lorenz, D., Maass, P., A generalized conditional gradient method for non-linear operator equations with sparsity constraints. Inverse Probl. **23**, 2041–2058 (2007)
5. Chegini, N., Dahlke, S., Friedrich, U., Stevenson, R.: Piecewise tensor product wavelet bases by extension and approximation rates. In: S. Dahlke et al. (eds.), Extraction of Quantifiable Information from Complex Systems, Lecture Notes in Computational Science and Engineering 102, doi: 10.1007/978-3-319-08159-5_3 (2014)
6. Cohen, A.: Numerical Analysis of Wavelet Methods. Studies in Mathematics and Its Applications, vol. 32, 1st edn. Elsevier, Amsterdam (2003)
7. Evans, L.C.: Partial Differential Equations. American Mathematical Society, Providence (2008)
8. Grasmair, M., Haltmeier, M., Scherzer, O.: Sparse regularization with l^q penalty term. Inverse Probl. **24**, 055020 (2008)
9. Haller-Dintelmann, R., Rehberg, J.: Maximal parabolic regularity for divergence operators including mixed boundary conditions. J. Differ. Equ. **247**, 1354–1396 (2009)
10. Jin, B., Maass, P.: A reconstruction algorithm for electrical impedance tomography based on sparsity regularization. ESAIM: Control Optim. Calc. Var. **18(4)**, 1027–1048 (2012)
11. Jin, B., Maass, P.: Sparsity regularization for parameter identification problems. Inverse Probl. **28**, 123001 (2012)
12. Mjolsness, E., Sharp, D., Reinitz, J.: A connectionist model of development. J. Theor. Biol. **152**, 429–453 (1991)
13. Reinitz, J., Sharp, D.: Mechanism of eve stripe formation. Mech. Dev. **49**, 133–158 (1995)
14. Ressel, R.: A parameter identification problem for a nonlinear parabolic differential equation, PhD-Thesis (Bremen) (2012)
15. Rondi, L., Santosa, F.: Enhanced electrical impedance tomography via the Mumford-Shah functional. ESAIM, Control Optim. Calc. Var. **6**, 517–538 (2001)
16. Runst, T., Sickel, W.: Sobolev Spaces of Fractional Order, Nemytskij Operators, and Nonlinear Partial Differential Equations. deGruyter, Berlin (1996)
17. Showalter, R.E.: Monotone Operators in Banach Space and Nonlinear Partial Differential Equations. Mathematical Surveys and Monographs, vol. 49. American Mathematical Society, Providence (1997)
18. Triebel, H., Interpolation Theory, Function Spaces, Differential Operators. Johann Ambrosius Barth Verlag, Heidelberg (1995)
19. Werner, D., Funktionalanalysis. Springer, Berlin/Heidelberg (2000)

Chapter 4
Piecewise Tensor Product Wavelet Bases by Extensions and Approximation Rates

Nabi G. Chegini, Stephan Dahlke, Ulrich Friedrich, and Rob Stevenson

Abstract In this chapter, we present some of the major results that have been achieved in the context of the DFG-SPP project "Adaptive Wavelet Frame Methods for Operator Equations: Sparse Grids, Vector-Valued Spaces and Applications to Nonlinear Inverse Problems". This project has been concerned with (nonlinear) elliptic and parabolic operator equations on nontrivial domains as well as with related inverse parameter identification problems. One crucial step has been the design of an efficient forward solver. We employed a spatially adaptive wavelet Rothe scheme. The resulting elliptic subproblems have been solved by adaptive wavelet Galerkin schemes based on generalized tensor wavelets that realize dimension-independent approximation rates. In this chapter, we present the construction of these new tensor bases and discuss some numerical experiments.

4.1 Introduction

The aim of the project "Adaptive Wavelet Frame Methods for Operator Equations" has been the development of optimal convergent adaptive wavelet schemes for elliptic and parabolic operator equations on nontrivial domains. Moreover, we have been concerned with the efficient treatment of related parameter identification problems. For the design of the efficient forward solver, we used variants of the recently developed adaptive wavelet schemes for elliptic operator equations, see., e.g., [4, 12]. (This list is clearly not complete). As usually, the construction of a suitable wavelet basis on the underlying domain is a nontrivial bottleneck. We attacked this problem by generalizing the construction of tensor wavelets on the hypercube to general domains. The resulting fully adaptive solver for the elliptic forward

N.G. Chegini • R. Stevenson
University of Amsterdam, Korteweg-de Vries (KdV) Institute for Mathematics, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands
e-mail: n.godarzvandchegini@uva.nl; R.P.Stevenson@uva.nl

S. Dahlke (✉) • U. Friedrich
Philipps-University of Marburg, Hans Meerwein Str., Lahnberge, 35032 Marburg, Germany
e-mail: dahlke@mathematik.uni-marburg.de; friedrich@mathematik.uni-marburg.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_4

problem realizes dimension-independent convergence rates. For the treatment of the related inverse parameter identification problems we used regularization techniques. In particular, we analyzed and developed Tikhonov-regularization schemes with sparsity constraints for such nonlinear inverse problems. As a model problem, we studied the parameter identification problem for a parabolic reaction-diffusion system which describes the gene concentration in embryos at an early state of development (embryogenesis). In this chapter, we will only be concerned with the analysis of the forward problem, and we will concentrate on the construction of the new tensor wavelets and their approximation properties. For the analysis of the inverse problem, in particular concerning the regularity of the associated control-to-state map, we refer to Chap. 3 of this book. Numerical examples of the overall, fully adaptive wavelet scheme can be found in [5].

The approach we will present has partially been inspired by the work of Z. Ciesielski and T. Figiel [3] and of W. Dahmen and R. Schneider [6] who constructed a basis for a range of Sobolev spaces on a domain Ω from corresponding bases on subdomains. The principle idea can be described as follows. Let $\Omega = \cup_{k=0}^{N} \Omega_k \subset \mathbb{R}^n$ be a non-overlapping domain decomposition. By the use of extension operators, we will construct isomorphisms from the Cartesian product of Sobolev spaces on the subdomains, which incorporate suitable boundary conditions, to Sobolev spaces on Ω. By applying such an isomorphism to the union of Riesz bases for the Sobolev spaces on the subdomains, the result is a Riesz basis for the Sobolev space on Ω.

Since the approach can be applied recursively, to understand the construction of such an isomorphism, it is sufficient to consider the case of having two subdomains. For $i \in \{1, 2\}$, let R_i be the restriction of functions on Ω to Ω_i, let η_2 be the extension by zero of functions on Ω_2 to functions on Ω, and let E_1 be some extension of functions on Ω_1 to functions on Ω which, for some $m \in \mathbb{N}_0$, is bounded from $H^m(\Omega_1)$ to the target space $H^m(\Omega)$. Then $\begin{bmatrix} R_1 \\ R_2(\mathrm{Id} - E_1 R_1) \end{bmatrix} : H^m(\Omega) \to H^m(\Omega_1) \times H^m_{0,\partial\Omega_1 \cap \partial\Omega_2}(\Omega_2)$ is boundedly invertible with inverse $[E_1 \ \eta_2]$. ($H^m_{0,\partial\Omega_1 \cap \partial\Omega_2}(\Omega_2)$ is the space of $H^m(\Omega_2)$ functions that vanish up to order $m-1$ at $\partial\Omega_1 \cap \partial\Omega_2$). Consequently, if Ψ_1 is a Riesz basis for $H^m(\Omega_1)$ and Ψ_2 is a Riesz basis for $H^m_{0,\partial\Omega_1 \cap \partial\Omega_2}(\Omega_2)$, then $E_1\Psi_1 \cup \eta_2\Psi_2$ is a Riesz basis for $H^m(\Omega)$.

Our main interest in the construction of a basis from bases on subdomains lies in the use of *piecewise tensor product approximation*. On the hypercube $\square := (0,1)^n$ one can construct a basis for the Sobolev space $H^m(\square)$ (or for a subspace incorporating Dirichlet boundary conditions) by taking an n-fold tensor product of a collection of univariate functions that forms a Riesz basis for $L_2(0,1)$ as well as, properly scaled, for $H^m(0,1)$. Thinking of a univariate wavelet basis of order $d > m$, the advantage of this approach is that the rate of nonlinear best M-term approximation of a sufficiently smooth function u is $d - m$, compared to $\frac{d-m}{n}$ for standard best M-term isotropic (wavelet) approximation of order d on $\square$. The multiplication of the one-dimensional rate $d - m$ by the factor $\frac{1}{n}$ is commonly

referred to as the *curse of dimensionality*. However, when it comes to practical applications one should keep in mind that also the constants depend on n – even exponentially in the worst case. This is an intrinsic problem that also holds for other discretizations, e.g., by sparse grids. Nonetheless, tensor wavelets are a tool by which, for moderate space dimensions, the curse of dimensionality is at least diminished.

In view of these results on $\square$, we consider a domain Ω whose closure is the union of subdomains $\alpha_k + \overline{\square}$ for some $\alpha_k \in \mathbb{Z}^n$, or a domain Ω that is a parametric image of such a domain under a piecewise sufficiently smooth, globally C^{m-1} diffeomorphism κ, being a homeomorphism when $m = 1$. We equip $H^m(\Omega)$ (or a subspace incorporating Dirichlet boundary conditions) with a Riesz basis that is constructed using extension operators as discussed before from tensor product wavelet bases of order d on the subdomains, or from push-forwards of such bases. Many topological settings are covered by our approach, i.e., we consider homogeneous Dirichlet boundary conditions on arbitrary Lipschitz domains in two dimensions, see also Example 4.1 below. Our restriction to decompositions of Ω into subdomains from a topological Cartesian partition will allow us to rely on univariate extensions. Indeed, in order to end up with locally supported wavelets, we will apply local, scale-dependent extension operators – i.e., only wavelets that are adapted to the boundary conditions on the interfaces will be extended. We will show the best possible approximation rate $d - m$ for any u that restricted to any of these subdomains has a pull-back that belongs to a weighted Sobolev space.

4.2 Approximation by Tensor Product Wavelets on the Hypercube

We will study non-overlapping domain decompositions, where the subdomains are either unit n-cubes or smooth images of those. Sobolev spaces on these cubes, that appear with the construction of a Riesz basis for a Sobolev space on the domain as a whole, will be equipped with tensor product wavelet bases. From [7], we recall the construction of those bases.

For $t \in [0,\infty)\backslash(\mathbb{N}_0+\frac{1}{2})$ and $\boldsymbol{\sigma} = (\sigma_\ell, \sigma_r) \in \{0,\dots,\lfloor t+\frac{1}{2}\rfloor\}^2$, with $\mathscr{I} := (0,1)$, let

$$H^t_{\boldsymbol{\sigma}}(\mathscr{I}) := \{v \in H^t(\mathscr{I}) : v(0) = \cdots = v^{(\sigma_\ell-1)}(0) = 0 = v(1) = \cdots = v^{(\sigma_r-1)}(1)\}.$$

With t and $\boldsymbol{\sigma}$ as above, and for $\tilde{t} \in [0,\infty) \setminus (\mathbb{N}_0 + \frac{1}{2})$ and $\tilde{\boldsymbol{\sigma}} = (\tilde{\sigma}_\ell, \tilde{\sigma}_r) \in \{0,\dots,\lfloor \tilde{t} + \frac{1}{2}\rfloor\}^2$, we assume that

$$\Psi_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}} := \big\{\psi^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}_\lambda : \lambda \in \nabla_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\big\} \subset H^t_{\boldsymbol{\sigma}}(\mathscr{I}), \quad \tilde{\Psi}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}} := \{\tilde{\psi}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}_\lambda : \lambda \in \nabla_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\} \subset H^{\tilde{t}}_{\tilde{\boldsymbol{\sigma}}}(\mathscr{I})$$

are biorthogonal Riesz bases for $L_2(\mathscr{I})$, and, by rescaling, for $H^t_{\boldsymbol{\sigma}}(\mathscr{I})$ and $H^{\tilde{t}}_{\tilde{\boldsymbol{\sigma}}}(\mathscr{I})$, respectively. Furthermore, denoting by $|\lambda| \in \mathbb{N}_0$ the *level* of λ, we assume that for some

$$\mathbb{N} \ni d > t,$$

$\mathscr{W}_1$. $|\langle \tilde{\psi}_\lambda^{(\sigma,\tilde{\sigma})}, u\rangle_{L_2(\mathscr{I})}| \lesssim 2^{-|\lambda|d}\|u\|_{H^d(\operatorname{supp}\tilde{\psi}^{(\sigma,\tilde{\sigma})})} \quad (u \in H^d(\mathscr{I}) \cap H^t_{\boldsymbol{\sigma}}(\mathscr{I}))$,

$\mathscr{W}_2$. $\rho := \sup_{\lambda\in\nabla_{\sigma,\tilde{\sigma}}} 2^{|\lambda|}\max(\operatorname{diam}\operatorname{supp}\tilde{\psi}_\lambda^{(\sigma,\tilde{\sigma})}, \operatorname{diam}\operatorname{supp}\psi_\lambda^{(\sigma,\tilde{\sigma})})$
$\eqsim \inf_{\lambda\in\nabla_{\sigma,\tilde{\sigma}}} 2^{|\lambda|}\max(\operatorname{diam}\operatorname{supp}\tilde{\psi}_\lambda^{(\sigma,\tilde{\sigma})}, \operatorname{diam}\operatorname{supp}\psi_\lambda^{(\sigma,\tilde{\sigma})})$,

$\mathscr{W}_3$. $\sup_{j,k\in\mathbb{N}_0} \#\{|\lambda| = j : [k2^{-j}, (k+1)2^{-j}] \cap (\operatorname{supp}\tilde{\psi}_\lambda^{(\sigma,\tilde{\sigma})} \cup \operatorname{supp}\psi_\lambda^{(\sigma,\tilde{\sigma})}) \neq \emptyset\} < \infty$.

These conditions (as well as ($\mathscr{W}_4$) – ($\mathscr{W}_7$) in Sect. 4.4) are satisfied by following the biorthogonal wavelet constructions on the interval outlined in [8, 11].

It holds that $L_2(\square) = \otimes_{i=1}^n L_2(\mathscr{I})$. Further with

$$\boldsymbol{\sigma} = (\boldsymbol{\sigma}_i = ((\sigma_i)_\ell, (\sigma_i)_r))_{1\le i\le n} \in (\{0, \dots, \lfloor t + \tfrac{1}{2}\rfloor\}^2)^n,$$

we define

$$H^t_{\boldsymbol{\sigma}}(\square) := H^t_{\boldsymbol{\sigma}_1}(\mathscr{I})\otimes L_2(\mathscr{I})\otimes\cdots\otimes L_2(\mathscr{I}) \cap \cdots \cap L_2(\mathscr{I})\otimes\cdots\otimes L_2(\mathscr{I})\otimes H^t_{\boldsymbol{\sigma}_n}(\mathscr{I}),$$

which is the space of $H^t(\square)$-functions whose normal derivatives of up to orders $(\sigma_i)_\ell$ and $(\sigma_i)_r$ vanish at the facets $\overline{\mathscr{I}}^{i-1} \times \{0\} \times \overline{\mathscr{I}}^{n-i}$ and $\overline{\mathscr{I}}^{i-1} \times \{1\} \times \overline{\mathscr{I}}^{n-i}$, respectively ($1 \le i \le n$) (the proof of this fact given in [7] for $t \in \mathbb{N}_0$ can be generalized to $t \in [0,\infty) \setminus (\mathbb{N}_0 + \frac{1}{2})$).

The *tensor product wavelet* collection

$$\boldsymbol{\Psi}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}} := \otimes_{i=1}^n \Psi_{\boldsymbol{\sigma}_i,\tilde{\boldsymbol{\sigma}}_i} = \{\boldsymbol{\psi}_{\boldsymbol{\lambda}}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} := \otimes_{i=1}^n \psi_{\lambda_i}^{(\boldsymbol{\sigma}_i,\tilde{\boldsymbol{\sigma}}_i)} : \boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}} := \prod_{i=1}^n \nabla_{\boldsymbol{\sigma}_i,\tilde{\boldsymbol{\sigma}}_i}\},$$

and its renormalized version $\{\big(\sum_{i=1}^n 4^{t|\lambda_i|}\big)^{-1/2}\boldsymbol{\psi}_{\boldsymbol{\lambda}}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} : \boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\}$ are Riesz bases for $L_2(\square)$ and $H^t_{\boldsymbol{\sigma}}(\square)$, respectively. The collection that is dual to $\boldsymbol{\Psi}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}$ reads as

$$\tilde{\boldsymbol{\Psi}}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}} := \otimes_{i=1}^n \tilde{\Psi}_{\boldsymbol{\sigma}_i,\tilde{\boldsymbol{\sigma}}_i} = \{\tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda}}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} := \otimes_{i=1}^n \tilde{\psi}_{\lambda_i}^{(\boldsymbol{\sigma}_i,\tilde{\boldsymbol{\sigma}}_i)} : \boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\},$$

and its renormalized version $\{\big(\sum_{i=1}^n 4^{|\lambda_i|}\big)^{-\tilde{t}/2}\tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda}}^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} : \boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\}$ is a Riesz basis for $H^{\tilde{t}}_{\tilde{\boldsymbol{\sigma}}}(\square)$.

For $\boldsymbol{\lambda} \in \boldsymbol{\nabla}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}$, we set $|\boldsymbol{\lambda}| := (|\lambda_1|, \dots, |\lambda_n|)$.

For $\theta \ge 0$, the *weighted Sobolev space* $\mathscr{H}^d_\theta(\mathscr{I})$ is defined as the space of all measurable functions u on $\mathscr{I}$ for which the norm

$$\|u\|_{\mathscr{H}^d_\theta(\mathscr{I})} := \left[\sum_{j=0}^{d} \int_{\mathscr{I}} |x^\theta (1-x)^\theta u^{(j)}(x)|^2 \,\mathrm{d}x\right]^{\frac{1}{2}}$$

is finite. For $m \in \{0, \ldots, \lfloor t \rfloor\}$, we will consider the weighted Sobolev space

$$\mathscr{H}^d_{m,\theta}(\square) := \cap_{p=1}^n \otimes_{i=1}^n \mathscr{H}^d_{\theta-\delta_{ip}\min(m,\theta)}(\mathscr{I}),$$

equipped with a squared norm that is the sum over $p = 1, \ldots, n$ of the squared norms on $\otimes_{i=1}^n \mathscr{H}^d_{\theta-\delta_{ip}\min(m,\theta)}(\mathscr{I})$.

4.3 Construction of Riesz Bases by Extension

Let $\{\square_0, \ldots, \square_N\}$ be a set of hypercubes from $\{\tau + \square : \tau \in \mathbb{Z}^n\}$, and let $\hat{\Omega}$ be a (reference) domain (i.e., open and connected) in $\mathbb{R}^n$ with $\cup_{k=0}^N \square_k \subset \hat{\Omega} \subset (\cup_{k=0}^N \overline{\square}_k)^{\mathrm{int}}$, and such that $\partial\hat{\Omega}$ is the union of (closed) facets of the $\square_k$'s. The case $\hat{\Omega} \subsetneq (\cup_{k=0}^N \overline{\square}_k)^{\mathrm{int}}$ corresponds to the situation that $\hat{\Omega}$ has one or more cracks. We will describe a construction of Riesz bases for Sobolev spaces on $\hat{\Omega}$ from Riesz bases for corresponding Sobolev spaces on the subdomains $\square_k$ using extension operators. We start with giving sufficient conditions ($\mathscr{D}_1$)–($\mathscr{D}_5$) such that suitable extension operators exist.

We assume that there exists a sequence $(\{\hat{\Omega}_k^{(q)} : q \le k \le N\})_{0\le q\le N}$ of sets of polytopes, such that $\hat{\Omega}_k^{(0)} = \square_k$ and where each next term in the sequence is created from its predecessor by joining two of its polytopes. More precisely, we assume that for any $1 \le q \le N$, there exists a $q \le \bar{k} = \bar{k}^{(q)} \le N$ and $q-1 \le k_1 = k_1^{(q)} \ne k_2 = k_2^{(q)} \le N$ such that

$\mathscr{D}_1$. $\hat{\Omega}_{\bar{k}}^{(q)} = \left(\overline{\hat{\Omega}_{k_1}^{(q-1)} \cup \hat{\Omega}_{k_2}^{(q-1)}} \setminus \partial\hat{\Omega}\right)^{\mathrm{int}}$ is connected, and the interface $J := \hat{\Omega}_{\bar{k}}^{(q)} \setminus (\hat{\Omega}_{k_1}^{(q-1)} \cup \hat{\Omega}_{k_2}^{(q-1)})$ is part of a hyperplane,

$\mathscr{D}_2$. $\{\hat{\Omega}_k^{(q)} : q \le k \le N, k \ne \bar{k}\} = \{\hat{\Omega}_k^{(q-1)} : q-1 \le k \le N, k \ne \{k_1, k_2\}\}$,

$\mathscr{D}_3$. $\hat{\Omega}_N^{(N)} = \hat{\Omega}$.

For some

$$t \in [0, \infty) \setminus (\mathbb{N}_0 + \tfrac{1}{2}),$$

to each of the *closed* facets of all the hypercubes $\square_k$, we associate a number in $\{0, \ldots, \lfloor t + \frac{1}{2} \rfloor\}$ indicating the order of the Dirichlet boundary condition on that facet (where a Dirichlet boundary condition of order 0 means no boundary condition). On facets on the boundary of $\hat{\Omega}$, this number can be chosen at one's convenience (it is dictated by the boundary conditions of the boundary value problem that one wants

to solve on $\hat{\Omega}$), and, as will follow from the conditions imposed below, on the other facets it should be either 0 or $\lfloor t + \frac{1}{2} \rfloor$.

By construction, each facet of any $\hat{\Omega}_k^{(q)}$ is a union of some facets of the $\square_{k'}$'s, that will be referred to as subfacets. Letting each of these subfacets inherit the Dirichlet boundary conditions imposed on the $\square_{k'}$'s, we define

$$\mathring{H}^t(\hat{\Omega}_k^{(q)}),$$

and so for $k = q = N$ in particular $\mathring{H}^t(\hat{\Omega}) = \mathring{H}^t(\hat{\Omega}_N^{(N)})$, to be the closure in $H^t(\hat{\Omega}_k^{(q)})$ of the smooth functions on $\hat{\Omega}_k^{(q)}$ that satisfy these boundary conditions. Note that for $0 \leq k \leq N$, for some $\boldsymbol{\sigma}(k) \in (\{0, \ldots, \lfloor t + \frac{1}{2} \rfloor\}^2)^n$,

$$\mathring{H}^t(\hat{\Omega}_k^{(0)}) = \mathring{H}^t(\square_k) = H^t_{\boldsymbol{\sigma}(k)}(\square_k).$$

The boundary conditions on the hypercubes that determine the spaces $\mathring{H}^t(\hat{\Omega}_k^{(q)})$, and the order in which polytopes are joined should be chosen such that

$\mathscr{D}_4$. on the $\hat{\Omega}_{k_1}^{(q-1)}$ and $\hat{\Omega}_{k_2}^{(q-1)}$ sides of J, the boundary conditions are of order 0 and $\lfloor t + \frac{1}{2} \rfloor$, respectively,

and, w.l.o.g. assuming that $J = \{0\} \times \check{J}$ and $(0, 1) \times \check{J} \subset \Omega_{k_1}^{(q-1)}$,

$\mathscr{D}_5$. for any function in $\mathring{H}^t(\hat{\Omega}_{k_1}^{(q-1)})$ that vanishes near $\{0, 1\} \times \check{J}$, its reflection in $\{0\} \times \mathbb{R}^{n-1}$ (extended with zero, and then restricted to $\hat{\Omega}_{k_2}^{(q-1)}$) is in $\mathring{H}^t(\hat{\Omega}_{k_2}^{(q-1)})$.

The condition ($\mathscr{D}_5$) is a compatibility condition on the subfacets adjacent to the interface, see Fig. 4.1 for an illustration.

Given $1 \leq q \leq N$, for $i \in \{1, 2\}$, let $R_i^{(q)}$ be the *restriction* of functions on $\hat{\Omega}_{\bar{k}}^{(q)}$ to $\hat{\Omega}_{k_i}^{(q-1)}$, and let $\eta_2^{(q)}$ be the *extension* of functions on $\hat{\Omega}_{k_2}^{(q-1)}$ to $\hat{\Omega}_{\bar{k}}^{(q)}$ by zero. Under the conditions ($\mathscr{D}_1$)–($\mathscr{D}_5$), the extensions $E_1^{(q)}$ of functions on $\hat{\Omega}_{k_1}^{(q-1)}$ to $\hat{\Omega}_{\bar{k}}^{(q)}$ can be constructed (essentially) as tensor products of *univariate extensions* with identity operators in the other Cartesian directions. In the remaining part of this

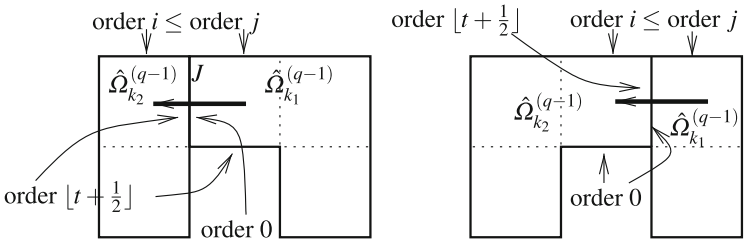


Fig. 4.1 Two illustrations with ($\mathscr{D}_1$)–($\mathscr{D}_5$). The *fat arrow* indicates the action of the extension $E_1^{(q)}$

chapter $[\cdot,\cdot]_{s,2}$ denotes the real interpolation space between two Hilbert spaces. For further information we refer to [1].

Proposition 4.1 ([2, Prop. 4.4.]). *Let G_1 be an extension operator of functions on $(0,1)$ to functions on $(-1,1)$ such that*

$$G_1 \in B(L_2(0,1), L_2(-1,1)), \quad G_1 \in B(H^t(0,1), H^t_{(\lfloor t+\frac{1}{2}\rfloor,0)}(-1,1)).$$

Let $E_1^{(q)}$ be defined by $R_2^{(q)} E_1^{(q)}$ being the composition of the restriction to $(0,1)\times \breve{J}$, followed by an application of

$$G_1 \otimes \mathrm{Id} \otimes \cdots \otimes \mathrm{Id},$$

followed by an extension by 0 to $\hat{\Omega}_{k_2}^{(q-1)} \setminus (-1,0)\times \breve{J}$. Then for $s \in [0,1]$

$$E^{(q)} := [E_1^{(q)}\ \eta_2^{(q)}] \in B\Big(\prod_{i=1}^{2} [L_2(\hat{\Omega}_{k_i}^{(q-1)}), \mathring{H}^t(\hat{\Omega}_{k_i}^{(q-1)})]_{s,2}, [L_2(\hat{\Omega}_{\tilde{k}}^{(q)}), \mathring{H}^t(\hat{\Omega}_{\tilde{k}}^{(q)})]_{s,2}\Big) \tag{4.1}$$

is boundedly invertible.

A Riesz basis on $\hat{\Omega}$ can now be constructed as follows.

Corollary 4.1 ([2, Cor. 4.6]). *For $0 \le k \le N$, let $\boldsymbol{\Psi}_k$ be a Riesz basis for $L_2(\square_k)$, that renormalized in $H^t(\square_k)$, is a Riesz basis for $\mathring{H}^t(\square_k) = H^t_{\boldsymbol{\sigma}(k)}(\square)$. Let E be the composition for $q = 1,\dots,N$ of the mappings $E^{(q)}$ defined in (4.1), trivially extended with identity operators in coordinates $k \in \{q-1,\dots,N\} \setminus \{k_1^{(q)}, k_2^{(q)}\}$. Then it holds that*

$$E \in B\Big(\prod_{k=0}^{n} [L_2(\square_k), \mathring{H}^t(\square_k)]_{s,2}, [L_2(\hat{\Omega}), \mathring{H}^t(\hat{\Omega})]_{s,2}\Big) \tag{4.2}$$

is boundedly invertible. Further, for $s \in [0,1]$, the collection $E(\prod_{k=0}^{N} \boldsymbol{\Psi}_k)$, normalized in the corresponding norm, is a Riesz basis for $[L_2(\hat{\Omega}), \mathring{H}^t(\hat{\Omega})]_{s,2}$.

For the dual basis $E^{-*}(\prod_{k=0}^{N} \tilde{\boldsymbol{\Psi}}_k)$ a similar result holds. In particular, for $s \in [0,1]$, it is, properly scaled, a Riesz basis for $[L_2(\hat{\Omega}), \mathring{H}^{\tilde{t}}(\hat{\Omega})]_{s,2}$. We refer to [2] for a detailed presentation.

The construction of Riesz bases on the reference domain $\hat{\Omega}$ extends to more general domains in a standard fashion. Let Ω be the image of $\hat{\Omega}$ under a homeomorphism κ. We define the *pull-back* κ^* by $\kappa^* w = w \circ \kappa$, and so its inverse κ^{-*}, known as the *push-forward*, satisfies $\kappa^{-*} v = v \circ \kappa^{-1}$.

Proposition 4.2 ([2, Prop. 4.11.]). *Let κ^* be boundedly invertible as a mapping both from $L_2(\Omega)$ to $L_2(\hat{\Omega})$ and from $H^t(\Omega)$ to $H^t(\hat{\Omega})$. Setting $\mathring{H}^t(\Omega) := \Im \kappa^{-*}|_{\mathring{H}^t(\hat{\Omega})}$, we have that $\kappa^{-*} \in B([L_2(\hat{\Omega}), \mathring{H}^t(\hat{\Omega})]_{s,2}, [L_2(\Omega), \mathring{H}^t(\Omega)]_{s,2})$ is*

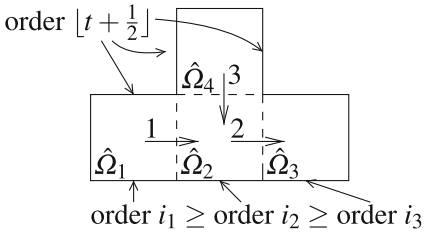


Fig. 4.2 Example extension directions and compatible boundary conditions

boundedly invertible ($s \in [0,1]$). *So if* $\boldsymbol{\Psi}$ *is a Riesz basis for* $L_2(\hat{\Omega})$ *and, properly scaled, for* $\mathring{H}^t(\hat{\Omega})$, *then for* $s \in [0,1]$, $\kappa^{-*}\boldsymbol{\Psi}$ *is, properly scaled, a Riesz basis for* $[L_2(\Omega), \mathring{H}^t(\Omega)]_{s,2}$. *If* $\tilde{\boldsymbol{\Psi}}$ *is the collection dual to* $\boldsymbol{\Psi}$, *then* $|\det D\kappa^{-1}(\cdot)|\kappa^{-*}\tilde{\boldsymbol{\Psi}}$ *is the collection dual to* $\kappa^{-*}\boldsymbol{\Psi}$.

We conclude this section by discussing some of the topological aspects of the construction.

Example 4.1. Consider a T-shaped domain decomposed into four subcubes as depicted in Fig. 4.2. In such a setting it is not possible to arrange the subdomains in a linear fashion. Further, when constructing a basis on such a domain, the ordering and directions of the extension operators are not unique. However, both aspects influence the boundary conditions that may be imposed. When proceeding as depicted, in the first step wavelets on $\hat{\Omega}_1$ are extended to $\hat{\Omega}_2$. Then the resulting basis is extended to $\hat{\Omega}_3$. Finally wavelets on $\hat{\Omega}_4$ are extended along the bottom interface. This set of extensions and its ordering is compatible with all boundary conditions that satisfy the restrictions depicted in Fig. 4.2, e.g., with homogeneous Dirichlet boundary conditions. In the second step of the construction only tensor wavelets on $\hat{\Omega}_2$ are extended. Consequently interchanging the ordering of the first two extensions does not change the resulting basis.

4.4 Approximation by – Piecewise – Tensor Product Wavelets

In the setting of Corollary 4.1 we select the bases on the subdomains $\square_k = \square + \alpha_k$, $\alpha_k \in \mathbb{Z}^n$, to be $\boldsymbol{\Psi}_{\boldsymbol{\sigma}(k),\tilde{\boldsymbol{\sigma}}(k)}(\cdot - \alpha_k)$, $\tilde{\boldsymbol{\Psi}}_{\boldsymbol{\sigma}(k),\tilde{\boldsymbol{\sigma}}(k)}(\cdot - \alpha_k)$, as constructed in Sect. 4.2. In this setting, for $m \in \{0, \dots, \lfloor t \rfloor\}$ we study the approximation of functions $u \in \mathring{H}^m(\Omega) := [L_2(\Omega), \mathring{H}^t(\Omega)]_{m/t,2}$, that also satisfy

$$u \in \kappa^{-*}\Big(\prod_{k=0}^{N} \mathscr{H}^d_{m,\theta}(\square_k)\Big) := \{v : \Omega \to \mathbb{R} : v \circ \kappa \in \prod_{k=0}^{N} \mathscr{H}^d_{m,\theta}(\square_k)\}, \tag{4.3}$$

by $\kappa^{-*}E\big(\prod_{k=0}^{N} \boldsymbol{\Psi}_{\boldsymbol{\sigma}(k),\tilde{\boldsymbol{\sigma}}(k)}(\cdot - \alpha_k)\big)$ in the $H^m(\Omega)$-norm. Since, as is assumed in Proposition 4.2, $\kappa^* \in B(\mathring{H}^m(\Omega), \mathring{H}^m(\hat{\Omega}))$ is boundedly invertible, it is sufficient to study this issue for the case that $\kappa = \mathrm{Id}$ and so $\Omega = \hat{\Omega}$.

We will apply extension operators $E_1^{(q)}$ that are built from univariate extension operators. The latter will be chosen such that the resulting primal and dual wavelets on $\hat{\Omega}$, restricted to each $\square_k \subset \hat{\Omega}$, are tensor products of collections of univariate functions. We make the following additional assumptions on the univariate wavelets. For $\boldsymbol{\sigma} = (\sigma_\ell, \sigma_r) \in \{0, \dots, \lfloor t + \frac{1}{2} \rfloor\}^2$, $\tilde{\boldsymbol{\sigma}} = (\tilde{\sigma}_\ell, \tilde{\sigma}_r) \in \{0, \dots, \lfloor \tilde{t} + \frac{1}{2} \rfloor\}^2$, and with $\mathbf{0} := (0, 0)$,

$\mathscr{W}_4$. $V_j^{(\boldsymbol{\sigma})} := \operatorname{span}\{\psi_\lambda^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} : \lambda \in \nabla_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}, |\lambda| \le j\}$ is independent of $\tilde{\boldsymbol{\sigma}}$, and $V_j^{(\boldsymbol{\sigma})} = V_j^{(\mathbf{0})} \cap H^t_{\boldsymbol{\sigma}}(\mathscr{I})$,

$\mathscr{W}_5$. $\nabla_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}$ is the disjoint union of $\nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}$, $\nabla^{(I)}$, $\nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}$ such that

i. $$\sup_{\lambda \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell},\, x \in \operatorname{supp}\psi_\lambda^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}} 2^{|\lambda|}|x| \lesssim \rho, \qquad \sup_{\lambda \in \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r},\, x \in \operatorname{supp}\psi_\lambda^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}} 2^{|\lambda|}|1 - x| \lesssim \rho,$$

ii. For $\lambda \in \nabla^{(I)}$, $\psi_\lambda^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} = \psi_\lambda^{(\mathbf{0},\mathbf{0})}$, $\tilde{\psi}_\lambda^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} = \tilde{\psi}_\lambda^{(\mathbf{0},\mathbf{0})}$, and the extensions of $\psi_\lambda^{(\mathbf{0},\mathbf{0})}$ and $\tilde{\psi}_\lambda^{(\mathbf{0},\mathbf{0})}$ by zero are in $H^t(\mathbb{R})$ and $H^{\tilde{t}}(\mathbb{R})$, respectively,

$\mathscr{W}_6$.
$$\begin{cases} \operatorname{span}\{\psi_\lambda^{(\mathbf{0},\mathbf{0})}(1-\cdot) : \lambda \in \nabla^{(I)},\ |\lambda| = j\} = \operatorname{span}\{\psi_\lambda^{(\mathbf{0},\mathbf{0})} : \lambda \in \nabla^{(I)},\ |\lambda| = j\}, \\ \operatorname{span}\{\psi_\lambda^{(\sigma_\ell,\sigma_r),(\tilde{\sigma}_\ell,\tilde{\sigma}_r)}(1-\cdot) : \lambda \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell},\ |\lambda| = j\} = \\ \qquad\qquad \operatorname{span}\{\psi_\lambda^{(\sigma_r,\sigma_\ell),(\tilde{\sigma}_r,\tilde{\sigma}_\ell)} : \lambda \in \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r},\ |\lambda| = j\}, \end{cases}$$

$\mathscr{W}_7$.
$$\begin{cases} \psi_\lambda^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}(2^l\cdot) \in \operatorname{span}\{\psi_\mu^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} : \mu \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}\} \quad (l \in \mathbb{N}_0,\ \lambda \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}), \\ \psi_\lambda^{(\mathbf{0},\mathbf{0})}(2^l\cdot) \in \operatorname{span}\{\psi_\mu^{(\mathbf{0},\mathbf{0})} : \mu \in \nabla^{(I)}\} \quad (l \in \mathbb{N}_0,\ \lambda \in \nabla^{(I)}). \end{cases}$$

In the setting of Proposition 4.1 we choose the univariate extension operator to be a Hestenes extension [6, 9, 10], that is,

$$\breve{G}_1 v(-x) = \sum_{l=0}^{L} \gamma_l (\zeta v)(\beta_l x) \quad (v \in L_2(\mathscr{I}),\ x \in \mathscr{I}), \tag{4.4}$$

(and, being an extension, $\breve{G}_1 v(x) = v(x)$ for $x \in \mathscr{I}$), where $\gamma_l \in \mathbb{R}$, $\beta_l > 0$, and $\zeta : [0, \infty) \to [0, \infty)$ is some smooth cut-off function with $\zeta \equiv 1$ in a neighborhood of 0, and $\operatorname{supp}\zeta \subset [0, \min_l(\beta_l, \beta_l^{-1})]$.

With such an extension operator at hand the obvious approach is to define $E_1^{(q)}$ according to Proposition 4.1 with $G_1 = \breve{G}_1$. A problem with the choice $G_1 = \breve{G}_1$ is that generally the desirable property $\operatorname{diam}(\operatorname{supp} G_1 u) \lesssim \operatorname{diam}(\operatorname{supp} u)$ is not implied. Indeed, think of the application of a Hestenes extension to a u with a small support that is not located near the interface.

To solve this and the corresponding problem for the adjoint extension, following [6] we will apply our construction using the modified, *scale-dependent* univariate extension operator

$$G_1 : u \mapsto \sum_{\lambda \in \nabla^{(\ell)}_{0,0}} \langle u, \tilde{\psi}_\lambda^{(\mathbf{0},\mathbf{0})} \rangle_{L_2(\mathscr{I})} \breve{G}_1 \psi_\lambda^{(\mathbf{0},\mathbf{0})} + \sum_{\lambda \in \nabla^{(I)} \cup \nabla^{(r)}_{0,0}} \langle u, \tilde{\psi}_\lambda^{(\mathbf{0},\mathbf{0})} \rangle_{L_2(\mathscr{I})} \eta_1 \psi_\lambda^{(\mathbf{0},\mathbf{0})}. \tag{4.5}$$

We focus on univariate extension operators with $\beta_l = 2^l$. This, together with ($\mathscr{W}_7$) ensures that the extended wavelets are locally (weighted sums of) univariate wavelets. Consequently most properties, like the locality on the primal and dual side in the sense of ($\mathscr{W}_2$), and ($\mathscr{W}_3$), as well as piecewise Sobolev smoothness are inherited by the extended wavelets. Further, by the symmetry assumption ($\mathscr{W}_6$), the extended part of a wavelet belongs to the span of boundary adapted wavelets. Therefore, together with ($\mathscr{W}_4$), we derive the technically useful property that extended wavelets $G_1\psi_\mu^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}$ belong piecewise to spaces $V_j^{(\mathbf{0})}$ with additionally $j \le |\mu| + 2L$. This property, together with the locality of the primal and dual wavelets, is key for our central approximation result in Theorem 4.1.

Proposition 4.3 ([2, Prop. 5.2]). *Assuming ρ to be sufficiently small, the scale-dependent extension G_1 from (4.5) satisfies, for $\boldsymbol{\sigma} \in \{0,\dots,\lfloor t+\frac{1}{2}\rfloor\}^2$, $\tilde{\boldsymbol{\sigma}} \in \{0,\dots,\lfloor \tilde{t}+\frac{1}{2}\rfloor\}^2$*

$$G_1\psi_\mu^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} = \begin{cases} \eta_1\psi_\mu^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} & \text{when } \mu \in \nabla^{(I)} \cup \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}, \\ \breve{G}_1\psi_\mu^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})} & \text{when } \mu \in \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}. \end{cases} \tag{4.6}$$

Assuming, additionally, $\breve{G}_1$ to be a Hestenes extension *with $\beta_l = 2^l$, the resulting adjoint extension $G_2 := (\mathrm{Id} - \eta_1 G_1^*)\eta_2$ satisfies*

$$G_2(\tilde{\psi}_\mu^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}(1+\cdot)) = \begin{cases} \eta_2(\tilde{\psi}_\mu^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}(1+\cdot)) & \text{when } \mu \in \nabla^{(I)} \cup \nabla^{(\ell)}_{\sigma_\ell,\tilde{\sigma}_\ell}, \\ \breve{G}_2(\tilde{\psi}_\mu^{(\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}})}(1+\cdot)) & \text{when } \mu \in \nabla^{(r)}_{\sigma_r,\tilde{\sigma}_r}. \end{cases} \tag{4.7}$$

We have $G_1 \in B(L_2(0,1), L_2(-1,1))$, $G_1 \in B(H^t(0,1), H^t(-1,1))$, and further $G_1^* \in B(H^{\tilde{t}}(-1,1), H^{\tilde{t}}_{(\lfloor \tilde{t}+\frac{1}{2}\rfloor,0)}(0,1))$. Finally, G_1 and G_2 are local in the following sense

$$\begin{cases} \operatorname{diam}(\operatorname{supp} R_2 G_1 u) \lesssim \operatorname{diam}(\operatorname{supp} u) & (u \in L_2(0,1)), \\ \operatorname{diam}(\operatorname{supp} R_1 G_2 u) \lesssim \operatorname{diam}(\operatorname{supp} u) & (u \in L_2(-1,0)). \end{cases} \tag{4.8}$$

A typical example of a Hestenes extension with $\beta_l = 2^l$ is the reflection, i.e., $L = 0, \gamma_0 = 1$.

Remark 4.1. Although implicitly claimed otherwise in [6, (4.3.12)], we note that (4.7), and so the second property in (4.8), *cannot* be expected for $\breve{G}_1$ being a general Hestenes extension as given by (4.4), without assuming that $\beta_l = 2^l$.

We may now formulate the central approximation result. Recall that by utilizing the scale-dependent extension operator in the construction presented in Sect. 4.3, we end up with a pair of biorthogonal wavelet Riesz bases

$$\big(E(\prod_{k=0}^{N} \boldsymbol{\Psi}_k), \quad E^{-*}(\prod_{k=0}^{N} \tilde{\boldsymbol{\Psi}}_k)\big) = \big(\{\boldsymbol{\psi}_{\boldsymbol{\lambda},p} : (\boldsymbol{\lambda},p) \in \nabla(\hat{\Omega})\}, \{\tilde{\boldsymbol{\psi}}_{\boldsymbol{\lambda},p} : (\boldsymbol{\lambda},p) \in \nabla(\hat{\Omega})\}\big)$$

for $L_2(\hat{\Omega})$, that is for $s \in [0,1]$ and properly scaled a pair of Riesz bases for $[L_2(\hat{\Omega}), \mathring{H}^t(\hat{\Omega})]_{s,2}$ and $[L_2(\hat{\Omega}), \mathring{H}^{\tilde{t}}(\hat{\Omega})]_{s,2}$, respectively. In particular the index set is given by $\nabla(\hat{\Omega}) = \bigcup_{k=0}^{N} \nabla_{\sigma(k),\tilde{\sigma}(k)} \times \{k\}$.

Theorem 4.1 ([2, Thm. 5.6]). *Let the $E_1^{(q)}$ be defined using the scale-dependent extension operators as in Proposition 4.3. Then for any $\theta \in [0,d)$, there exists a (nested) sequence $(\nabla_M)_{M\in\mathbb{N}} \subset \nabla(\hat{\Omega})$ with $\#\nabla_M \eqsim M$, such that*

$$\inf_{v\in\operatorname{span}\{\psi_{\lambda,p}:(\lambda,p)\in\nabla_M\}} \|u-v\|_{H^m(\hat{\Omega})} \lesssim M^{-(d-m)} \sqrt{\sum_{k=0}^{N} \|u\|^2_{\mathscr{H}^d_{m,\theta}(\square_k)}}, \tag{4.9}$$

for any $u \in \mathring{H}^m(\hat{\Omega})$ for which the right-hand side is finite, i.e., that satisfies (4.3) (with $\kappa = \mathrm{Id}$). For $m = 0$, the factor $M^{-(d-m)}$ in (4.9) has to be read as $(\log M)^{(n-1)(\frac{1}{2}+d)} M^{-d}$.

The issue whether we may expect (4.3) for u to hold is nontrivial. Fortunately, in [2], we were able to prove that this property holds for the solutions of a large class of boundary value problems over polygonal or polyhedral domains.

4.5 Numerical Results

As domains, we consider the *slit domain* $\Omega = (0,2)^2 \backslash \{1\} \times [1,2)$, the 3-dimensional *L-shaped domain* $\Omega = (0,2)^2 \times (0,1) \setminus [1,2)^2 \times (0,1)$, and the *Fichera corner domain* $\Omega = (0,2)^3 \setminus [1,2)^3$. The corresponding domain decompositions and the directions in which the extension operator is applied are illustrated in Fig. 4.3.

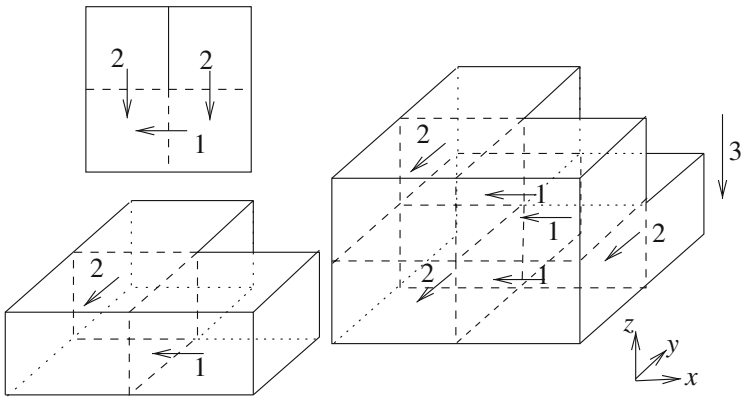


Fig. 4.3 The direction and ordering of the extensions

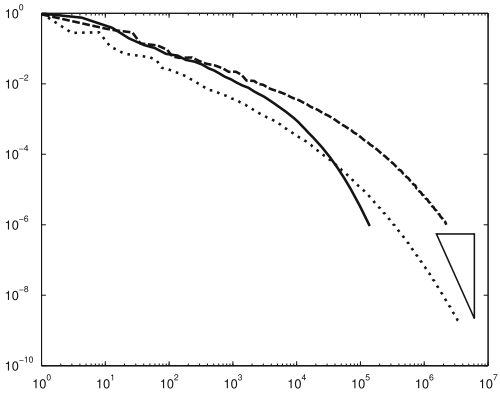


Fig. 4.4 Support length vs. relative residual on the slit domain (*line*), L-shaped domain (*dotted*) and Fichera corner domain (*dashed*)

As extension operator, we apply the reflection suited for $\frac{1}{2} < t < \frac{3}{2}$, $0 < \tilde{t} < \frac{1}{2}$, which is sufficient for our aim of constructing a Riesz basis for $H_0^1(\Omega)$.

Using piecewise tensor product bases, we solved the problem of finding $u \in H_0^1(\Omega)$ such that

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = f(v) \quad (v \in H_0^1(\Omega)) \tag{4.10}$$

by applying the *adaptive wavelet-Galerkin method* [4, 12]. We choose the forcing vector $f = 1$.

As the univariate bases for the tensor wavelet construction we choose C^1, piecewise quartic ($\mathrm{d} = 5$) (multi-) wavelets. The chosen solver is known to produce a sequence of approximations that converges in the $H^1(\Omega)$-norm with the same rate as best M-term wavelet approximation. We therefore expect the approximation rate $d - m = 5 - 1 = 4$.

The numerical results are presented in Fig. 4.4.

References

1. Bergh, J., Löfström, J.: Interpolation Spaces. An Introduction. Grundlehren der mathematischen Wissenschaften, vol. 223. Springer, Berlin (1976)
2. Chegini, N., Dahlke, S., Friedrich, U., Stevenson, R.: Piecewise tensor product wavelet bases by extensions and approximation rates. Math. Comput. **82**(284), 2157–2190 (2013)
3. Ciesielski, Z., Figiel, T.: Spline bases in classical function spaces on compact C^∞ manifolds. I and II. Stud. Math. **76**(2), 1–58, 95–136 (1983)
4. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods for elliptic operator equations: convergence rates. Math. Comput. **70**(233), 27–75 (2001)
5. Dahlke, S., Friedrich, U., Maass, P., Raasch, T., Ressel, R.A.: An adaptive wavelet solver for a nonlinear parameter identification problem for a parabolic differential equation with sparsity constraints. J. Inverse Ill-Posed Probl. **20**(2), 213–251 (2012)

6. Dahmen, W., Schneider, R.: Wavelets on manifolds. I: construction and domain decomposition. SIAM J. Math. Anal. **31**(1), 184–230 (1999)
7. Dauge, M., Stevenson, R.: Sparse tensor product wavelet approximation of singular functions. SIAM J. Math. Anal. **42**(5), 2203–2228 (2010)
8. Dijkema, T.: Adaptive tensor product wavelet methods for solving pdes. Ph.D. thesis, Utrecht University (2009)
9. Hestenes, M.: Extension of the range of a differentiable function. Duke Math. J. **8**, 183–192 (1941)
10. Kunoth, A., Sahner, J.: Wavelets on manifolds: an optimized construction. Math. Comput. **75**(255), 1319–1349 (2006)
11. Primbs, M.: New stable biorthogonal spline-wavelets on the interval. Results Math. **57**(1–2), 121–162 (2010)
12. Stevenson, R.: Adaptive wavelet methods for solving operator equations: an overview. In: Multiscale, Nonlinear and Adaptive Approximation. Dedicated to Wolfgang Dahmen on the Occasion of his 60th Birthday, pp. 543–597. Springer, Berlin (2009)

Chapter 5
Adaptive Wavelet Methods for SPDEs

Petru A. Cioica, Stephan Dahlke, Nicolas Döhring, Stefan Kinzel, Felix Lindner, Thorsten Raasch, Klaus Ritter, and René L. Schilling

Abstract We review a series of results that have been obtained in the context of the DFG-SPP 1324 project "Adaptive wavelet methods for SPDEs". This project has been concerned with the construction and analysis of adaptive wavelet methods for second order parabolic stochastic partial differential equations on bounded, possibly nonsmooth domains $\mathcal{O} \subset \mathbb{R}^d$. A detailed regularity analysis for the solution process u in the scale of Besov spaces $B^s_{\tau,\tau}(\mathcal{O})$, $1/\tau = s/d + 1/p$, $\alpha > 0$, $p \geq 2$, is presented. The regularity in this scale is known to determine the order of convergence that can be achieved by adaptive wavelet algorithms and other nonlinear approximation schemes. As it turns out, in general, for solutions of SPDEs this regularity exceeds the $L_p(\mathcal{O})$-Sobolev regularity, which determines the order of convergence for uniform approximation schemes. We also study nonlinear wavelet approximation of elliptic boundary value problems on $\mathcal{O}$ with random right-hand side. Such problems appear naturally when applying Rothe's method to the parabolic stochastic equation. A general stochastic wavelet model for the right-hand side is introduced and its Besov regularity as well as linear and nonlinear approximation is studied. The results are matched by computational experiments.

R.L. Schilling (✉)
Technical University of Dresden, Zellescher Weg 12–14, 01069 Dresden, Germany
e-mail: rene.schilling@tu-dresden.de

N. Döhring • F. Lindner • K. Ritter
Technical University of Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany
e-mail: doehring@mathematik.uni-kl.de; lindner@mathematik.uni-kl.de; ritter@mathematik.uni-kl.de

T. Raasch
University of Mainz, Staudingerweg 9, 55099 Mainz, Germany
e-mail: raasch@uni-mainz.de

P.A. Cioica • S. Dahlke • S. Kinzel
Philipps-University of Marburg, Hans Meerwein Str., Lahnberge, 35032 Marburg, Germany
e-mail: cioica@mathematik.uni-marburg.de; dahlke@mathematik.uni-marburg.de; kinzel@mathematik.uni-marburg.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_5

5.1 Introduction

In this paper we survey a series of results towards the construction and analysis of adaptive wavelet methods for the pathwise approximation of parabolic stochastic differential equations (SPDEs) on bounded Lipschitz domains, see [4–9, 32]. This kind of analysis has to be built on a regularity analysis in suitable scales of Besov spaces and the study of nonlinear approximation for the following reason. It is well-known that the approximation order, which can be achieved by adaptive wavelet algorithms, is determined by the regularity of the exact solution u in the scale

$$B^s_{\tau,\tau}(\mathscr{O}), \quad \frac{1}{\tau} = \frac{s}{d} + \frac{1}{p}, \quad s > 0, \tag{5.1}$$

of Besov spaces. In contrast, the efficiency of nonadaptive (uniform) schemes depend on the $L_p(\mathscr{O})$-Sobolev regularity of u. Therefore, the use of adaptive algorithms to solve SPDEs is completely justified, whenever the Besov regularity of u in the scale (5.1) exceeds its $L_p(\mathscr{O})$-Sobolev regularity. For deterministic PDEs, promising results in this direction have already been developed; we refer to [11, 13, 14, 19] for a detailed discussion. However, for their stochastic counterparts this line of research has only been started a few years ago.

Generally speaking, the numerics of parabolic SPDEs has been rapidly developing during the last decade, and different problems, like strong and weak approximation and the quadrature problem, are being studied. While the vast majority of papers in this field is devoted to uniform discretizations in space and time, spatially adaptive methods or nonuniform time discretizations have been considered in [5, 7] and in [34], respectively.

The SPDEs under consideration in this paper are given as follows. Let $\mathscr{O} \subset \mathbb{R}^d$ be a bounded Lipschitz domain, $T \in (0,\infty)$, and let $(w^k_t)_{t\in[0,T]}$ be an independent family of one-dimensional standard Wiener processes. We consider parabolic SPDEs with zero Dirichlet boundary conditions of the form

$$\left.\begin{aligned} du &= \Big(\sum_{i,j=1}^{d} a^{ij} u_{x^i x^j} + f\Big)dt + \sum_{k=1}^{\infty} g^k\, dw^k_t \quad \text{on } \Omega\times[0,T]\times\mathscr{O},\\ u &= 0 \qquad\qquad \text{on } \Omega\times(0,T]\times\partial\mathscr{O},\\ u(0,\cdot) &= u_0 \qquad\qquad \text{on } \Omega\times\mathscr{O}. \end{aligned}\right\} \tag{5.2}$$

In Sect. 5.2.1 we recall key results on the existence and uniqueness and the spatial regularity of the solution $u = (u(t,\cdot))_{t\in[0,T]}$ to Eq. (5.2) in weighted Sobolev spaces, as far as they are needed for our purposes. This approach to parabolic SPDEs was initiated in [26]. In Sect. 5.2.2 we briefly discuss Besov spaces and their characterization by means of wavelet expansions.

Results on the Besov or Hölder-Besov regularity of u, i.e., sufficient conditions on the parameters a^{ij}, f, g^k, and u_0 of Eq. (5.2) for

$$\mathbb{E}\Big[\int_0^T \|u(t,\cdot)\|^q_{B^s_{\tau,\tau}(\mathscr{O})}\,dt\Big] < \infty$$

or

$$\mathbb{E}\|u\|^q_{\mathscr{C}^r([0,T];B^s_{\tau,\tau}(\mathscr{O}))} < \infty$$

to hold, together with explicit upper bounds in terms of the parameters of Eq. (5.2), are presented in Sect. 5.3. For the justification to use adaptive approximation schemes it is important to know whether the smoothness in the scale (5.1) exceeds the smoothness in the scale of Sobolev spaces $W^s_p(\mathscr{O})$, $s > 0$; specific examples of SPDEs with this property are discussed in Sect. 5.3.4.

Rothe's method for Eq. (5.2) leads to a sequence of elliptic subproblems with random right-hand sides g, say. As a model problem of this type, we consider the Poisson equation with Dirichlet boundary conditions on bounded Lipschitz domains in Sect. 5.4. In Sect. 5.4.1, we introduce a stochastic model for g that is based on a wavelet expansion with random coefficients. The latter are products of normally distributed and Bernoulli distributed random variables, and the model allows to explicitly control the Besov regularity of the realizations of g. Furthermore, we determine the asymptotic behaviour of the linear approximation error and the error of the best (average) N-term wavelet approximation in Sect. 5.4.1. In Sect. 5.4.2 we present upper bounds for the error of best N-term wavelet approximation for the solution of the Poisson equation, together with numerical experiments, which very well match the asymptotic error analysis.

5.2 Preliminaries

Throughout this paper $\mathscr{O} \subset \mathbb{R}^d$ denotes a bounded Lipschitz domain. The underlying filtered probability space $(\Omega, \mathscr{A}, (\mathscr{F}_t)_{t\in[0,T]}, \mathbb{P})$ for Eq. (5.2) is assumed to satisfy the usual conditions. To simplify the exposition we assume that the coefficients a^{ij} do *not* depend on $(\omega, t, x) \in \Omega \times [0,T] \times \mathscr{O}$, and the matrix $(a^{ij}) \in \mathbb{R}^{d\times d}$ is assumed to be symmetric and positive definite. The force terms f and g^k are real-valued functions on $\Omega \times [0,T] \times \mathscr{O}$. For a better readability, we omit the notation of the sums $\sum_{i,j}^d$ and $\sum_{k=1}^\infty$ in the sequel and use the summation convention on the repeated indices i, j, k. Let us stress that most of the regularity results presented in Sect. 5.3 extend to more general equations, including, e.g., the case of multiplicative noise, cf. [8, Appendix B].

5.2.1 SPDEs in Weighted Sobolev Spaces

Our regularity analysis of the solution to Eq. (5.2) in the nonlinear approximation scale (5.1) relies on the theory of SPDEs in weighted Sobolev spaces initiated in [26] and developed further, e.g., in [30,31] and [22,23]. In this section, we describe the basic concepts of this theory and state an existence result from [23] which we will use later on in Theorem 5.1. We are interested in solutions to equations of the form

$$du = (a^{ij}u_{x^ix^j} + f)dt + g^k dw_t^k, \qquad u(0,\cdot) = u_0, \tag{5.3}$$

in certain stochastic parabolic weighted Sobolev spaces $\mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T)$ with smoothness, weight and summability parameters

$$\gamma \in \mathbb{R}, \quad \theta \in \mathbb{R}, \quad p,q \in [2,\infty). \tag{5.4}$$

If the parameters are chosen properly, all elements in these spaces fulfill the zero Dirichlet boundary condition in (5.2), cf. Remark 5.3 below. In the context of these spaces, the assumption (5.4) is taken for granted throughout the whole article.

5.2.1.1 Weighted Sobolev Spaces

We give the definition of weighted Sobolev spaces $H^{\gamma}_{p,\theta}(\mathscr{O})$ as introduced in [33].

Let $\rho(x) := \operatorname{dist}(x,\partial\mathscr{O})$ be the distance between $x \in \mathscr{O}$ and the boundary $\partial\mathscr{O}$ of $\mathscr{O}$. In the sequel, let $n \in \mathbb{Z}$. Fix $k_0 > 0$ and consider the layers $\mathscr{O}_n \subset \mathscr{O}$ of points x whose distance to $\partial\mathscr{O}$ lies in the interval (e^{-n-k_0}, e^{-n+k_0}), i.e.,

$$\mathscr{O}_n := \{x \in \mathscr{O} : e^{-n-k_0} < \rho(x) < e^{-n+k_0}\}.$$

Let $\mathscr{C}_0^\infty(\mathscr{G})$ be the space of smooth functions with compact support in $\mathscr{G} \subset \mathbb{R}^d$. Consider a sequence of nonnegative functions $\zeta_n \in \mathscr{C}_0^\infty(\mathscr{O}_n)$ satisfying

$$\inf_{x\in\mathscr{O}} \sum_{n\in\mathbb{Z}} \zeta_n(x) > 0 \text{ and}$$

$$\sup_{x\in\mathscr{O},n\in\mathbb{Z}} |D^\alpha \zeta_n(x)| e^{-|\alpha|n} < \infty \text{ for all } \alpha \in \mathbb{N}_0^d,$$

where $D^\alpha = \partial^\alpha/(\partial x_1^{\alpha_1}\cdots\partial x_d^{\alpha_d})$ and $|\alpha| = \alpha_1 + \ldots + \alpha_d$. If $\mathscr{O}_n$ is empty we set $\zeta_n = 0$. For a construction of such functions see, e.g., [9, Section 2]. The weighted Sobolev spaces $H^{\gamma}_{p,\theta}(\mathscr{O})$ consist of (generalized) functions $u \in \mathscr{D}'(\mathscr{O})$ such that the localized and dilated functions $u^{(n)} := \zeta_n(e^{-n}\cdot)u(e^{-n}\cdot)$ belong to the Bessel potential space $H^{\gamma}_p(\mathbb{R}^d) = (1-\Delta)^{-\gamma/2}L_p(\mathbb{R}^d)$ and the sequence of their norms lies

in a certain weighted ℓ_p-space. As usual, $\mathscr{D}'(\mathscr{O})$ denotes the space of distributions on $\mathscr{O}$.

Definition 5.1. We define

$$H^\gamma_{p,\theta}(\mathscr{O}) := \Big\{u \in \mathscr{D}'(\mathscr{O}) \,:\, \|u\|^p_{H^\gamma_{p,\theta}(\mathscr{O})} := \sum_{n\in\mathbb{Z}} e^{-n\theta} \big\|(1-\Delta)^{\gamma/2} u^{(n)}\big\|^p_{L_p(\mathbb{R}^d)} < \infty\Big\}.$$

Remark 5.1. If $\gamma = m \in \mathbb{N}_0$, then the spaces $H^\gamma_{p,\theta}(\mathscr{O})$ can be characterized as

$$H^0_{p,\theta}(\mathscr{O}) = L_{p,\theta}(\mathscr{O}) := L_p(\mathscr{O}, \rho(x)^{\theta-d}dx),$$
$$H^m_{p,\theta}(\mathscr{O}) = \Big\{u \in L_{p,\theta}(\mathscr{O}) \,:\, \rho^{|\alpha|} D^\alpha u \in L_{p,\theta}(\mathscr{O}) \text{ for all } \alpha \in \mathbb{N}^d_0 \text{ with } |\alpha| \le m\Big\},$$

and one has the norm equivalence

$$\sum_{\alpha\in\mathbb{N}^d_0,\,|\alpha|\le m} \int_{\mathscr{O}} \Big|\rho(x)^{|\alpha|} D^\alpha u(x)\Big|^p \rho(x)^{\theta-d}\,dx \;\asymp\; \|u\|^p_{H^m_{p,\theta}(\mathscr{O})}.$$

By $f \asymp g$ we indicate that there is a constant $c > 0$ such that $c^{-1} f(x) \le g(x) \le c g(x)$ for all x. The spaces $H^\gamma_{p,\theta}(\mathscr{O})$ with noninteger parameter γ can also be obtained by complex interpolation. Moreover, one has a monotonicity in the parameters: if $\gamma_1 \le \gamma_2$, $p_1 \le p_2$ and $\theta_1 \le \theta_2$, then $H^{\gamma_2}_{p_2,\theta_1}(\mathscr{O}) \hookrightarrow H^{\gamma_1}_{p_1,\theta_2}(\mathscr{O})$. All these facts and further details can be found in [33].

For the formulation of the assumption on the noise term in Eq. (5.3) we use analogously defined spaces of $\ell_2(\mathbb{N})$-valued functions $g = (g^k)_{k\in\mathbb{N}}$. The norm in $\ell_2 = \ell_2(\mathbb{N})$ is denoted by $|\cdot|_{\ell_2}$.

Definition 5.2. We define

$$H^\gamma_{p,\theta}(\mathscr{O};\ell_2) := \Big\{g \in (H^\gamma_{p,\theta}(\mathscr{O}))^{\mathbb{N}} \,:\, \|g\|^p_{H^\gamma_{p,\theta}(\mathscr{O};\ell_2)} := \sum_{n\in\mathbb{Z}} e^{-n\theta} \Big\| \Big|\big((1-\Delta)^{\gamma/2}(g^k)^{(n)}\big)_{k\in\mathbb{N}}\Big|_{\ell_2}\Big\|_{L_p(\mathbb{R}^d)} < \infty\Big\},$$

with $v^{(n)} = \zeta_n(e^{-n}\cdot)v(e^{-n}\cdot)$ for $v \in \mathscr{D}'(\mathscr{O})$ as above.

5.2.1.2 Stochastic Parabolic Spaces

The stochastic parabolic weighted Sobolev spaces $\mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O}, T)$ are defined in terms of the following spaces of stochastic processes and initial conditions u_0. By $\mathscr{P}$ we denote the predictable σ-algebra w.r.t. the filtration $(\mathscr{F}_t)_{t\in[0,T]}$.

Definition 5.3. We set

$$\mathbb{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T) := L_q(\Omega\times[0,T],\mathscr{P},\mathbb{P}\otimes dt;H^{\gamma}_{p,\theta}(\mathscr{O})),$$
$$\mathbb{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T;\ell_2) := L_q(\Omega\times[0,T],\mathscr{P},\mathbb{P}\otimes dt;H^{\gamma}_{p,\theta}(\mathscr{O};\ell_2)),$$
$$U^{\gamma,q}_{p,\theta}(\mathscr{O}) := L_q(\Omega,\mathscr{F}_0,\mathbb{P};H^{\gamma-2/q}_{p,\theta-p(1-2/q)}(\mathscr{O})).$$

If $p=q$ we also write $\mathbb{H}^{\gamma}_{p,\theta}(\mathscr{O},T)$, $\mathbb{H}^{\gamma}_{p,\theta}(\mathscr{O},T;\ell_2)$, $U^{\gamma}_{p,\theta}(\mathscr{O})$ instead of $\mathbb{H}^{\gamma,p}_{p,\theta}(\mathscr{O},T)$, $\mathbb{H}^{\gamma,p}_{p,\theta}(\mathscr{O},T;\ell_2)$ and $U^{\gamma,p}_{p,\theta}(\mathscr{O})$ respectively.

Remark 5.2. The restrictions on the parameters in the definition of $U^{\gamma,q}_{p,\theta}(\mathscr{O})$ is due to considerations concerning the trace at $t=0$ of solutions to equations of the type (5.3), cf. [29, Remarks 3.3 and 4.2].

Definition 5.4. We write $u\in\mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T)$ if, and only if, we have $u\in\mathbb{H}^{\gamma,q}_{p,\theta-p}(\mathscr{O},T)$, $u(0,\cdot)\in U^{\gamma,q}_{p,\theta}(\mathscr{O})$, and there exist $f\in\mathbb{H}^{\gamma-2,q}_{p,\theta+p}(\mathscr{O},T)$ and $g\in\mathbb{H}^{\gamma-1,q}_{p,\theta}(\mathscr{O},T;\ell_2)$ such that

$$du = f\,dt + g^k\,dw^k_t$$

in the sense of distributions. That is, for any $\varphi\in\mathscr{C}^{\infty}_0(\mathscr{O})$, with probability one, the equality

$$(u(t,\cdot),\varphi) = (u(0,\cdot),\varphi) + \int_0^t (f(s,\cdot),\varphi)\,ds + \sum_{k=1}^{\infty}\int_0^t (g^k(s,\cdot),\varphi)\,dw^k_s \tag{5.5}$$

holds for all $t\in[0,T]$. If $p=q$, we also write $\mathfrak{H}^{\gamma}_{p,\theta}(\mathscr{O},T)$ instead of $\mathfrak{H}^{\gamma,p}_{p,\theta}(\mathscr{O},T)$.

We add that, due to our assumption $p,q\geq 2$, the series of stochastic integrals in (5.5) converges in the space of continuous, square-integrable $(\mathscr{F}_t)_{t\in[0,T]}$-martingales, compare [27, Remark 3.2] or [8, Appendix A]. Moreover, given u, the coefficients f and g^k are determined uniquely and we write $\mathbb{D}u := f$ and $\mathbb{S}u := g$. The space $\mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T)$ is a Banach space when endowed with the norm

$$\|u\|_{\mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T)} := \|u\|_{\mathbb{H}^{\gamma,q}_{p,\theta-p}(\mathscr{O},T)} + \|\mathbb{D}u\|_{\mathbb{H}^{\gamma-2,q}_{p,\theta+p}(\mathscr{O},T)} + \|\mathbb{S}u\|_{\mathbb{H}^{\gamma-1,q}_{p,\theta}(\mathscr{O},T;\ell_2)} + \|u(0,\cdot)\|_{U^{\gamma,q}_{p,\theta}(\mathscr{O})},$$

compare [29, Remark 3.8].

5.2.1.3 Solution Concept and Existence of Solutions

We use the following notion of a solution.

Definition 5.5. A stochastic process $u \in \mathfrak{H}^{\gamma,q}_{p,\theta}(G,T)$ is a *solution* to Eq. (5.3) if, and only if, $\mathbb{D}u = a^{ij}u_{x^ix^j} + f$, $\mathbb{S}u = (g^k)_{k\in\mathbb{N}}$ and $u(0,\cdot) = u_0$ in the sense of Definition 5.4.

Remark 5.3. A solution $u \in \mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T)$ to Eq. (5.3) fulfills the boundary condition in Eq. (5.2) in the sense of traces whenever $\gamma > 1/p$ and $d-1 < \theta < d+p-1$, see [9, Remark 6.7]. To simplify the exposition we focus on existence results for solutions in $\mathfrak{H}^{\gamma,q}_{p,d}(\mathscr{O},T)$, i.e., on the case $\theta = d$. However, we note that there are more general existence results available, compare Remark 5.5.

Remark 5.4. In many cases our solution concept fits to the one used in the semigroup approach to SPDEs, cf. [18, 36]. Assume for example that Eq. (5.3) admits a solution $u \in \mathfrak{H}^{1,q}_{2,d}(\mathscr{O},T)$, that the process f is $L_2(\mathscr{O})$-valued, predictable, with locally Bochner integrable trajectories, and that $g = (g^k)_{k\in\mathbb{N}} \in H^0_{2,d}(\mathscr{O};\ell_2)$ is constant in (ω,t). Then this solution coincides with the unique weak solution in the sense of [18, Chapter 5] to the equation

$$du(t) = (Au(t) + f(t))dt + G\,dW(t), \quad u(0,\cdot) = u_0, \quad t \in [0,T],$$

compare [8, Remark 2.14]. Here $A : D(A) \subset L_2(\mathscr{O}) \to L_2(\mathscr{O})$ is the unbounded linear operator defined by $Ah := a^{ij}h_{x^ix^j}$, $h \in D(A) := \{\zeta \in \mathring{W}^1_2(\mathscr{O}) : a^{ij}\zeta_{x^ix^j} \in L_2(\mathscr{O})\}$, G is the Hilbert-Schmidt operator from $\ell_2 = \ell_2(\mathbb{N})$ to $L_2(\mathscr{O})$ defined by $G\,\mathbf{a} := g^k\mathbf{a}^k$, $\mathbf{a} = (\mathbf{a}^k)_{k\in\mathbb{N}} \in \ell_2$, and $W = (w^1, w^2, \ldots)$ is the cylindrical Wiener process on ℓ_2 with coordinate processes $w^k = (w^k_t)_{t\in[0,T]}$, $k \in \mathbb{N}$. Here and below we use the notation $\mathring{W}^m_2(\mathscr{O})$ for the closure of the space $\mathscr{C}^\infty_0(\mathscr{O})$ in the L_2-Sobolev space $W^m_2(\mathscr{O})$ of order $m \in \mathbb{N}$.

The following result concerning existence and uniqueness of a solution to Eq. (5.3) in $\mathfrak{H}^{\gamma}_{p,d}(\mathscr{O},T) = \mathfrak{H}^{\gamma,p}_{p,d}(\mathscr{O},T)$ summarizes Theorem 2.12, Remark 2.13 and Theorem 2.15 of [23]. Detailed proofs can be found in [23, Section 4] and [23, Section 5].

Theorem 5.1. *There exists $p_0 > 2$, such that for all $p \in [2,p_0)$, $f \in \mathbb{H}^{\gamma}_{p,d+p}(\mathscr{O},T)$, $g \in \mathbb{H}^{\gamma+1}_{p,d}(\mathscr{O},T;\ell_2)$ and $u_0 \in U^{\gamma+2}_{p,d}(\mathscr{O})$, Eq. (5.3) has a unique solution u in the class $\mathfrak{H}^{\gamma+2}_{p,d}(\mathscr{O},T)$. For this solution*

$$\|u\|_{\mathfrak{H}^{\gamma+2}_{p,d}(\mathscr{O},T)} \le C\left(\|f\|_{\mathbb{H}^{\gamma}_{p,d+p}(\mathscr{O},T)} + \|g\|_{\mathbb{H}^{\gamma+1}_{p,d}(\mathscr{O},T;\ell_2)} + \|u_0\|_{U^{\gamma+2}_{p,d}(\mathscr{O})}\right),$$

where the constant C depends only on d, p, γ, (a^{ij}), T and $\mathscr{O}$.

Remark 5.5. Theorem 1 remains valid if we replace d in all weight parameters by $\theta \in (d-\kappa, d+\kappa)$ for a certain $\kappa \in (0,1)$ that depends on the shape of the domain $\mathscr{O}$, see [23]. If $\mathscr{O}$ is a C^1-domain, the result remains valid with $p_0 = \infty$ and one can choose an arbitrary $\theta \in (d-1, d+p-1)$ instead of d in the weight parameters, see [22]. This is important since we obtain even higher spatial regularity in the scale (5.1) for smaller values of θ, see Sect. 5.3.

A regularity result for the stochastic heat equation in $\mathfrak{H}^{\gamma,q}_{p,d}(\mathscr{O},T)$ with $q > p$ is presented in Sect. 5.3.2 below. It will lead to a Hölder-Besov regularity result in Sect. 5.3.3, see Theorem 5.6 (ii).

5.2.2 Besov Spaces and Wavelets

We consider Besov spaces $B^{\gamma}_{p,q}(\mathscr{O})$ with smoothness parameter $\gamma > 0$ and summability parameters $p, q \in (0,\infty)$, see, e.g., [8, Section 2.2] and the references therein.

In general, a wavelet basis $\Psi := \{\psi_\lambda : \lambda \in \nabla\}$ is a Riesz basis for an L_2-space with specific properties, cf. [10]. The indices $\lambda \in \nabla$ typically encode several types of information, namely the scale, often denoted by $|\lambda|$, the spatial location, and also the type of the wavelet. For instance, on the real line, $|\lambda| = j \in \mathbb{Z}$ denotes the dyadic refinement level and $2^{-j}k$ with $k \in \mathbb{Z}$ stands for the location of the wavelet. We ignore any explicit dependence on the type of the wavelet, since this only produces additional constants. Hence, we use $\lambda = (j,k)$ and

$$\nabla = \{(j,k) : j \geq j_0, k \in \nabla_j\},$$

where ∇_j is some countable index set and $|(j,k)| = j$. Moreover, $\tilde{\Psi} := \{\tilde{\psi}_\lambda : \lambda \in \nabla\}$ denotes the dual wavelet basis.

In Sect. 5.4 we assume, that the domain $\mathscr{O}$ under consideration enables us to construct a wavelet basis Ψ, which have local support, i.e., that $\operatorname{diam}(\operatorname{supp}\ \psi_\lambda) \asymp 2^{-|\lambda|}$, and satisfy the cancellation property. Furthermore we assume that the cardinalities of the index sets ∇_j satisfy $\#\nabla_j \asymp 2^{jd}$ and that the wavelet basis induces the norm equivalence

$$\|v\|_{B^{\gamma}_{p,q}(\mathscr{O})} \asymp \left(\sum_{j=j_0}^{\infty} 2^{j(\gamma+d(\frac{1}{2}-\frac{1}{p}))q} \left(\sum_{k\in\nabla_j} |\langle v, \tilde{\psi}_{j,k}\rangle_{L_2(\mathscr{O})}|^p \right)^{q/p} \right)^{1/q}, \tag{5.6}$$

see [7]. Suitable constructions of wavelets on domains can be found in [3,15–17,35].

Remark 5.6. For the proofs of the Besov regularity results for SPDEs in [4, 8, 9], cf. Sect. 5.3 below, the existence of a suitable wavelet on the domain is not needed. We use the fact that for bounded Lipschitz domains $\mathscr{O} \subset \mathbb{R}^d$ the spaces $B^{\gamma}_{p,q}(\mathscr{O})$,

$\gamma > \max\{0, d(1/p-1)\}$, can be characterized by means of extension operators and wavelet expansions on all of $\mathbb{R}^d$: For a suitable wavelet basis $\Psi = \{\psi_\lambda : \lambda \in \nabla\}$ on $\mathbb{R}^d$ with dual basis $\tilde{\Psi} = \{\tilde{\psi}_\lambda : \lambda \in \nabla\}$ and a linear and bounded extension operator $\mathscr{E} : B^\gamma_{p,q}(\mathscr{O}) \to B^\gamma_{p,q}(\mathbb{R}^d)$ one has the norm equivalence

$$\|v\|_{B^\gamma_{p,q}(\mathscr{O})} \asymp \left(\sum_{j=j_0}^{\infty} 2^{j(\gamma+d(\frac{1}{2}-\frac{1}{p}))q} \left(\sum_{k\in\nabla_j} |\langle \mathscr{E}v, \tilde{\psi}_{j,k}\rangle_{L_2(\mathbb{R}^d)}|^p \right)^{q/p} \right)^{1/q}, \tag{5.7}$$

see, e.g., [8] for details. Note that the inner products appearing in (5.7) are the wavelet coefficients of $\mathscr{E}v$ w.r.t. the basis Ψ.

5.3 Regularity Analysis in Besov Spaces

Throughout this section let $d \geq 2$.

5.3.1 Spatial Besov Regularity

The spatial regularity in the Besov scale (5.1) of solutions to SPDEs of the type (5.3) has been analyzed for the first time in [8], where weighted Sobolev norm estimates as in Theorem 5.1 have been combined with the wavelet technique as inspired by the results in [14]. The results in [8] have been improved in the subsequent paper [9], which we follow here.

In order to gain some insight into the strategy to obtain regularity results in the scale (5.1), let us denote by $\Theta(\mathscr{O})$ the set of harmonic functions on $\mathscr{O}$, so that we can reformulate the main result in [14] as follows:

$$\Theta(\mathscr{O}) \cap B^\nu_{p,p}(\mathscr{O}) \subset B^s_{\tau,\tau}(\mathscr{O}), \quad \frac{1}{\tau} = \frac{s}{d} + \frac{1}{p}, \quad \text{for all } 0 < s < \nu\frac{d}{d-1}.$$

The key idea to prove this is to make use of the fact that the Besov smoothness of a function can be described in terms of decay properties of its wavelet coefficients, see the norm equivalence (5.7). Further, one has to use the fact that harmonic functions contained in $B^\nu_{p,p}(\mathscr{O})$ have finite weighted Sobolev seminorm

$$|u|_{H^m_{p,d-\nu p}(\mathscr{O})} := \Big(\sum_{\substack{\alpha\in\mathbb{N}_0^d \\ |\alpha|=m}} \int_{\mathscr{O}} \big|\rho(x)^{|\alpha|} D^\alpha u(x)\big|^p \rho(x)^{-\nu p} dx \Big)^{1/p}$$

for any $\nu < m \in \mathbb{N}$. This together with the Lipschitz character of $\partial\mathscr{O}$ allow for a certain control of the decay of the wavelet coefficients. It turns out that, using similar arguments as in [14], one can even show that for $\nu > 0$ and $m \in \mathbb{N}$,

$$H^m_{p,d-\nu p}(\mathscr{O}) \cap B^\nu_{p,p}(\mathscr{O}) \hookrightarrow B^s_{\tau,\tau}(\mathscr{O}), \quad \frac{1}{\tau} = \frac{s}{d} + \frac{1}{p}, \quad 0 < s < \min\Big\{m, \nu\frac{d}{d-1}\Big\},$$

where '$\hookrightarrow$' denotes a continuous embedding. The Besov regularity results for SPDEs in [8] are implicitly based on this embedding, although the embedding is not explicitly stated there. It has been substantially improved and generalized in [9, Section 6]: On the one hand, by using refined estimates for the wavelet coefficients, the embedding can be generalized to arbitrary smoothness parameters $\gamma > 0$ instead of $m \in \mathbb{N}$. (This generalization is needed for the Hölder-Besov regularity results below.) On the other hand, by using interpolation arguments, one can show that $H^\gamma_{p,d-\nu p}(\mathscr{O}) \hookrightarrow B^{\gamma\wedge\nu}_{p,p}(\mathscr{O})$. As a consequence, see [9, Theorem 6.9], one obtains

Theorem 5.2. *Let $p \in [2,\infty)$ and γ and ν be positive numbers. Then*

$$H^\gamma_{p,d-\nu p}(\mathscr{O}) \hookrightarrow B^s_{\tau,\tau}(\mathscr{O}), \quad \frac{1}{\tau} = \frac{s}{d} + \frac{1}{p}, \quad \textit{for all} \quad 0 < s < \min\Big\{\gamma, \nu\frac{d}{d-1}\Big\}.$$

In the sequel, we use the shorthand notation

$$L_q(\Omega_T; X) := L_q(\Omega \times [0,T], \mathscr{P}, \mathbb{P} \otimes dt; X),$$

where X is a quasi-Banach space, e.g., $X = B^s_{\tau,\tau}(\mathscr{O})$. Thus, $L_q(\Omega_T; X)$ is the space of (equivalence classes) of strongly $\mathscr{P}$-measurable functions $f : \Omega \times [0,T] \to X$ such that $\mathbb{E}\int_0^T \|f\|_X^q \, dt$ is finite. As a consequence of Theorem 5.2 and the definition of $\mathfrak{H}^{\gamma,q}_{p,d}(\mathscr{O},T)$, we have the embedding

$$\mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T) \hookrightarrow L_q(\Omega_T; B^s_{\tau,\tau}(\mathscr{O})), \quad \frac{1}{\tau} = \frac{s}{d} + \frac{1}{p}, \tag{5.8}$$

holding for $\gamma > 0$ and

$$0 < s < \min\Big\{\gamma, \Big(1 + \frac{d-\theta}{p}\Big)\frac{d}{d-1}\Big\}.$$

The combination of this embedding and Theorem 5.1 with $\gamma = 0$ leads to the following spatial Besov regularity result for SPDEs of the type (5.3), see [9, Theorem 7.2].

Theorem 5.3. *Let $\gamma + 2 \in (0,\infty)$, $p_0 > 2$ and $p \in [2, p_0)$ as in Theorem 5.1. Then, for any $f \in \mathbb{H}^\gamma_{p,d+p}(\mathscr{O},T)$, $g \in \mathbb{H}^{\gamma+1}_{p,d}(\mathscr{O},T;\ell_2)$ and $u_0 \in U^{\gamma+2}_{p,d}(\mathscr{O})$, the unique solution $u \in \mathfrak{H}^{\gamma+2}_{p,d}(\mathscr{O},T)$ to Eq. (5.3) satisfies*

$$u \in L_p(\Omega_T; B^s_{\tau,\tau}(\mathcal{O})), \quad \frac{1}{\tau} = \frac{s}{d} + \frac{1}{p}, \quad \text{for all } 0 < s < \min\left\{\gamma + 2, \frac{d}{d-1}\right\}. \tag{5.9}$$

Moreover, for any s in (5.9), there exists a constant C, which does not depend on u, f, g and u_0, such that

$$\mathbb{E}\Big[\int_0^T \|u(t,\cdot)\|^p_{B^s_{\tau,\tau}(\mathcal{O})}\,dt\Big] \le C\left(\|f\|^p_{\mathbb{H}^{\gamma}_{p,d+p}(\mathcal{O},T)} + \|g\|^p_{\mathbb{H}^{\gamma+1}_{p,d}(\mathcal{O},T;\ell_2)} + \|u_0\|^p_{U^{\gamma+2}_{p,d}(\mathcal{O})}\right).$$

Remark 5.7. The result in Theorem 5.3 is the crucial step in our regularity analysis for SPDEs of the type (5.3). Now we have an explicit assertion concerning the spatial regularity of the solution u in the nonlinear approximation scale (5.1). In Sect. 5.3 we will see that the spatial regularity of u in the corresponding linear approximation scale $W^{\gamma}_p(\mathcal{O})$, $\gamma > 0$, is in general lower than the regularity in the nonlinear approximation scale (5.1). Thus, we can justify the use of spatially adaptive approximation methods for u in concrete situations.

Remark 5.8. The assertion of Theorem 1 concerning the existence of a solution to Eq. (5.3) in $\mathfrak{H}^{\gamma+2}_{p,d}(\mathcal{O},T)$ also holds for a more general class of linear SPDEs, cf. [23]. Moreover it has been extended to a class of semilinear SPDEs in [4]. Due to Theorem 5.2, both results lead to corresponding spatial Besov regularity results.

5.3.2 Regularity in Weighted Sobolev Spaces Revisited

For bounded Lipschitz domains $\mathcal{O}$, first regularity results in the spaces $\mathfrak{H}^{\gamma,q}_{p,d}(\mathcal{O},T)$ with $q > p$ have been derived recently in [9, Chapter 4]. For $q \gg p$ we are able to improve our results concerning the spatial regularity in the scale (5.1) w.r.t. regularity in time, cf. Sect. 5.3.3. We present a result from [9, Chapter 4] for the model equation

$$du = (\Delta u + f)\,dt + g^k\,dw^k_t, \qquad u(0,\cdot) = 0, \tag{5.10}$$

see Theorem 4.4 therein. The proof combines methods from the analytic approach to SPDEs as used in [28] with functional analytic results from the semigroup approach to SPDEs in [36] and results concerning the Dirichlet Laplacian in $L_p(\mathcal{O})$ from [37].

Theorem 5.4. *Let $\gamma \ge 0$. There exists an exponent p_0 with $p_0 > 3$ when $d \ge 3$ and $p_0 > 4$ when $d = 2$ such that for $p \in [2, p_0)$ and $p \le q < \infty$, Eq. (5.10) has a unique solution $u \in \mathfrak{H}^{\gamma+2,q}_{p,d}(\mathcal{O},T)$, provided*

$$f \in \mathbb{H}^{\gamma,q}_{p,d+p}(\mathcal{O},T) \cap \mathbb{H}^{0,q}_{p,d}(\mathcal{O},T) \quad \text{and} \quad g \in \mathbb{H}^{\gamma+1,q}_{p,d}(\mathcal{O},T;\ell_2) \cap \mathbb{H}^{1,q}_{p,d-p}(\mathcal{O},T;\ell_2).$$

Moreover, there exists a constant $C \in (0,\infty)$*, which does not depend on* f *and* g*, such that*

$$\|u\|_{\mathfrak{H}^{\gamma+2,q}_{p,d}(\mathscr{O},T)} \leq C\Big(\|f\|_{\mathbb{H}^{\gamma,q}_{p,d+p}(\mathscr{O},T)} + \|f\|_{\mathbb{H}^{0,q}_{p,d}(\mathscr{O},T)} + \|g\|_{\mathbb{H}^{\gamma+1,q}_{p,d}(\mathscr{O},T;\ell_2)} + \|g\|_{\mathbb{H}^{1,q}_{p,d-p}(\mathscr{O},T;\ell_2)}\Big).$$

5.3.3 Hölder-Besov Regularity

For $r \in (0,1)$ and a (quasi-)Banach space $(X, \|\cdot\|_X)$ we denote by $\mathscr{C}^r([0,T];X)$ the Hölder space of continuous functions $v : [0,T] \to X$ with finite (quasi-)norm $\|v\|_{\mathscr{C}^r([0,T];X)}$ defined by

$$[v]_{\mathscr{C}^r([0,T];X)} := \sup_{s,t\in[0,T]} \frac{\|v(t)-v(s)\|_X}{|t-s|^r},$$

$$\|v\|_{\mathscr{C}([0,T];X)} := \sup_{t\in[0,T]} \|v(t)\|_X,$$

$$\|v\|_{\mathscr{C}^r([0,T];X)} := \|v\|_{\mathscr{C}([0,T];X)} + [v]_{\mathscr{C}^r([0,T];X)}.$$

We want to improve the spatial Besov regularity results for SPDEs of the type (5.3) w.r.t. regularity in time and derive assertions of the form $u \in L_q(\Omega;\mathscr{C}^r([0,T];B^s_{\tau,\tau}(\mathscr{O})))$, $1/\tau = s/d + 1/p$, for certain $r \in (0,1)$ and $s > 0$. The key to achieve this is the following result concerning the Hölder regularity in time of elements of the spaces $\mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T)$, which has been shown in [9, Theorem 5.1]. The proof uses [29, Proposition 4.1], which covers the assertion of Theorem 5.5 with $\mathbb{R}^d_+$ instead of $\mathscr{O}$, and a quite technical boundary flattening argument exploiting the Lipschitz character of $\partial\mathscr{O}$.

Theorem 5.5. *Let* $2 \leq p \leq q < \infty$*,* $\gamma \in \mathbb{N}$ *and* $u \in \mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T)$*. Moreover, let*

$$2/q < \bar{r} < r \leq 1.$$

Then there exists a constant $C \in (0,\infty)$*, which does not depend on* T *and* u*, such that*

$$\mathbb{E}[u]^q_{\mathscr{C}^{\bar{r}/2-1/q}([0,T];H^{\gamma-r}_{p,\theta-p(1-r)}(\mathscr{O}))} \leq C\,T^{(r-\bar{r})q/2}\Big(\|u\|^q_{\mathbb{H}^{\gamma,q}_{p,\theta-p}(\mathscr{O},T)} + \|\mathbb{D}u\|^q_{\mathbb{H}^{\gamma-2,q}_{p,\theta+p}(\mathscr{O},T)} + \|\mathbb{S}u\|^q_{\mathbb{H}^{\gamma-1,q}_{p,\theta}(\mathscr{O},T;\ell_2)}\Big),$$

and

$$\mathbb{E}\|u\|^q_{\mathscr{C}^{\bar r/2-1/q}([0,T];H^{\gamma-r}_{p,\theta-p(1-r)}(\mathscr{O}))} \le C\,T^{(r-\bar r)q/2}\|u\|^q_{\mathfrak{H}^{\gamma,q}_{p,\theta}(\mathscr{O},T)}.$$

Remark 5.9. For the case that the summability parameters in time and space coincide, i.e., $q = p$, a result similar to Theorem 5.5 has been proved in [23, Theorem 2.9]. However, the technique used there does not work in the case $q > p$. This case is important for us, since it allows a wider range of parameters $\bar r$ and r and enables us to obtain Hölder-Besov regularity results for SPDEs, see Theorem 5.6 (ii).

The combination of Theorems 5.2, 5.4 and 5.5 leads to Besov and Hölder-Besov regularity results for the solution to Eq. (5.10).

Theorem 5.6. *Let $p \in [2, p_0)$ and $p \le q < \infty$, and let p_0 with $p_0 > 3$ when $d \ge 3$ and $p_0 > 4$ when $d = 2$ be as in Theorem 5.4. For $f \in \mathbb{H}^{0,q}_{p,d}(\mathscr{O},T)$ and $g \in \mathbb{H}^{1,q}_{p,d-p}(\mathscr{O},T;\ell_2)$, let u be the unique solution in the class $\mathfrak{H}^{2,q}_{p,d}(\mathscr{O},T)$ to Eq. (5.10). Then, the following assertions hold:*

(i) *We have*

$$u \in L_q(\Omega_T; B^s_{\tau,\tau}(\mathscr{O})), \quad \frac{1}{\tau} = \frac{s}{d} + \frac{1}{p}, \quad \text{for all } 0 < s < \frac{d}{d-1}. \tag{5.11}$$

For any s in (5.11), there exists a constant $C \in (0,\infty)$, which does not depend on u, f, and g, such that

$$\mathbb{E}\Big[\int_0^T \|u(t,\cdot)\|^q_{B^s_{\tau,\tau}(\mathscr{O})}\,dt\Big] \le C\Big(\|f\|_{\mathbb{H}^{0,q}_{p,d}(\mathscr{O},T)} + \|g\|_{\mathbb{H}^{1,q}_{p,d-p}(\mathscr{O},T;\ell_2)}\Big).$$

(ii) *Assume furthermore that $2/q < \bar r < 1$, and that s and τ fulfill*

$$\frac{1}{\tau} = \frac{s}{d} + \frac{1}{p} \quad \text{and} \quad 0 < s < \min\Big\{2-\bar r, (1-\bar r)\,\frac{d}{d-1}\Big\}.$$

Then, with a constant $C \in (0,\infty)$ that does not depend on u, f and g, we also have

$$\mathbb{E}\|u\|^q_{\mathscr{C}^{\bar r/2-1/q}([0,T];B^s_{\tau,\tau}(\mathscr{O}))} \le C\Big(\|f\|_{\mathbb{H}^{0,q}_{p,d}(\mathscr{O},T)} + \|g\|_{\mathbb{H}^{1,q}_{p,d-p}(\mathscr{O},T;\ell_2)}\Big). \tag{5.12}$$

Proof. By Remark 5.1 we have the continuous embeddings $H^0_{p,d}(\mathscr{O}) \hookrightarrow H^0_{p,d+p}(\mathscr{O})$ and $H^1_{p,d-p}(\mathscr{O}) \hookrightarrow H^1_{p,d}(\mathscr{O})$, so that Theorem 5.4 with $\gamma = 0$ implies the existence of a unique solution $u \in \mathfrak{H}^{2,q}_{p,d}(\mathscr{O},T)$ and the estimate

$$\|u\|_{\mathfrak{H}^{2,q}_{p,d}(\mathscr{O},T)} \le C\Big(\|f\|_{\mathbb{H}^{0,q}_{p,d}(\mathscr{O},T)} + \|g\|_{\mathbb{H}^{1,q}_{p,d-p}(\mathscr{O},T;\ell_2)}\Big). \tag{5.13}$$

The combination of (5.13) and Theorem 5.2 yields the assertion in (i), compare the embedding (5.8). Further, the combination of (5.13), Theorem 5.5 and Theorem 5.2 implies that the estimate (5.12) holds for $0 < s < \min\{2-r, (1-r)d/(d-1)\}$ with r as in Theorem 5.5. Since r can be chosen arbitrarily close to $\bar{r} < r$, we obtain the assertion in (ii). □

5.3.4 *Comparison with the Regularity in the Linear Approximation Scale*

The results on the spatial regularity of solutions to SPDEs in the nonlinear approximation scale (5.1) justify the use of adaptive schemes only if we can show that the regularity in the corresponding linear approximation scale $W_p^s(\mathscr{O})$, $p > 0$, is lower than the regularity in (5.1). We give two simple concrete examples in the Hilbert space setting ($p = 2$) where this is the case.

In both examples we consider equations of the type

$$du = \Delta u dt + g^k dw_t^k, \quad u(0,\cdot) = u_0,$$

and assume that $u_0 \in L_2(\Omega; \mathring{W}_2^2(\mathscr{O})) \subset U_{2,d}^2(\mathscr{O})$ and that $g = (g^k)_{k\in\mathbb{N}} \in H_{2,d}^1(\mathscr{O}; \ell_2)$ is constant in (ω, t). Recall that $\mathring{W}_2^n(\mathscr{O})$ denotes the closure of the space $\mathscr{C}_0^\infty(\mathscr{O})$ in the L_2-Sobolev space $W_2^n(\mathscr{O})$ of order $n \in \mathbb{N}$. By Theorem 5.1, there exists a unique solution in the class $\mathfrak{H}_{2,d}^2(\mathscr{O}, T)$. Since we use results derived within the semigroup approach to SPDEs, we recall that, by Remark 5.4, the solution in $\mathfrak{H}_{2,d}^2(\mathscr{O}, T)$ coincides with the weak solution in the sense of [18, Chapter 5] to the equation

$$du(t) = \Delta_{\mathscr{O}}^{\mathrm{D}} u(t) dt + G\, dW(t), \quad u(0,\cdot) = u_0,\ t \in [0,T], \tag{5.14}$$

where $\Delta_{\mathscr{O}}^{\mathrm{D}} : D(\Delta_{\mathscr{O}}^{\mathrm{D}}) \subset L_2(\mathscr{O}) \to L_2(\mathscr{O})$ is the Dirichlet-Laplacian on $\mathscr{O}$ with domain $D(\Delta_{\mathscr{O}}^{\mathrm{D}}) := \{v \in \mathring{W}_2^1(\mathscr{O}) : \Delta v \in L_2(\mathscr{O})\}$, and G and W are as in Remark 5.4.

In the first example, the spatial regularity of u in the linear approximation scale $W_2^s(\mathscr{O})$, $s > 0$, is limited due to the shape of the domain $\mathscr{O}$. In the second example, the limitation is due the incompatibility of the noise with the zero Dirichlet boundary condition. In view of the results in Sect. 5.4, we remark that another reason for a limited regularity in the linear approximation scale can be the limited Sobolev regularity of the noise term itself.

Example 5.1. Let $d = 2$ and let $\mathscr{O} \subset \mathbb{R}^2$ be a bounded, nonconvex, polygonal domain. Denote by ϖ the largest interior angle at a vertex of $\partial\mathscr{O}$. Assume that the range of the operator G is dense in $L_2(\mathscr{O})$. Then, it follows from [32, Example 3.6], that the weak solution u to Eq. (5.14) is such that

$$u(\omega,\cdot) \notin L_2([0,T];W_2^{1+\pi/\varpi}(\mathscr{O})) \quad \text{for } \mathbb{P}\text{-almost all } \omega \in \Omega. \tag{5.15}$$

We stress that the density of $G(L_2(\mathscr{O}))$ in $L_2(\mathscr{O})$ is essential here: it implies that the stochastic source term in (5.14) induces the appearance of corner singularities of the solution u, cf. Remark 5.10 below. Note that the smoothness parameter $1+\pi/\varpi$ in (5.15) is less than 2, since $\mathscr{O}$ is nonconvex. However, we have $u \in L_2(\Omega_T;B^s_{\tau,\tau}(\mathscr{O}))$, $1/\tau = s/2+1/2$ for all $s<2$ by Theorem 5.3.

Remark 5.10. Example 3.6 in [32] is a consequence of a general decomposition result concerning the weak solution u to a stochastic heat equation of the form

$$du(t) = [\Delta^{\mathrm{D}}_{\mathscr{O}}\, u(t) + F(u(t))]dt + G(u(t))\, dW(t), \quad u(0,\cdot) = u_0,\ t \in [0,T],$$

see [32, Theorem 3.3]. Under certain technical assumptions, u can be decomposed into a regular part u_{R} and a singular part u_{S}. Both parts have negative regularity in time. While u_{R} has full L_2-Sobolev regularity of order 2 in space, the spatial L_2-Sobolev regularity of the singular part u_S is in general limited by $1+\pi/\varpi$, since u_S may contain the corner singularities for the Poisson problem on $\mathscr{O}$.

Example 5.2. Let $\mathscr{O} \subset \mathbb{R}^d$ be a bounded $\mathscr{C}^\infty$-domain. Let $g^1 = 1$ and $g^k = 0$ for all $k \geq 2$. Then the weak solution to Eq. (5.14) is such that

$$u(t) \notin L_2(\Omega;W_2^\gamma(\mathscr{O})) \quad \text{for all } t \in (0,T] \text{ and } \gamma > 3/2.$$

This is a straightforward application of Itô's isometry, using the representation $u(t) = e^{tA}u_0 + \int_0^t e^{(t-s)A}G\,dW(s)$, the properties of the operator semigroup $(e^{tA})_{t\geq 0}$ generated by A, and the explicit characterization of the domains of fractional powers of A in [20]. Combining Theorem 5.2 and the general existence results mentioned in Remark 5.5, one can show that in the described situation, for $d=2$, we even have $u \in L_2(\Omega_T;B^s_{\tau,\tau}(\mathscr{O}))$, $1/\tau = s/2+1/2$ for all $s<3$.

5.4 Nonlinear Approximation for Elliptic Equations

One approach to approximate the solutions to equations of the type (5.3) is the horizontal method of lines, also known as Rothe's method, which starts with a discretization first in time and then in space. The parabolic equation can be interpreted as an abstract Cauchy problem, i.e., as an ordinary stochastic differential equation in some suitable function spaces. This immediately provides a way to employ adaptive strategies. Indeed, in time direction we might potentially even use SDE-solvers with step size control. This solver must be based on an implicit discretization scheme since the equation under consideration is usually stiff. Then, in each time step, a system of elliptic equations with random right-hand sides has to be solved. To this end, a second level of adaptivity, well-established adaptive

numerical schemes based, e.g., on wavelets or finite elements, can be used, see [6]. More details and results on error propagation and complexity estimates for Rothe's method can be found in [5], compare also [25].

Therefore, in this section, we study adaptive wavelet algorithms for the Poisson equation

$$\begin{aligned} -\Delta u &= g \quad \text{in} \quad \mathscr{O}, \\ u &= 0 \quad \text{on} \quad \partial\mathscr{O}, \end{aligned} \tag{5.16}$$

which serves as a model problem for elliptic equations with random right-hand side g. More precisely, g is a random function (random field) with realizations at least in $L_2(\mathscr{O})$, and we wish to approximate the realizations of the random function u in $W_2^1(\mathscr{O})$. We investigate a new stochastic model for g that provides an explicit control of the Besov regularity, and we analyze nonlinear approximations of both, g and u. An average N-term approximation of the right-hand side g can be simulated efficiently, and the nonlinear approximation rate for the solution u is achieved by means of an adaptive wavelet algorithm. More precisely, we apply optimally convergent wavelet algorithms, see [11–13], in a stochastic setting. These algorithms realize the convergence order of the best N-term approximation.

Rates of convergence for the approximation of stochastic evolution equations using uniform wavelet schemes in space are presented in [21]. In [24] wavelet methods have been used to simplify the additive noise term of linear stochastic evolution equations. In [25] a splitting scheme for a semilinear stochastic heat equation using uniform wavelet approximation for the stochastic part and an adaptive wavelet method for the deterministic part is considered. An algorithm for solving stochastic heat equations by nonuniform time discretization of the driving Brownian motion is presented in [34].

5.4.1 Random Functions in Besov Spaces

In this section we discuss linear and nonlinear approximations as well as the Besov regularity of random functions $g : \Omega \to L_2(\mathscr{O})$. The random functions are defined in terms of wavelet expansions according to a stochastic model that provides an explicit control for the Besov regularity and, in particular, induces sparsity of the wavelet coefficients. In the context of Bayesian nonparametric regression this model was introduced and analyzed in [1] and generalized in [2] in the case $\mathscr{O} = [0, 1]$ for Besov spaces with parameters $p, q \geq 1$.

The stochastic model is based on the wavelet expansion described in Sect. 5.2.2, where the coefficients are given as independent random variables $Y_{j,k}$ and $Z_{j,k}$ for $j \geq j_0$ and $k \in \nabla_j$ on some probability space $(\Omega, \mathscr{A}, \mathbb{P})$. The variables $Y_{j,k}$ are Bernoulli distributed with parameter

$$p_j := \min\left(1, C_1 2^{-\beta j d}\right), \quad \text{with} \quad \mathbb{P}(Y_{j,k} = 1) = p_j \text{ and } \mathbb{P}(Y_{j,k} = 0) = 1 - p_j,$$

where $\beta \in [0,1], C_1 > 0$. The random variables $Z_{j,k}$ are standard-normal $N(0,1)$-distributed and, in order to rescale their variances, we put

$$\sigma_j^2 := C_2 j^{\gamma d} 2^{-\alpha j d},$$

with parameters $\alpha, C_2 > 0, \gamma \in \mathbb{R}$. Now we define

$$g := \sum_{j=j_0}^{\infty} \sum_{k \in \nabla_j} \sigma_j Y_{j,k} Z_{j,k} \psi_{j,k}, \tag{5.17}$$

which converges $\mathbb{P}$-a.s. in $L_2(\mathcal{O})$.

Remark 5.11. Classical examples for Gaussian random functions on $\mathcal{O} = [0,1]^d$ are the Brownian sheet, which, in terms of smoothness, corresponds to $\alpha = 2$, $\beta = 0$ and $\gamma = 2(d-1)/d$, and Lévy's Brownian motion, which, in terms of smoothness, corresponds to $\alpha = (d+1)/d, \beta = 0$ and $\gamma = 0$.

Theorem 5.7. *Suppose that $s > \max\{0, d/p - d\}$. We have $g \in B^s_{p,q}(\mathcal{O})$ with probability one if, and only if,*

$$s < d\left(\frac{\alpha - 1}{2} + \frac{\beta}{p}\right) \tag{5.18}$$

or

$$s = d\left(\frac{\alpha - 1}{2} + \frac{\beta}{p}\right) \qquad \text{and} \qquad q\gamma d < -2. \tag{5.19}$$

Furthermore,

$$\mathbb{E}\|g\|^q_{B^s_{p,q}(\mathcal{O})} < \infty$$

if (5.18) *or* (5.19) *is satisfied.*

We set

$$V_{j_1} = \max_{j \le j_1} \max_{k \in \nabla_j} \sigma_j Y_{j,k} |Z_{j,k}|$$

in order to normalize the absolute values of the coefficients of the truncated wavelet expansion

$$\hat{g}_{j_1} := \sum_{j=j_0}^{j_1} \sum_{k \in \nabla_j} \sigma_j Y_{j,k} Z_{j,k} \psi_{j,k}.$$

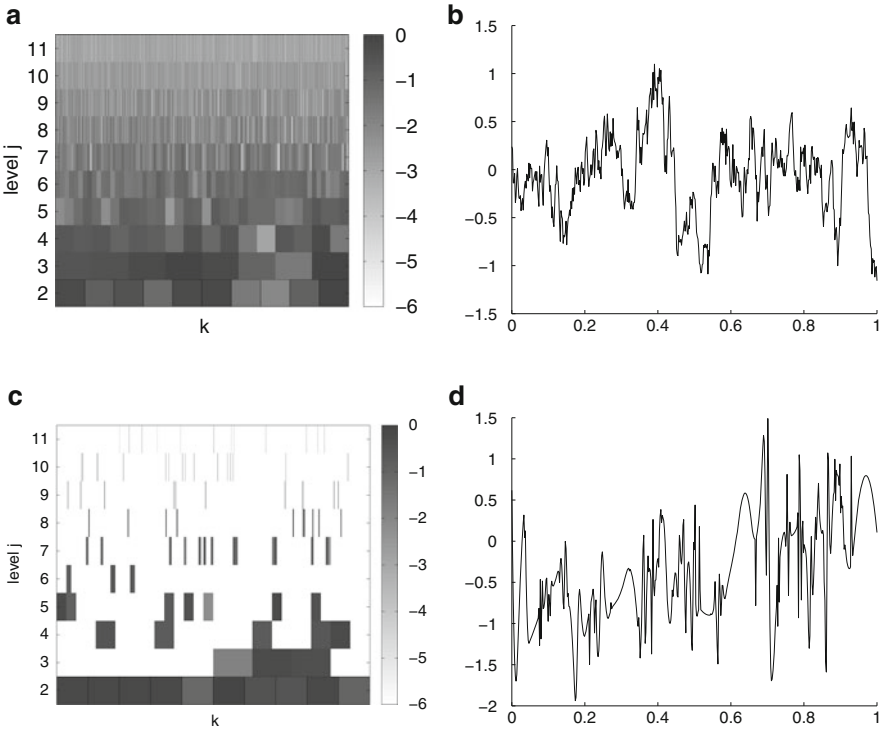


Fig. 5.1 *Left*: absolute values of normalized coefficients. *Right*: respective realization. (**a**) $\alpha = 2.0$, $\beta = 0.0$. (**b**) $\alpha = 2.0$, $\beta = 0.0$. (**c**) $\alpha = 1.2$, $\beta = 0.8$. (**d**) $\alpha = 1.2$, $\beta = 0.8$

Remark 5.12. In Fig. 5.1, the left column shows realizations of the normalized absolute values $\sigma_j Y_{j,k}|Z_{j,k}|/V_{j_1}$ of all coefficients on $\mathscr{O} = [0,1]$ up to level $j_1 = 11$. It exhibits that the parameter β induces sparsity patterns, for larger values of β more coefficients are zero and the wavelet expansion of g is sparser. The right column shows the corresponding realization of g. We observe that for $\beta = 0$ the realization is irregular everywhere, and by increasing β the irregularities become more isolated. The first row in Fig. 5.1 corresponds, in terms of smoothness, to a Brownian motion, see Remark 5.11. By keeping $\alpha + \beta = 2$, in the second row we obtain the same L_2-Sobolev smoothness, whereas it is well-known that piecewise smooth functions with isolated singularities have a higher Besov smoothness on the nonlinear approximation scale.

As a special case of Theorem 5.7, we consider the specific scale of Besov spaces $B^s_{\tau,\tau}(\mathscr{O})$ with

$$\frac{1}{\tau} = \frac{s-t}{d} + \frac{1}{2} \tag{5.20}$$

for some smoothness parameter t with $s > t \geq 0$, which determines the approximation order of best N-term wavelet approximation with respect to $W_2^t(\mathscr{O})$. For $t = 0$ this is the Besov scale (5.1) considered in the previous sections.

Corollary 5.1. *Suppose that*

$$0 \leq t < d\,\frac{\alpha+\beta-1}{2}$$

and $s > t$ as well as

$$(1-\beta)s < d\,\frac{\alpha+\beta-1}{2} - \beta t.$$

Let τ be given by (5.20). Then $g \in B^s_{\tau,\tau}(\mathscr{O})$ holds with probability one, and

$$\mathbb{E}\|g\|^2_{B^s_{\tau,\tau}(\mathscr{O})} < \infty.$$

Remark 5.13. It follows from Corollary 5.1 that by choosing the sparsity parameter β close to one we get an arbitrarily high regularity in the nonlinear approximation scale of Besov spaces, cf. (5.20), provided that the wavelet basis is sufficiently smooth. This is obviously not possible in the classical L_2-Sobolev scale, since, by Theorem 5.7 with $p = q = 2$, the Sobolev regularity is bounded by $d/2(\alpha+\beta-1)$.

In the following we study the approximation of g with respect to the L_2-norm. Any linear approximation method employs a fixed finite-dimensional linear subspace of $L_2(\mathscr{O})$ to approximate all realizations of g. The corresponding linear approximation error of g is given by

$$e_N^{\mathrm{lin}}(g) = \inf\left(\mathbb{E}\|g-\hat{g}\|^2_{L_2(\mathscr{O})}\right)^{1/2}$$

with the infimum taken over all measurable mappings $\hat{g} : \Omega \to L_2(\mathscr{O})$ such that

$$\dim(\mathrm{span}(\hat{g}(\Omega))) \leq N.$$

Theorem 5.8. *The linear approximation error satisfies*

$$e_N^{\mathrm{lin}}(g) \asymp (\ln N)^{\frac{\gamma d}{2}}\, N^{-\frac{\alpha+\beta-1}{2}}.$$

The best N-term wavelet approximation imposes a restriction only on the number

$$\eta(g) = \#\Big\{\lambda \in \nabla : c_\lambda \neq 0, g = \sum_{\lambda\in\nabla} c_\lambda\,\psi_\lambda\Big\}$$

of nonzero wavelet coefficients of g. Hence the corresponding error of best N-term approximation for g is given by

$$e_N(g) = \inf \left(\mathbb{E}\|g - \hat{g}\|^2_{L_2(\mathscr{O})}\right)^{1/2}$$

with the infimum taken over all measurable mappings $\hat{g} : \Omega \to L_2(\mathscr{O})$ such that

$$\eta(\hat{g}(\omega)) \leq N \quad \mathbb{P}\text{-a.s.}$$

The analysis of $e_N(g)$ is based on the Besov regularity of g and the underlying wavelet basis Ψ. For deterministic functions v on $\mathscr{O}$ the error of best N-term approximation with respect to the L_2-norm is defined by

$$\sigma_N(v) = \inf\{\|v - \hat{v}\|_{L_2(\mathscr{O})} : \hat{v} \in L_2(\mathscr{O}), \eta(\hat{v}) \leq N\}. \tag{5.21}$$

Clearly

$$e_N(g) = \left(\mathbb{E}(\sigma_N^2(g))\right)^{1/2}.$$

Theorem 5.9. *For every $\varepsilon > 0$, the error of best N-term approximation satisfies*

$$e_N(g) \preceq \begin{cases} N^{-1/\varepsilon}, & \text{if } \beta = 1 \\ N^{-\frac{\alpha+\beta-1}{2(1-\beta)}+\varepsilon}, & \text{otherwise.} \end{cases}$$

By $f \preceq g$ we indicate that there is a constant $c > 0$ such that $f(x) \leq cg(x)$ for all x. For random functions it is also reasonable to impose a constraint on the average number of nonzero wavelet coefficients only, and to study the error of best average N-term (wavelet) approximation

$$e_N^{\text{avg}}(g) = \inf \left(\mathbb{E}\|g - \hat{g}\|^2_{L_2(\mathscr{O})}\right)^{1/2}$$

with the infimum taken over all measurable mappings $\hat{g} : \Omega \to L_2(\mathscr{O})$ such that

$$\mathbb{E}(\eta(\hat{g})) \leq N.$$

Theorem 5.10. *The error of best average N-term approximation satisfies*

$$e_N^{\text{avg}}(g) \preceq \begin{cases} N^{\frac{\gamma d}{2}} 2^{-\frac{\alpha d N}{2}}, & \text{if } \beta = 1 \\ (\ln N)^{\frac{\gamma d}{2}} N^{-\frac{\alpha+\beta-1}{2(1-\beta)}}, & \text{otherwise.} \end{cases}$$

Remark 5.14. One can show that for $\beta < 1$ the upper bound is sharp.

Remark 5.15. The asymptotic behavior of $e_N^{\mathrm{lin}}(g)$ is essentially determined by the parameter $\alpha + \beta$. According to Theorem 5.7, this quantity also determines the regularity of g in the scale of Sobolev spaces $W_2^s(\mathscr{O})$. The asymptotic behavior of $e_N^{\mathrm{avg}}(g)$ is essentially determined by the parameter $(\alpha + \beta - 1)/(1 - \beta)$, which also determines the regularity of g in the scale (5.20) of Besov spaces $B^s_{\tau,\tau}(\mathscr{O})$, according to Corollary 5.1. For $\beta \in (0, 1]$ nonlinear approximation is superior to linear approximation.

5.4.2 Nonlinear Approximation for Elliptic Boundary Value Problems

We are interested in best N-term wavelet approximation for elliptic boundary value problems with random right-hand sides. As a particular, but nevertheless very important, model problem we are concerned with the Poisson equation (5.16), where the right-hand side $g : \Omega \to L_2(\mathscr{O}) \subset W_2^{-1}(\mathscr{O})$ is a random function as described in Sect. 5.4.1. However, g is given as expansion in the dual basis $\tilde{\Psi}$, on which we impose the same assumptions as in Sect. 5.2.2 on the basis Ψ. Analogous to (5.21) we introduce

$$\sigma_{N,W_2^1(\mathscr{O})}(u) = \inf\{\|u - \hat{u}\|_{W_2^1(\mathscr{O})} : \hat{u} \in W_2^1(\mathscr{O}), \eta(\hat{u}) \le N\}.$$

The quantity $\sigma_{N,W_2^1(\mathscr{O})}(u(\omega))$, where $u(\omega)$ is the exact solution to (5.16), serves as benchmark for the performance of the adaptive wavelet algorithms. In order to analyze the power of these algorithms in the stochastic setting, we investigate the error

$$e_{N,W_2^1(\mathscr{O})}(u) = \inf(\mathbb{E}\|u - \hat{u}\|^2_{W_2^1(\mathscr{O})})^{1/2},$$

with the infimum taken over all measurable mappings $\hat{u} : \Omega \to W_2^1(\mathscr{O})$ such that

$$\eta(\hat{u}(\omega)) \le N \quad \mathbb{P}\text{-a.s.}$$

Clearly

$$e_{N,W_2^1(\mathscr{O})}(u) = \left(\mathbb{E}(\sigma^2_{N,W_2^1(\mathscr{O})}(u))\right)^{1/2}.$$

Theorem 5.11. *Suppose that $d \in \{2, 3\}$ and that the right-hand side g in (5.16) is of the form (5.17). Put*

$$\varrho = \min\left(\frac{1}{2(d-1)}, \frac{\alpha + \beta - 1}{6} + \frac{2}{3d}\right).$$

Then, for every $\varepsilon > 0$, the error of the best N-term approximation satisfies

$$e_{N,W_2^1(\mathscr{O})}(u) \preceq N^{-\varrho+\varepsilon}.$$

In case $\mathscr{O}$ is a $\mathscr{C}^\infty$-domain, the problem is completely regular. Therefore, similar to Corollary 5.1, it is remarkable that an arbitrarily high order of convergence (subject to the maximal approximation order afforded by the wavelet basis Ψ) can be realized by choosing β close to one.

Theorem 5.12. *Suppose that $\mathscr{O}$ is a bounded $\mathscr{C}^\infty$-domain in $\mathbb{R}^d$ and that the right-hand side g in (5.16) is of the form (5.17). Moreover, assume that $\beta < 1$ and put*

$$\varrho = \frac{1}{1-\beta}\left(\frac{\alpha-1}{2}+\beta\right)+\frac{1}{d}.$$

Then, for every $\varepsilon > 0$, the error of best N-term approximation satisfies

$$e_{N,W_2^1(\mathscr{O})}(u) \preceq N^{-\varrho+\varepsilon}.$$

5.4.2.1 Numerical Experiments

Here, we illustrate the impact of α and β on approximation rates of elliptic stochastic equations as outlined in Sect. 5.4.2. In the one-dimensional case the equation is given by

$$-u''(\cdot,\omega) = g(\cdot,\omega), \qquad u(0,\omega) = u(1,\omega) = 0 \qquad \text{on } \mathscr{O} = [0,1].$$

The numerical experiment is carried out and evaluated as follows. On input $\delta > 0$ the adaptive wavelet scheme, see [11–13], computes an N-term approximation $\hat{u}(\cdot,\omega)$ to $u(\cdot,\omega)$, whose error with respect to the W_2^1-norm is at most δ. The number N of terms depends on δ as well as on ω via the right-hand side $g(\cdot,\omega)$. We determine $u(\cdot,\omega)$ in a master computation with very high accuracy and then use the norm equivalence (5.6) for the space $W_2^1(\mathscr{O}) = B^s_{2,2}(\mathscr{O})$. The master computation employs a uniform approximation with refinement level $j_1 = 11$. To get a reliable estimate for the average number $\mathbb{E}(\eta(\hat{u}))$ of nonzero wavelet coefficients of $\hat{u}$ and for the error $(\mathbb{E}\|u-\hat{u}\|^2_{W_2^1(\mathscr{O})})^{1/2}$ we use 1,000 independent samples of right-hand sides. This procedure is carried out for 18 different values of δ; the results are presented together with a regression line, whose slope yields an estimate for the order of convergence. For the uniform scheme, we use 1,000 independent samples for 6 different refinement levels, $j = 4,\dots,9$. We add that confidence intervals for the level 0.95 are of length less than 3 % of the estimate in all cases.

In the first experiment we choose $\alpha = 0.9, \beta = 0.2$, i.e., the right-hand side is contained in $W_2^s(\mathscr{O})$ only for $s < 0.05$. Consequently, since in the univariate case the problem is completely regular, the solution is contained in $W_2^s(\mathscr{O})$ with

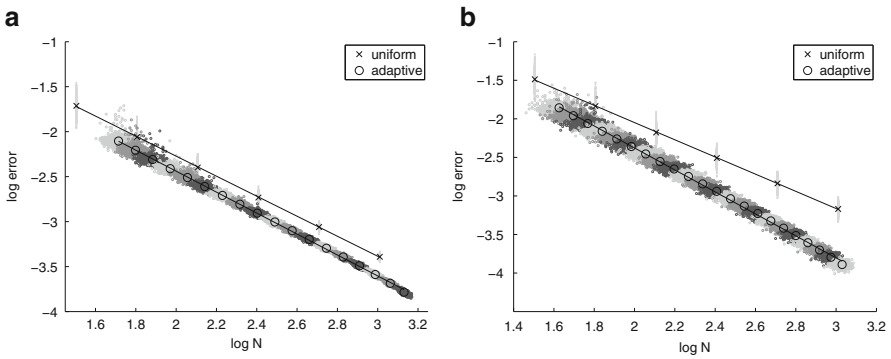


Fig. 5.2 Error and expected number of nonzero coefficients. (**a**) $\alpha = 0.9$, $\beta = 0.2$. (**b**) $\alpha = 0.4$, $\beta = 0.7$

$s < 2.05$. An optimal uniform approximation scheme with respect to the $W_2^1(\mathcal{O})$-norm yields the approximation order $1.05 - \varepsilon$ for every $\varepsilon > 0$. This is observed in Fig. 5.2a, where the empirical order of convergence for the uniform approximation is 1.113. For the relatively small value of $\beta = 0.2$, the Besov smoothness, and therefore the order of best N-term approximation, is not much higher. In fact, by inserting the parameters into Theorem 5.12 with $d = 1$, we get the approximation order $\varrho - \varepsilon$ with $\varrho = 19/16 = 1.1875$. This is also reflected in Fig. 5.2a, where the empirical order of convergence for the adaptive wavelet scheme is 1.164. In both cases the numerical results match very well the asymptotic error analysis, and both methods exhibit almost the same order of convergence. Let us point out, that even in this case adaptivity slightly pays off for the same regularity parameter, since the Besov norm is smaller than the Sobolev norm, which yields smaller constants.

The picture changes for higher values of β. As a second test case, we choose $\alpha = 0.4, \beta = 0.7$. Then, the Besov regularity is considerably higher. In fact, from Theorem 5.12 with $d = 1$ we expect the convergence rate $\varrho - \varepsilon$ with $\varrho = 7/3$, provided that the wavelet basis indeed characterizes the corresponding Besov spaces. It is well known that a tensor product spline wavelet basis of order m in dimension d has this very property for $B^s_{\tau,\tau}(\mathcal{O})$ with $1/\tau = s - 1/2$ and $s < s_1 = m/d$. In our case, $s_1 = 3$, so $\varrho = 2$ is the best we can expect. From Fig. 5.2b, we observe that the empirical order of convergence is slightly lower, namely 1.425. The reason is that the Besov smoothness of the solution is only induced by the right-hand side, which, in a Galerkin approach, is expanded in the dual wavelet basis. Estimating the Hölder regularity of the dual wavelet basis $\tilde{\Psi}$, it turns out that this wavelet basis is only contained in $W^s_\infty(\mathcal{O})$ for $s < 0.55$. Therefore, by using classical embeddings of Besov spaces, it is only ensured that this wavelet basis characterizes Besov spaces $B^s_{\tau,\tau}(\mathcal{O})$, with the same smoothness parameter. Consequently, the solution u, which is obtained by the master computation, is only contained in the spaces $B^s_{\tau,\tau}(\mathcal{O})$ with $1/\tau = s - 1/2$ and $s < 2.55$ which gives an approximation order $\varrho - \varepsilon$ with $\varrho = 1.55$. This is captured very well in Fig. 5.2b.

For uniform approximation the empirical order of convergence is 1.115 and thus does not differ from the result in the first experiment.

References

1. Abramovich, F., Sapatinas, T., Silverman, B.W.: Wavelet thresholding via a Bayesian approach. J. R. Stat. Soc., Ser. B, Stat. Methodol. **60**, 725–749 (1998)
2. Bochkina, N.: Besov regularity of functions with sparse random wavelet coefficients (2006, unpublished preprint). arXiv:1310.3720
3. Canuto, C., Tabacco, A., Urban, K.: The wavelet element method. I: construction and analysis. Appl. Comput. Harmon. Anal. **6**, 1–52 (1999)
4. Cioica, P.A., Dahlke, S.: Spatial Besov regularity for semilinear stochastic partial differential equations on bounded Lipschitz domains. Int. J. Comput. Math. **89**, 2443–2459 (2012)
5. Cioica, P.A., Dahlke, S., Döhring, N., Friedrich, U., Kinzel, S., Lindner, F., Raasch, T., Ritter, K., Schilling, R.L.: On the convergence analysis of Rothe's method, DFG-SPP 1324. **124** (2012, preprint)
6. Cioica, P.A., Dahlke, S., Döhring, N., Friedrich, U., Kinzel, S., Lindner, F., Raasch, T., Ritter, K., Schilling, R.L.: Convergence analysis of spatially adaptive Rothe methods. Found. Comput. Math. (2014). doi: 10.1007/s10208-013-9183-7
7. Cioica, P.A., Dahlke, S., Döhring, N., Kinzel, S., Lindner, F.,Raasch, T., Ritter K., Schilling, R.L.: Adaptive wavelet methods for the stochastic Poisson equation. BIT **52**, 589–614 (2011)
8. Cioica, P.A., Dahlke, S., Kinzel, S., Lindner, F., Raasch, T., Ritter, K., Schilling, R.L.: Spatial Besov regularity for stochastic partial differential equations on Lipschitz domains. Studia Math. **207**, 197–234 (2011)
9. Cioica, P.A., Kim, K.-H., Lee, K., Lindner, F.: On the $L_q(L_p)$-regularity and Besov smoothness of stochastic parabolic equations on bounded Lipschitz domains. Electron. J. Probab. **18**, 1–41 (2013)
10. Cohen, A.: Numerical Analysis of Wavelet Methods. Studies in Mathematics Applications, vol. 32, 1st edn. Elsevier, Amsterdam (2003)
11. Cohen, A., Dahmen, W., DeVore, R.A.: Adaptive wavelet methods for elliptic operator equations: convergence rates. Math. Comput. **70**, 27–75 (2001)
12. Cohen, A., Dahmen, W., DeVore, R.A.: Adaptive wavelet methods II: beyond the elliptic case. Found. Comput. Math. **2**, 203–245 (2002)
13. Dahlke, S., Dahmen, W., DeVore, R.A.: Nonlinear approximation and adaptive techniques for solving elliptic operator equations. In: Multiscale Wavelet Methods for Partial Differential Equations, pp. 237–284. Academic, San Diego (1997)
14. Dahlke, S., DeVore, R.A.: Besov regularity for elliptic boundary value problems. Commun. Partial Differ. Equ. **22**, 1–16 (1997)
15. Dahmen, W., Kunoth, A., Urban, K.: Biorthogonal spline wavelets on the interval – stability and moment conditions. Appl. Comput. Harmon. Anal. **6**, 132–196 (1999)
16. Dahmen, W., Schneider, R.: Composite wavelet bases for operator equations. Math. Comput. **68**, 1533–1567 (1999)
17. Dahmen, W., Schneider, R.: Wavelets on manifolds I: construction and domain decomposition. SIAM J. Math. Anal. **31**, 184–230 (1999)
18. Da Prato, G., Zabczyk, J.: Stochastic Equations in Infinite Dimensions. Encyclopedia of Mathematics and Its Applications, vol. 44. Cambridge University Press, Cambridge (1992)
19. DeVore, R.A.: Nonlinear approximation. Acta Numer. **8**, 51–150 (1998)
20. Fujiwara, D.: Concrete characterization of the domains of fractional powers of some elliptic differential operators of the second order. Proc. Jpn. Acad. **43**, 82–86 (1967)

21. Gyöngy, I., Millet, A.: Rate of convergence of space time approximations for stochastic evolution equations. Potential Anal. **30**, 29–64 (2008)
22. Kim, K.-H.: On stochastic partial differential equations with variable coefficients in C^1 domains. Stoch. Proc. Appl. **112**, 261–283 (2004)
23. Kim, K.-H.: A weighted Sobolev space theory of parabolic stochastic PDEs on non-smooth domains. J. Theor. Probab. **27**, 107–136 (2014)
24. Kovács, M., Larsson, S., Lindgren, F.: Spatial approximation of stochastic convolutions. J. Comput. Appl. Math. **235**, 3554–3570 (2011)
25. Kovács, M., Larsson, S., Urban, K.: On Wavelet-Galerkin Methods for Semilinear Parabolic Equations with Additive Noise. Springer Proceedings in Mathematics and Statistics, vol. 65, pp. 481–499. Springer, Berlin (2013)
26. Krylov, N.V.: A W_2^n-theory of the Dirichlet problem for SPDEs in general smooth domains. Probab. Theory Relat. Fields. **98**, 389–421 (1994)
27. Krylov, N.V.: An analytic approach to SPDEs. In: Carmona, R., Rozovskii, B.L. (eds.) Stochastic Partial Differential Equations: Six Perspectives, pp. 185–242. American Mathematical Society, Providence, RI (1999)
28. Krylov, N.V.: SPDEs in $L_q((0, \tau], L_p)$ spaces. Electron. J. Probab. **5**, 1–29 (2000)
29. Krylov, N.V.: Some properties of traces for stochastic and deterministic parabolic weighted Sobolev spaces. J. Funct. Anal. **183**, 1–41 (2001)
30. Krylov, N.V., Lototsky, S.V.: A Sobolev space theory of SPDE with constant coefficients on a half line. SIAM J. Math. Anal. **30**, 298–325 (1999)
31. Krylov, N.V., Lototsky, S.V. : A Sobolev space theory of SPDEs with constant coefficients in a half space. SIAM J. Math. Anal. **31**, 19–33 (1999)
32. Lindner, F.: Singular behavior of the solution to the stochastic heat equation on polygonal domain. Stoch. PDE: Anal. Comp. (2014). doi:10.1007/s40072-014-0030-x
33. Lototsky, S.V.: Sobolev spaces with weights in domains and boundary value problems for degenerate elliptic equations. Methods Appl. Anal. **7**, 195–204 (2000)
34. Müller-Gronbach, T., Ritter, K.: An implicit Euler scheme with non-uniform time discretization for heat equations with multiplicative noise. BIT **47**, 393–418 (2007)
35. Primbs, M.: New stable biorthogonal spline-wavelets on the interval. Results Math. **57**, 121–162 (2010)
36. van Neerven, J., Veraar, M.C., Weis, L.: Maximal L^p-regularity for stochastic evolution equations. SIAM J. Math. Anal. **44**, 1372–1414 (2012)
37. Wood, I.: Maximal L_p-regularity for the Laplacian on Lipschitz domains. Math. Z. **255**, 855–875 (2007)

Chapter 6
Constructive Quantization and Multilevel Algorithms for Quadrature of Stochastic Differential Equations

Martin Altmayer, Steffen Dereich, Sangmeng Li, Thomas Müller-Gronbach, Andreas Neuenkirch, Klaus Ritter, and Larisa Yaroslavtseva

Abstract In this article we summarise the progress made in the project *Constructive Quantization and Multilevel Algorithms for Quadrature of Stochastic Differential Equations*. Research was conducted along the following three lines. First we focus on deterministic quadrature formulas to approximate expectations with respect to marginal distributions of SDEs. Here we provide a complexity analysis for deterministic algorithms in a worst case setting with respect to classes of SDEs that are defined in terms of smoothness constraints on the coefficients, and we present an algorithm that is based on weak Itô-Talyor steps and performs almost asymptotically optimal. Next, we are concerned with computing expectations of quantities that depend discontinuously on the SDE at the terminal time. We present an efficient method for quadrature in the Heston model based on multilevel schemes and a Malliavin calculus-based payoff smoothing. Finally, we consider expected values of quantities that depend on the whole trajectory of a Lévy-driven SDE. We establish error estimates and central limit theorems for a multilevel Monte Carlo algorithm that achieves error rates of order $N^{-\frac{1}{2}+o(1)}$ as the runtime N of the algorithm tends to infinity.

M. Altmayer (✉) • A. Neuenkirch
University of Mannheim, A5,6, 68131 Mannheim, Germany
e-mail: altmayer@uni-mannheim.de; neuenkirch@kiwi.math.uni-mannheim.de

S. Dereich • S. Li
WWU Münster, Einsteinstr. 62, 48149, Münster, Germany
e-mail: steffen.dereich@wwu.de; li.sangmeng@wwu.de

T. Müller-Gronbach • L. Yaroslavtseva
University of Passau, 94032 Passau, Germany
e-mail: thomas.mueller-gronbach@uni-passau.de; larisa.yaroslavtseva@uni-passau.de

K. Ritter
Technical University of Kaiserslautern, Postfach 3049, 67653, Kaiserslautern, Germany
e-mail: ritter@mathematik.uni-kl.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_6

6.1 Introduction

In this article we introduce and analyse new methods for the numerical computation of

$$S(f) = \mathbf{E}[f(X)],$$

where $X = (X_t)_{t\in[0,T]}$ is a solution to the stochastic differential equation

$$\begin{aligned} dX_t &= a(X_{t-})\,dt + b(X_{t-})\,dY_t, \qquad t \in [0,T], \\ X_0 &= x_0. \end{aligned} \tag{6.1}$$

Here $Y = (Y_t)_{t\in[0,T]}$ is a Wiener process or a Lévy process depending on the branch of the project, x_0 is the deterministic initial value, and a and b are, in general, vector-valued and matrix-valued functions, respectively, that satisfy appropriate regularity conditions. In general we distinguish two kinds of settings. In the *marginal setting/case* the functional f depends only on the value of the path at terminal time T and we rather conceive f as a function on the state space of the differential equation. Conversely in the *path-dependent setting/case* f is a measurable function mapping trajectories into reals. Motivated by financial applications, we shall refer to the function f as *payoff* in the subsequent discussion.

Research was carried out along three different lines and the exposition is arranged accordingly.

In line (I) of the project we consider the marginal setting for a d-dimensional SDE (6.1) driven by an m-dimensional Brownian motion. We study quantization of the distribution $\mathbf{P}_{X_T}$, i.e., construction of quadrature formulas for integration on $\mathbb{R}^d$ with respect to $\mathbf{P}_{X_T}$, by means of deterministic algorithms that are based on finitely many evaluations of the coefficients of the SDE. We provide a worst case complexity analysis with respect to classes of equations (x_0, a, b) and classes of payoffs f that are defined in terms of the degree of smoothness s_1, s_2, r of a, b, f, respectively, and we present an algorithm that performs almost asymptotically optimal in a large number of cases. The main results can be summarised as follows.

- If $m \geq d$ then the minimal errors that can be achieved are of order $N^{-\eta}$ with $\eta = \min(s_1, s_2)/d, r/d, \min(s_1, s_2, r)/d$ for N being the worst case number of evaluations of a and b, the worst case support size of the quadrature formulas and the worst case computational cost, respectively, see Theorem 6.1.
- For coefficients of smoothness at least 6 and Lipschitz continuous payoffs the optimal orders are achieved, up to an arbitrary small power, by a method that combines weak order 2.0 Itô-Taylor steps with strategies for the reduction of the size and the diameter of the support of a discrete measure, see Theorem 6.2 and Remark 6.4.

In line (II) of the project we establish an integration by parts formula for the quadrature of discontinuous payoffs in a multidimensional Heston model. We use

integration by parts of Malliavin calculus to smoothen the original functional. In combination with a payoff-splitting we obtain very efficient multilevel Monte Carlo schemes that achieve errors of order $N^{-\frac{1}{2}+o(1)}$ in the runtime N of the algorithms. We work under mild integrability conditions on the payoff and under a non-negativity assumption on the volatility.

In line (III) of the project we consider the path-dependent and the marginal setting for a Lévy-driven SDE. We consider multilevel schemes that make use of approximations based on Euler schemes with (nondeterministic) jump-adapted update times. The main results can be summarised as follows.

- The root mean squared error achieved by appropriate multilevel schemes is of order $N^{-\frac{1}{2}+o(1)}$ in the runtime N in the path-dependent setting with payoffs that are Lipschitz continuous with respect to the supremum norm, see Theorem 6.6.
- In the case of Lévy-driven SDEs with Gaussian component one has stable convergence of certain error processes, see Theorem 6.7, which implies a central limit theorem for the marginal setting, see Theorem 6.8.

These three lines of research aim at enhancing our understanding of quadrature problems for SDEs in various directions. Although a synthesis of these directions has not yet been carried out we would like to point out a number of potential questions that arise through their interplay.

Line (I) considers deterministic methods for classical stochastic differential equations and it is natural to ask for extensions to the case of discontinuous driving processes as considered in line (III). Line (II) develops specific multilevel methods for classical stochastic differential equations and hence it is natural to consider central limit theorems for these schemes as in line (III). Conversely, the ideas in (II) can be equally well employed for Lévy-driven SDEs with a Gaussian component and it is natural to ask whether a central limit theorem as in (III) holds true for the resulting numerical schemes.

6.2 Constructive Quantization of Systems of SDEs

In this section we summarize the results that have been obtained within line (I) of the project on the construction and the analysis of methods for quantization of systems of SDEs. For proofs and further details regarding Sects. 6.2.1 and 6.2.2 we refer to [35] and [36], respectively.

We consider a d-dimensional equation (6.1) with Y given by an m-dimensional Brownian motion and Lipschitz continuous coefficients a and b, and for simplicity we assume $T = 1$.

The computational task is to approximate the distribution

$$S(x_0, a, b) = \mathbf{P}_{X_1}$$

of the solution X of (6.1) at time $t = 1$ by a probability measure

$$\hat{S}(x_0, a, b) = \sum_{i=1}^{N} c_i \delta_{x_i}$$

on $\mathbb{R}^d$ with finite support based on the initial value x_0 and finitely many evaluations of a and b or derivatives of a and b.

The constructive approximation of marginal distributions of SDEs has been studied in [28–31, 34] with a focus on upper bounds, and we refer to [33] for the analogous problem on the path space. A technique to prove lower bounds has so far been developed only in a particular setting in [40].

In the sequel $M(\mathbb{R}^d)$ and $M_1(\mathbb{R}^d)$ denote the sets of finite measures and probability measures on $\mathbb{R}^d$ with finite support, respectively. For sequences u_n and v_n in $[0, \infty)$ we write $u_n \preceq v_n$ if there exists $c > 0$ such that $u_n \leq c\, v_n$ for all $n \in \mathbb{N}$.

6.2.1 Complexity Analysis of Marginal Quantization

Consider a class $\mathscr{C}$ of equations (x_0, a, b) given by

$$\mathscr{C} = \mathscr{C}_0 \times \mathscr{C}_1 \times \mathscr{C}_2,$$

where $\mathscr{C}_0 \subset \mathbb{R}^d$ and $\mathscr{C}_1$ and $\mathscr{C}_2$ are classes of functions $a\colon \mathbb{R}^d \to \mathbb{R}^d$ and $b\colon \mathbb{R}^d \to \mathbb{R}^{d\times m}$, respectively, that are at least Lipschitz continuous.

We study deterministic algorithms

$$\hat{S}\colon \mathscr{C} \to M_1(\mathbb{R}^d)$$

in the real number model of computation that use x_0 and finitely many evaluations of a and b or derivatives of these functions to compute an approximation $\hat{S}(x_0, a, b) \in M_1(\mathbb{R}^d)$ to the distribution $S(x_0, a, b)$. The class of all deterministic algorithms $\hat{S}$ is denoted by $\mathbf{S}$.

We employ the following three notions of the cost of an algorithm $\hat{S} \in \mathbf{S}$. For any equation $(x_0, a, b) \in \mathscr{C}$ we use $\mathbf{n}(\hat{S}, (x_0, a, b))$ to denote the number of evaluations of $a, b, \partial a_i/\partial x_p, \partial b_{i,j}/\partial x_p$ etc. that are carried out by $\hat{S}$ for this equation and we define the information cost of $\hat{S}$ by

$$\mathrm{cost}^{\mathrm{inf}}(\hat{S}) = \sup\{\mathbf{n}(\hat{S}, (x_0, a, b))\colon (x_0, a, b) \in \mathscr{C}\}.$$

Furthermore,

$$\mathrm{cost}^{\mathrm{supp}}(\hat{S}) = \sup\{\#\mathrm{supp}(\hat{S}(x_0, a, b))\colon (x_0, a, b) \in \mathscr{C}\}$$

is the maximum support size of the probability measures computed by $\hat{S}$. Finally, we consider the total computational cost of $\hat{S}$ given by

$$\mathrm{cost}^{\mathrm{total}}(\hat{S}) = \mathrm{cost}^{\mathrm{inf}}(\hat{S}) + \mathrm{cost}^{\mathrm{supp}}(\hat{S}) + \sup\{\mathrm{op}(\hat{S}(x_0, a, b)) \colon (x_0, a, b) \in \mathscr{C}\},$$

where $\mathrm{op}(\hat{S}(x_0, a, b))$ is the number of all basic computational operations, i.e., arithmetical operations, jumps, assignments and evaluations of elementary functions, that are carried out by $\hat{S}$ for the input (x_0, a, b).

For the definition of the error of $\hat{S}$ we employ a class $\mathscr{F}$ of payoffs $f \colon \mathbb{R}^d \to \mathbb{R}$ that are measurable and satisfy a polynomial growth condition. We use

$$\rho(\hat{S}(x_0, a, b), S(x_0, a, b)) = \sup_{f \in \mathscr{F}} \left\{ \left| \int_{\mathbb{R}^d} f \, d\hat{S}(x_0, a, b) - \int_{\mathbb{R}^d} f \, dS(x_0, a, b) \right| \right\}$$

to quantify the error of each single approximation $\hat{S}(x_0, a, b)$ and we define

$$\mathrm{error}(\hat{S}) = \sup\{\rho(\hat{S}(x_0, a, b), S(x_0, a, b)) \colon (x_0, a, b) \in \mathscr{C}\}.$$

The key quantities of the complexity analysis are the N-th minimal errors

$$e_N^* = e_N^*(\mathscr{C}, \mathscr{F}) = \inf\{\mathrm{error}(\hat{S}) \colon \hat{S} \in \mathbf{S},\ \mathrm{cost}^*(\hat{S}) \le N\},$$

where $* \in \{\mathrm{inf}, \mathrm{supp}, \mathrm{total}\}$.

For $r \in \mathbb{N}$, $K > 0$ and $\beta \ge 0$ we use $\mathscr{F}(r, K, \beta)$ to denote the class of functions $h \in C^r(\mathbb{R}^d; \mathbb{R})$ that satisfy $|h^{(\alpha)}(x)| \le K \cdot (1 + \|x\|_\infty^\beta)$ for every $x \in \mathbb{R}^d$ and every $\alpha \in \mathbb{N}_0^d$ with $1 \le \|\alpha\|_1 \le r$. Moreover, we put

$$\mathscr{F}^0(r, K) = \{h \in \mathscr{F}(r, K, 0) \colon |h(0)| \le K\}.$$

Theorem 6.1. *Assume $m \ge d$. Let $r, s_1, s_2 \in \mathbb{N}$, $K > 0$, $\beta \ge 0$ and consider the class*

$$\mathscr{C} = \mathscr{C}(s_1, s_2, K) = [-K, K]^d \times (\mathscr{F}^0(s_1, K))^d \times (\mathscr{F}^0(s_2, K))^{d \times m} \tag{6.2}$$

of equations and the class $\mathscr{F} = \mathscr{F}(r, K, \beta)$ of payoffs. Then

$$N^{-\eta^*} \preceq e_N^*(\mathscr{C}, \mathscr{F}) \preceq N^{-\eta^* + \varepsilon}$$

for every $\varepsilon > 0$ and $* \in \{\mathrm{inf}, \mathrm{supp}, \mathrm{total}\}$ with

$$\eta^* = \begin{cases} \min(s_1, s_2)/d, & \text{if } * = \mathrm{inf}, \\ r/d, & \text{if } * = \mathrm{supp}, \\ \min(s_1, s_2, r)/d, & \text{if } * = \mathrm{total}. \end{cases}$$

We conjecture that the upper bound in Theorem 6.1 actually holds with $\varepsilon = 0$.

Remark 6.1. The lower bounds in Theorem 6.1 are derived by relating the respective minimal errors to minimal errors for suitably chosen weighted integration problems on $\mathbb{R}^d$ and use results on corresponding lower bounds from [9, 38].

In the particular case of $* = \inf$ we consider the two subclasses of $\mathscr{C}$ given by

$$\mathscr{C}^{(1)} = \{0\} \times (\mathscr{F}^0(s, K))^d \times \{K \cdot E_d\},$$
$$\mathscr{C}^{(2)} = \{0\} \times \{0\}^d \times \{\mathrm{diag}(h) \mid h \in (\mathscr{F}^0(s, K))^d\},$$

where E_d denotes the identity matrix in $\mathbb{R}^{d\times d}$ and $\mathrm{diag}(h)\colon \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is defined by $(h(x))_{i,j} = h_i(x)$ if $i = j$ and $(h(x))_{i,j} = 0$ otherwise. We then essentially follow the approach in [40] and use a series expansion of the Lebesgue density of $\mathbf{P}_{X_1}$ by means of the parametrix method, see [14], to obtain

$$e_N^{\inf}(\mathscr{C}^{(i)}, \{f\}) \succeq N^{-s/d} \tag{6.3}$$

for any non-constant $f \in \mathscr{F}(1, K, \beta)$ in the case $i = 1$ and any non-multilinear $f \in \mathscr{F}(2, K, \beta)$ in the case $i = 2$.

Remark 6.2. Let $\varepsilon > 0$. For the proof of the upper bounds in Theorem 6.1 we construct a sequence of algorithms $\hat{S}_{n,\varepsilon}$ such that $\lim_{n\to\infty} \mathrm{cost}^*(\hat{S}_{n,\varepsilon}) = \infty$ as well as

$$\mathrm{error}(\hat{S}_{n,\varepsilon}) \preceq (\mathrm{cost}^*(\hat{S}_{n,\varepsilon}))^{-\eta^*+\varepsilon} \tag{6.4}$$

simultaneously for $* = \inf$, $* = \mathrm{supp}$ and $* = \mathrm{total}$.

However, implementing $\hat{S}_{n,\varepsilon}$ requires at least $n^{d^2 n^{d\,\min(s_1,s_2)}}$ basic computational operations in a precomputation step, so that these algorithms are of no practical use. The main purpose of constructing and analysing $\hat{S}_{n,\varepsilon}$ is to show that the lower bounds for the minimal errors in Theorem 6.1 are essentially sharp. In Sect. 6.2.2 we present algorithms that do not rely on precomputation, are easy to implement and perform almost asymptotically optimal for a large subclass $\tilde{\mathscr{C}}$ of $\mathscr{C}$ in the particular case of $r = 1$ and $\min(s_1, s_2) \geq 6$, see Theorem 6.2 and Remark 6.4.

Remark 6.3. Assume $m < d$. One can show that the upper bounds in Theorem 6.1 are still valid in this case. Moreover, it is easy to see that the lower bounds for the minimal errors e_N^{supp} in Theorem 6.1 hold true as well. For the minimal errors $e_N^{\inf}$ we only have the lower bounds $c\,N^{-\min(s_1,s_2)/m}$ with a positive constant $c > 0$, up to now. The precise order of convergence of the quantities $e_N^{\inf}$ is unknown to us.

6.2.2 *Marginal Quantization Based on a Weak Itô-Taylor Scheme*

It is natural to ask whether the upper bounds in Theorem 6.1 can be achieved by implementable algorithms that do not rely on heavy precomputation. The answer

to this question is positive, at least, if further restrictions are imposed on the input equations. In the following we assume $m \geq d$ and $\min(s_1, s_2) \geq 6$ in (6.2). Let $\Lambda > 0$ and consider the subclass $\tilde{\mathscr{C}} \subset \mathscr{C}$ of equations given by

$$\tilde{\mathscr{C}} = \tilde{\mathscr{C}}(K, \Lambda) = \{(x_0, a, b) \in \mathscr{C} \colon \|a\|_\infty, \|b\|_\infty \leq K, \inf_{x \in \mathbb{R}^d} \lambda_{\min}(bb^{\mathrm{T}}(x)) \geq \Lambda\},$$

where $\lambda_{\min}(bb^{\mathrm{T}}(x))$ denotes the smallest eigenvalue of the matrix $bb^{\mathrm{T}}(x) \in \mathbb{R}^{d \times d}$.

For every $n \in \mathbb{N}$ and $\varepsilon > 0$ we introduce a deterministic method

$$\tilde{S}_{n,\varepsilon} : \tilde{\mathscr{C}} \to M_1(\mathbb{R}^d)$$

that computes an approximation $\tilde{S}_{n,\varepsilon}(x_0, a, b)$ in the following way. Starting with the one-point measure in x_0 it iteratively applies, with time-steps τ according to a non-uniform discretization of the time interval $[0, 1]$, a linear transition $T_\tau^{a,b}$ on $M_1(\mathbb{R}^d)$, which yields an approximation to the distribution of the solution of the equation (x, a, b) at time τ for any $x \in \mathbb{R}^d$, together with strategies D_τ^ε and R_τ to reduce the diameter and the size of the support of a probability measure in $M_1(\mathbb{R}^d)$, respectively, in order to avoid an explosion of the computational cost. The number of time-steps is given by n and the parameter ε specifies the diameter reduction strategy.

Fix $(x_0, a, b) \in \tilde{\mathscr{C}}$. The definition of the transition $T_\tau^{a,b}$ is based on corresponding simplified weak order 2.0 Itô-Taylor steps $Y_\tau^{x,a,b}$ of length τ starting in $x \in \mathbb{R}^d$ that have been introduced in [43] in the case $d = m = 1$ and extended to arbitrary dimensions $d, m \in \mathbb{N}$ in [27, Sec. 14.2].

Consider independent random variables

$$\xi_i,\ \eta_{j,\ell}, \quad 1 \leq i \leq m,\ 1 \leq j < \ell \leq m,$$

with

$$\xi_i \sim 1/6 \cdot \delta_{-\sqrt{3\tau}} + 2/3 \cdot \delta_0 + 1/6 \cdot \delta_{\sqrt{3\tau}}, \quad \eta_{j,\ell} \sim 1/2 \cdot (\delta_{-\tau} + \delta_\tau).$$

Put $J_{(0),\tau} = \tau$, $J_{(0,0),\tau} = \tau^2/2$ and

$$J_{(i),\tau} = \xi_i,\ J_{(i,0),\tau} = J_{(0,i),\tau} = \tau \cdot \frac{\xi_i}{2},\ J_{(i,j),\tau} = \begin{cases} (\xi_i \xi_j + \eta_{i,j})/2, & \text{if } i < j, \\ (\xi_i \xi_j - \eta_{j,i})/2, & \text{if } i > j, \\ (\xi_i^2 - \tau)/2, & \text{if } i = j, \end{cases}$$

for $i, j \neq 0$, and define

$$Y_\tau^{x,a,b} = x + \sum_{\alpha \in G} \varphi_\alpha^{a,b}(x)\, J_{\alpha,\tau}, \tag{6.5}$$

where $G = \{0,\dots,m\} \cup \{0,\dots,m\}^2$, $\varphi_\alpha^{a,b}\colon \mathbb{R}^d \to \mathbb{R}$ denotes the Itô-Taylor coefficient function corresponding to a, b and $\alpha \in G$ and $J_{\alpha,\tau}$ serves as a discrete approximation to the iterated Itô-integral corresponding to α up to time τ.

We define $T_\tau^{a,b}\colon M_1(\mathbb{R}^d) \to M_1(\mathbb{R}^d)$ by

$$T_\tau^{a,b}(\nu) = \sum_{x \in \mathrm{supp}(\nu)} \nu(\{x\})\, \mathbf{P}_{Y_\tau^{x,a,b}}.$$

For the reduction of the diameter of a support we project points $x \in \mathbb{R}^d$ onto the cube $[-s^{-\varepsilon}, s^{-\varepsilon}]^d$ by taking

$$\lfloor x \rfloor_{\varepsilon,\tau} = \big(x_i + (-x_i - \tau^{-\varepsilon})_+ - (x_i - \tau^{-\varepsilon})_+\big)_{i=1,\dots,d}$$

and we define $D_\tau^\varepsilon\colon M_1(\mathbb{R}^d) \to M_1(\mathbb{R}^d)$ by

$$D_\tau^\varepsilon(\nu) = \sum_{x \in \mathrm{supp}(\nu)} \nu(\{x\}) \cdot \delta_{\lfloor x \rfloor_{\varepsilon,\tau}}.$$

Next we explain the strategy to reduce the size of a support. Let $q \in \mathbb{N}$. By a well-known sequential support point elimination procedure due to [13] we obtain an algorithm $R\colon M(\mathbb{R}^d) \to M(\mathbb{R}^d)$ that satisfies

$$\mathrm{supp}(R(\nu)) \subset \mathrm{supp}(\nu), \quad |\mathrm{supp}(R(\nu))| \le \binom{d+q}{d}$$

as well as

$$\int_{\mathbb{R}^d} p\, dR(\nu) = \int_{\mathbb{R}^d} p\, d\nu$$

for every polynomial p on $\mathbb{R}^d$ up to order q. The number of arithmetical operations needed to carry out one sequential step of the computation of $R(\nu)$ is proportional to $\binom{d+q}{d}^3$. Hence q should be small. On the other hand, q should be large enough in order that the reduced measure $R(\nu)$ stays close to ν with respect to taking expectations of smooth functions. In fact, a direct application of this algorithm would lead to a suboptimal relation of error and cost for any choice of the parameter q. We therefore use the following variant of a localized version of R, which has been introduced in [30] for the Wiener cubature approach. Take $q = 5$, put

$$A_{j,\tau} = \bigtimes_{i=1}^d [j_i\sqrt{\tau},\, (j_i+1)\sqrt{\tau})$$

for $j \in \mathbb{Z}^d$ and define $R_\tau : M_1(\mathbb{R}^d) \to M_1(\mathbb{R}^d)$ by

$$R_\tau(\nu) = \sum_{j\in\mathbb{Z}^d} R\Big(\sum_{x\in \mathrm{supp}(\nu)\cap A_{j,\tau}} \nu(\{x\})\, \delta_x \Big).$$

Finally, we employ the non-uniform time discretization

$$t_i = 1 - (1 - i/n)^3, \quad i = 0, \dots, n,$$

which, in a more general form, has been introduced in [28] and is also used in [30, 31] for the Wiener cubature approach.

Put $\tau_i = t_i - t_{i-1}$ for $i = 1, \dots, n$ and define

$$\tilde{S}_{n,\varepsilon}(x_0, a, b) = R_{\tau_n} \circ D^\varepsilon_{\tau_n} \circ T^{a,b}_{\tau_n} \circ \dots \circ R_{\tau_1} \circ D^\varepsilon_{\tau_1} \circ T^{a,b}_{\tau_1}(\delta_{x_0}).$$

Proposition 6.1. *We have*

$$\mathrm{cost}^*(\tilde{S}_{n,\varepsilon}) \le c n^{3d(1+2\varepsilon)/2}$$

for $* \in \{supp, total\}$ *and*

$$\Big| \int f \, d\tilde{S}_{n,\varepsilon}(x_0, a, b) - \int f \, dS(x_0, a, b) \Big| \le c \cdot \|f\|_{\mathrm{Lip}} \cdot n^{-3/2},$$

for every $(x_0, a, b) \in \tilde{\mathscr{C}}$ *and every Lipschitz continuous* $f : \mathbb{R}^d \to \mathbb{R}$, *where* $c > 0$ *only depends on* $d, m, K, \Lambda, \varepsilon$ *and* $\|f\|_{\mathrm{Lip}}$ *is the Lipschitz semi-norm of* f.

Proposition 6.1 implies the following worst case error estimate of $\tilde{S}_{n,\varepsilon}$.

Theorem 6.2. *For every* $\varepsilon > 0$ *the worst case error of* $\tilde{S}_{n,\varepsilon}$ *on* $\tilde{\mathscr{C}}$ *with respect to the class of payoffs* $\mathscr{F} = \mathscr{F}(1, K, 0)$ *satisfies for* $* \in \{\mathrm{supp}, \mathrm{total}\}$,

$$\mathrm{error}(\tilde{S}_{n,\varepsilon}) \preceq (\mathrm{cost}^*(\tilde{S}_{n,\varepsilon}))^{-\frac{1}{d(1+2\varepsilon)}}. \tag{6.6}$$

Remark 6.4. Using piecewise polynomial interpolation of the coefficients a and b one obtains a modified version of $\tilde{S}_{n,\varepsilon}$, which achieves (6.6) as well as

$$\mathrm{error}(\tilde{S}_{n,\varepsilon}) \preceq (\mathrm{cost}^{\mathrm{inf}}(\tilde{S}_{n,\varepsilon}))^{-\frac{\min(s_1,s_2)}{d(1+2s\varepsilon)}}.$$

It is easy to see that the lower bounds from Theorem 6.1 are still valid with $\mathscr{C}$ replaced by $\tilde{\mathscr{C}}$, i.e.,

$$e^*_N(\tilde{\mathscr{C}}, \mathscr{F}) \succeq \begin{cases} N^{-1/d}, & \text{if } * \in \{\mathrm{supp}, \mathrm{total}\}, \\ N^{-\min(s_1,s_2)/d}, & \text{if } * = \mathrm{inf}. \end{cases}$$

Consequently, the sequence of algorithms $\tilde{S}_{n,\varepsilon}$ performs asymptotically optimal, up to an arbitrary small exponent, with respect to cost* for $* \in \{\text{supp}, \text{inf}, \text{total}\}$.

Remark 6.5. The method $\tilde{S}_{n,\varepsilon}$ is already applicable to equations $(x_0, a, b) \in \mathscr{C}(s_1, s_2, K)$ with $s_1, s_2 \geq 2$ since only Itô-Taylor coefficient functions up to the iteration order 2 are involved in the construction of the underlying weak Itô-Taylor steps, see (6.5). However, it is an open question, whether $\tilde{S}_{n,\varepsilon}$ performs asymptotically optimal also in the case $2 \leq \min(s_1, s_2) < 6$, which can not appropriately be treated by our error analysis techniques. Similarly, it is unclear to us, whether the optimality properties of $\tilde{S}_{n,\varepsilon}$ carry over to classes of payoffs $\mathscr{F}(r, K, \beta)$ with $r > 1$. We conjecture that in the latter case asymptotically optimal algorithms can be constructed by using simplified Itô-Taylor steps of higher weak order if the coefficients of the equation are sufficiently smooth.

6.3 Multilevel Methods for Discontinuous Payoffs in the Generalized Heston Model

In this part of the project we aim to compute expectations of discontinuous path-independent functionals in the generalized Heston model, which is a very popular stochastic volatility model in mathematical finance. Thus the goal is to efficiently compute

$$S(f) = \mathbf{E}[f(X_T)],$$

where $X = (X_t)_{t\in[0,T]}$ is the generalized Heston price process. An efficient method for Lipschitz continuous functionals is the multilevel Monte Carlo method, see [22, 24], and also Sect. 1.3 in the context of Lévy driven SDEs. Combining approximations using different step-sizes in a way that reduces the overall variance this method usually is significantly more efficient than standard Monte Carlo. However, the method requires a good L^2-convergence rate for the approximations which is often not easy to achieve for discontinuous functionals, see [8, 23].

To overcome this problem we use the integration by parts formula from Malliavin calculus to replace the discontinuous functional by a continuous one, i.e.

$$S(f) = \mathbf{E}[f(X_T)] = \mathbf{E}[G(X_T) \cdot \Pi],$$

where G is Lipschitz continuous and Π is a stochastic weight term (see Theorem 6.4 below). Combined with a payoff-splitting to reduce the variance of the weight, this estimator outperforms the direct multilevel Monte Carlo estimator for $S(f)$ in our numerical experiments. This is also supported by the error analysis we have carried out so far.

The underlying SDE for the Heston model has non-Lipschitz coefficients. Thus neither are smoothness assumptions of Sects. 1.1 and 1.3 satisfied here nor are

standard results for the numerical analysis of SDEs applicable, see e.g. [27]. We mention that discontinuous functionals also appear when estimating distribution functions which is the topic of [21].

6.3.1 Malliavin Calculus

Malliavin calculus adds a derivative operator to stochastic analysis. Basically, if X is a random variable and $(W_t)_{t\in[0,T]}$ a d-dimensional Brownian motion, then the Malliavin derivative measures the dependence of X on W. The Malliavin derivative is defined by a standard extension procedure: Let $\mathscr{S}$ be the set of smooth random variables of the form

$$S = \varphi\left(\int_0^T h_1(s)dW_s, \dots, \int_0^T h_k(s)dW_s\right)$$

with $h_i \in L^2([0,T];\mathbb{R}^d)$, $i = 1,\dots,k$, $\int_0^T h_j(s)dW(s) = \sum_{i=1}^d h_j^{(i)} dW_s^{(i)}$ and $\varphi \in C^\infty(\mathbb{R}^k;\mathbb{R})$ bounded with bounded derivatives. The derivative operator D of such a smooth random variable is defined as

$$DS = \sum_{i=1}^k \frac{\partial\varphi}{\partial x_i}\left(\int_0^T h_1(s)dW_s, \dots \int_0^T h_k(s)dW_s\right) h_i.$$

This operator is closable from $L^p(\Omega)$ into $L^p\big(\Omega; L^2([0,T];\mathbb{R}^d)\big)$ and the Sobolev space $\mathbf{D}^{1,p}$ denotes the closure of $\mathscr{S}$ with respect to the norm

$$\|X\|_{1,p}^p := \mathbf{E}|X|^p + \mathbf{E}\left|\int_0^T |D_s X|^2 ds\right|^p.$$

In particular, if $D^{(i)}$ denotes the i-th component of the Malliavin derivative, i.e. the derivative with respect to $W^{(i)}$, we have

$$D^{(i)} W_t^{(j)} = \begin{cases} 1_{[0,t]} & \text{if} \quad i = j, \\ 0 & \text{else,} \end{cases}$$

for $i,j = 1,\dots,d$. The derivative operator follows rules similar to ordinary calculus. For example, for a random variable $X \in \mathbf{D}^{1,2}$ and $g \in C^1(\mathbb{R};\mathbb{R})$ with bounded derivative the chain rule reads as

$$Dg(X) = g'(X)\, DX. \tag{6.7}$$

The divergence operator δ is the adjoint of the derivative operator. If a random variable $u \in L^2\big(\Omega; L^2([0,T];\mathbb{R}^d)\big)$ belongs to $\mathrm{dom}(\delta)$, the domain of the divergence operator, then $\delta(u)$ is defined by the duality (or integration by parts) relationship

$$\mathbf{E}[X\delta(u)] = \mathbf{E}\left[\int_0^T \langle D_s X, u_s\rangle ds\right] \qquad \text{for all} \quad X \in \mathbf{D}^{1,2}. \tag{6.8}$$

Moreover, if $u \in \mathrm{dom}(\delta)$ and $X \in \mathbf{D}^{1,2}$ such that $Xu \in L^2\big(\Omega; L^2([0,T];\mathbb{R}^d)\big)$, then we have the following formula:

$$\delta(Xu) = X\delta(u) - \int_0^T \langle D_s X, u_s\rangle ds. \tag{6.9}$$

For more details on Malliavin calculus we refer to [39].

6.3.2 Generalized Heston Model

We study a generalized version of Heston's stochastic volatility model, where the volatility $v = (v_t)_{t\in[0,T]}$ is given by the SDE

$$\mathrm{d}v_t = \kappa(\lambda - v_t)\mathrm{d}t + \theta v_t^{\gamma}\mathrm{d}W_t^{(1)}$$

with parameters $\kappa, \lambda, \theta > 0$, $\gamma \in [1/2, 1)$, initial value $v_0 > 0$ and a Brownian motion $W^{(1)}$. In the case $\gamma = 1/2$ the volatility process is a Cox-Ingersoll-Ross process (CIR) and this leads to the standard Heston model, while for $\gamma > 1/2$ the volatility process is known as mean-reverting constant elasticity of variance process (CEV) [5]. It turns out that for our integration by parts procedure it is necessary that the volatility processes have strictly positive sample paths. In the CEV case this requires no further condition, but in the CIR case we need the parameters to satisfy $2\kappa\lambda > \theta^2$ which is however often fulfilled in practice (see e.g. [1, 19]). In both cases the price process is given by

$$\mathrm{d}X_t = bX_t\mathrm{d}t + \sqrt{v_t}X_t\mathrm{d}\Big(\rho W_t^{(1)} + \sqrt{1-\rho^2}W_t^{(2)}\Big),$$

where $b \geq 0$, $\rho \in [-1, 1]$ and $W^{(2)}$ is a Brownian motion independent of $W^{(1)}$.

Often it will be easier to consider the transformed processes $\sigma_t := v_t^{1-\gamma}$ and $Z_t := \log(X_t)$:

$$\begin{aligned}
\mathrm{d}\sigma_t &= (1-\gamma)\Big(\kappa\lambda\sigma_t^{-\frac{\gamma}{1-\gamma}} - \kappa\sigma_t - \frac{\gamma\theta^2}{2}\sigma_t^{-1}\Big)\mathrm{d}t + \theta(1-\gamma)\mathrm{d}W_t,\\
\mathrm{d}Z_t &= \Big(b - \frac{1}{2}v_t\Big)\mathrm{d}t + \theta\sqrt{v_t}\mathrm{d}\Big(\rho W_t^{(1)} + \sqrt{1-\rho^2}W_t^{(2)}\Big).
\end{aligned}$$

Neither the original nor the transformed processes satisfy the global Lipschitz assumption for its coefficients that is required by most standard results for the numerical approximation of SDEs as well as in Malliavin calculus. However,

by approximating σ and Z by suitable processes which do fulfill this crucial assumption, we can prove that v_t and X_t are in fact Malliavin differentiable and can explicitly calculate their derivatives. It turns out that the derivative of the Heston price with respect to $W^{(2)}$ is particularly simple:

Theorem 6.3 ([4, Theorem 4.3]). *If $X_T \in L^{p+\varepsilon}(\Omega)$ for some $p \geq 1$, $\varepsilon > 0$, then $X_T \in \mathbf{D}^{1,p}$ and*

$$D_t^{(2)} X_T = X_T \cdot D_t^{(2)} Z_T = \sqrt{1-\rho^2} X_T \sqrt{v_t}, \qquad t \in [0,T].$$

Note that the integrability condition on the Heston price can be expressed in terms of its parameters, see e.g. [5]. For conciseness we omit this.

6.3.3 Quadrature Formula

The Malliavin differentiability of X_T allows us to replace a functional f by one of its antiderivatives F:

Theorem 6.4 ([4, Theorem 5.3]). *Assume that $f : [0,\infty) \to \mathbb{R}$ is measurable and of at most quadratic growth. Let $F : [0,\infty) \to \mathbb{R}$ be given by $F(x) := \int_0^x f(z)\mathrm{d}z$. Assume that either $\gamma > 1/2$ or $2\kappa\lambda > \theta^2$ and that there is an $\varepsilon > 0$ such that $X_T \in L^{2+\varepsilon}(\Omega)$. Then*

$$\mathbf{E}[f(X_T)] = \mathbf{E}\left[\frac{F(X_T)}{X_T} \cdot \left(1 + \frac{1}{\sqrt{1-\rho^2}T} \cdot \int_0^T \frac{1}{\sqrt{v_t}}\, \mathrm{d}W_t^{(2)}\right)\right].$$

The idea of the proof is as follows: To replace f by F we use the chain rule (6.7) in the 2nd equality and the integration by parts formula (6.8) in the 3rd equality:

$$\begin{aligned}
\mathbf{E}[f(X_T)] &= \frac{1}{T} \cdot \mathbf{E}\left[\int_0^T f(X_T) \cdot D_t^{(2)} X_T \cdot \frac{1}{D_t^{(2)} X_T} \mathrm{d}t\right] \\
&= \frac{1}{T} \cdot \mathbf{E}\left[\int_0^T D_t^{(2)}(F(X_T)) \cdot \frac{1}{D_t^{(2)} X_T} \mathrm{d}t\right] \\
&= \frac{1}{T} \cdot \mathbf{E}\left[F(X_T) \cdot \delta^{(2)}\left(\frac{1}{D^{(2)} X_T}\right)\right].
\end{aligned}$$

Now it remains to compute the Skorohod integral which is possible thanks to the simple form of $D_t^{(2)} X_T$ and formula (6.9).

For digital options, i.e. $f = \mathbf{1}_{[0,K]}$ for some $K > 0$, the resulting functional $F(x)/x$ is bounded and globally Lipschitz continuous. The price to pay is that the weight term $\Pi = 1 + (\sqrt{1-\rho^2}T)^{-1} \int_0^T v_t^{-1/2} \mathrm{d}W_t^{(2)}$ typically has a very high

variance due to the usually low average volatility λ. To solve this problem, we split discontinuous payoffs into $f = g + h$ with a Lipschitz continuous part g and a discontinuous part with small support h. The quadrature formula is then applied only to the discontinuous part:

$$\mathbf{E}[f(X_T)] = \mathbf{E}\left[g(X_T) + \frac{H(X_T)}{X_T} \cdot \left(1 + \frac{1}{\sqrt{1-\rho^2}T} \cdot \int_0^T \frac{1}{\sqrt{v_t}}\, \mathrm{d}W_t^{(2)}\right)\right],$$

where $H(x) = \int_0^x h(z)\mathrm{d}z$.

6.3.4 Multidimensional Heston Models

A multidimensional Heston model consists of several one-dimensional models which are correlated via the correlation of the driving Brownian motions. For $i = 1, \ldots, d$ the i-th volatility and price process are given as

$$\begin{aligned}
\mathrm{d}v_t^{(i)} &= \kappa_i(\lambda_i - v_t^{(i)})\mathrm{d}t + \theta_i (v_t^{(i)})^{\gamma_i}\mathrm{d}W_t^{(i,1)}, \\
\mathrm{d}X_t^{(i)} &= b_i X_t^{(i)}\mathrm{d}t + \sqrt{v_t^{(i)}} X_t^{(i)}\mathrm{d}W_t^{(i,2)},
\end{aligned}$$

where $W^{(i,1)}, W^{(i,2)}$, $i = 1, \ldots, d$, can be arbitrarily correlated Brownian motions, as long as their covariance matrix is positive definite. In this case there exists an invertible upper $2d \times 2d$ triangular matrix R such that

$$Z = R^{-1}\left(W^{(1,1)}, \ldots, W^{(d,1)}, W^{(1,2)}, \ldots, W^{(d,2)}\right)^*$$

is a $2d$-standard Brownian motion. Then a slightly modified quadrature formula holds:

Theorem 6.5 ([4, Theorem 6.5]). *Assume that $f\colon [0,\infty)^d \to \mathbb{R}$ is measurable and of at most quadratic growth. Let $F\colon [0,\infty)^d \to \mathbb{R}$ be given by*

$$F(x) := \int_0^{x_1} f(\xi, x_2, \ldots, x_d)\mathrm{d}\xi.$$

Assume that for each $j = 1, \ldots, d$ either $\gamma_j > 1/2$ or $2\kappa_j\lambda_j > \theta_j^2$ and that there is an $\varepsilon > 0$ such that $X_T^{(j)} \in L^{2+\varepsilon}(\Omega)$ for all $j = 1, \ldots, d$. Then

$$\mathbf{E}[f(X_T)] = \mathbf{E}\left[\frac{F(X_T)}{X_T^{(1)}} \cdot \left(1 + \frac{1}{R_{11}T} \cdot \int_0^T \frac{1}{\sqrt{v_t^{(1)}}}\, \mathrm{d}Z_t^{(1)}\right)\right].$$

Note that a reordering of the price processes allows to apply the integration by parts procedure to arbitrary components of f.

6.3.5 Discretization

To approximate v we use the drift-implicit Euler scheme for σ because it is easy to implement, preserves positivity and leads to a strong convergence order of one for the approximation of v at fixed time points as shown in e.g. [3, 37]. Note that this is the only scheme for v for which such a sharp convergence result is known at the moment. For simplicity we consider only the one-dimensional case here. The scheme is given by $\hat{\sigma}_0 := v_0^{1-\gamma}$ and

$$\hat{\sigma}_{k+1} := \hat{\sigma}_k + (1-\gamma)\Big(\kappa\lambda\hat{\sigma}_{k+1}^{-\frac{\gamma}{1-\gamma}} - \kappa\hat{\sigma}_{k+1} + \frac{\theta^2\gamma}{2}\hat{\sigma}_{k+1}^{-1}\Big)\Delta + (1-\gamma)\theta\Delta_k W^{(1)},$$

$$\hat{v}_k := \hat{\sigma}_k^{\frac{1}{1-\gamma}}$$

with $\Delta_k W^{(1)} := W^{(1)}_{(k+1)\Delta} - W^{(1)}_{k\Delta}$ and $\Delta > 0$. For the Heston model, $\gamma = 1/2$, the implicit equation for $\hat{\sigma}_{k+1}$ can be solved explicitly while in the general case it can be solved using standard root-finding methods.

In any case the price is then approximated using Euler's method on the log-price, i.e. $\hat{Z}_0 := \log(X_0)$ and

$$\hat{Z}_{k+1} := \hat{Z}_k + \Big(b - \frac{1}{2}\hat{v}_k\Big)\Delta + \sqrt{\hat{v}_k}\Big(\rho\Delta_k W^{(1)} + \sqrt{1-\rho^2}\Delta_k W^{(2)}\Big),$$

$$\hat{X}_k := e^{\hat{Z}_k}.$$

Using these approximations we can formulate two estimators. The first uses the discontinuous functional directly, while the second one uses the quadrature formula and payoff-splitting.

$$\hat{P}^1 = f(\hat{X}_{\lfloor T/\Delta\rfloor}),$$

$$\hat{P}^2 = g(\hat{X}_{\lfloor T/\Delta\rfloor}) + \frac{H(\hat{X}_{\lfloor T/\Delta\rfloor})}{\hat{X}_{\lfloor T/\Delta\rfloor}} \cdot \left(1 + \frac{1}{\sqrt{1-\rho^2}\,T} \cdot \sum_{k=0}^{\lfloor T/\Delta\rfloor} \frac{1}{\sqrt{\hat{v}_k}}\Delta_k W^{(2)}\right).$$

To use multilevel Monte Carlo good L^2-convergence properties are important, see [22], which are provided by the following result for our estimator based on the integration by parts procedure.

Proposition 6.2. *Let $g, h\colon [0,\infty) \to \mathbb{R}$ be measurable and bounded with g being Lipschitz-continuous. If $\gamma = \frac{1}{2}$ and $2\kappa\lambda/\theta^2 > 3$ or $\gamma > 1/2$, then*

$$\mathbf{E}\left|\hat{P}^2 - \left(g(X_T) + \frac{H(X_T)}{X_T} \cdot \Pi\right)\right|^2 \leq C \cdot \Delta$$

for some constant $C > 0$.

The main difficulty for $\gamma = 1/2$ is to control the inverse moments for the CIR process and its approximations. The same problem also appears when analyzing the approximation of the CIR process itself, see [3, 37]. For the bias of $\hat{P}^2$ numerical tests indicate a weak order of one. Our current research focuses on establishing this weak convergence rate and also on an improvement on the condition for $\gamma = 1/2$ in the above Proposition.

6.4 Multilevel Methods for Lévy-Driven SDEs

In this section we summarize our research on multilevel Monte Carlo methods for Lévy-driven stochastic differential equations. Let us first fix the notation. Let $X = (X_t)_{t\in[0,T]}$ be the solution of the stochastic differential equation

$$X_t = x_0 + \int_0^t a(X_{s-})\,\mathrm{d}Y_s, \tag{6.10}$$

where a is a Lipschitz continuous coefficient and $Y = (Y_t)_{t\in[0,T]}$ is a square integrable Lévy process. For ease of notation we will assume that the processes X and Y are one-dimensional, although this is not mandatory for most results. In particular, one could add a Lipschitz drift term and get analogous results.

The aim of this branch of the project was to design and analyse new numerical multilevel methods for the computation of expectations $S(f) = \mathbf{E}[f(X)]$, in the path-dependent and marginal setting.

The distribution of the L^2-Lévy process Y is characterized by the Lévy triplet (b, σ^2, ν) constituted by

- The drift $b \in \mathbb{R}$,
- The diffusion coefficient $\sigma^2 \in [0, \infty)$ and
- The Lévy measure ν, a measure on $\mathbb{R}\backslash\{0\}$ with $\int x^2\,\nu(dx) < \infty$

via its characteristic exponent $\mathbf{E}[e^{izY_t}] = \exp\{t\psi(z)\}$ $(z \in \mathbb{R})$, where

$$\psi(z) := ib - \frac{1}{2}\sigma^2 z + \int (e^{izx} - 1 - izx)\,\nu(dx).$$

The Lévy process is naturally connected to a simple Poisson point process Π on the Borel sets of $(0, \infty) \times (\mathbb{R}\backslash\{0\})$ given by

$$\Pi = \sum_{s>0:\Delta Y_s \neq 0} \delta_{(s,\Delta Y_s)},$$

where $\Delta Y_s = Y_s - Y_{s-}$ denotes the displacement of Y in s. It has intensity $\ell|_{(0,\infty)} \otimes \nu$, where $\ell|_{(0,\infty)}$ denotes Lebesgue measure on $(0,\infty)$, and we will also consider its compensated version, the random signed measure $\bar{\Pi}$ on $(0,\infty) \times (\mathbb{R}\backslash\{0\})$ given by

$$\bar{\Pi} = \Pi - \ell|_{(0,\infty)} \otimes \nu.$$

Now we can represent (Y_t) as the limit

$$Y_t = bt + \sigma W_t + \lim_{\delta \downarrow 0} \int_{(0,t]\times B(0,\delta)^c} x \, \mathrm{d}\bar{\Pi}(s,x), \tag{6.11}$$

where (W_t) denotes an independent Brownian motion and the limit is to be understood locally uniformly in L^2. For further details concerning Lévy processes, we refer the reader to [6, 11] and [42]. For ease of notation, we write

$$\int_{(0,t]\times A} x \, \mathrm{d}\bar{\Pi}(s,x) = \lim_{\delta \downarrow 0} \int_{(0,t]\times (A \cap B(0,\delta)^c)} x \, \mathrm{d}\bar{\Pi}(s,x),$$

for a Borel set A of $\mathbb{R}\backslash\{0\}$ and note that the limit always exists locally uniformly in L^2.

6.4.1 Multilevel Monte Carlo with an Jump-Adapted Euler Scheme

Let us introduce the numerical scheme explicitly. We use approximate solutions that are parametrised by three positive parameters

- $h \in (0,\infty)$, the threshold for the size of the jumps being considered large or small,
- $\varepsilon \in (0,\infty)$ with $\frac{T}{\varepsilon} \in \mathbb{N}$, the fixed update intervals for the Brownian motion, drift and large jumps,
- $\varepsilon' \in \varepsilon\mathbb{N}$, the update intervals for the contribution of the small jumps.

Let us first explain how the Lévy process is simulated. We will represent the Lévy process Y as the sum of four independent processes. We denote by N^h and M^h the process constituted by the compensated small jumps, resp. large jumps, that is

$$N_t^h = \int_{(0,t]\times B(0,h)^c} x \, \mathrm{d}\bar{\Pi}(s,x), \quad M_t^h = \int_{(0,t]\times (B(0,h)\backslash\{0\})} x \, \mathrm{d}\bar{\Pi}(s,x), \quad \text{for } t \in [0,T],$$

and note that $Y_t = bt + \sigma W_t + N_t^h + M_t^h$, see (6.11). In general simulation of increments of (M_t^h) is not straight-forward. With precomputation it is possible to design fast simulation algorithms on intervals of fixed length provided that the

dimension of the Lévy process is small. Conversely, we would like to consider large jumps immediately when they occur. We work with two sets of update times, namely

$$\mathbf{I} = (\varepsilon \mathbb{N}_0 \cap [0,T]) \cup \{t \in (0,T] : \Delta N^h_t \neq 0\} = \{T_0, T_1, \dots\},$$

where $T_0, T_1, \dots$ is an enumeration of the times in $\mathbf{I}$ in increasing order. The contribution of the small jumps M^h is only taken into account at the deterministic times $\mathbf{J} = \varepsilon' \mathbb{N}_0 \cap [0,T]$.

In practise, one simulation of an approximate solution is based on the simulation of W and N^h on the random set of times $\mathbf{I}$ and a simulation of M^h on the (deterministic) time grid $\mathbf{J}$. Simulation of N^h is achieved via a simulation of the point process $\Pi|_{(0,T]\times B(0,h)^c}$ and the simulation of the Brownian motion is straightforward since it is independent of N^h. To simulate the increments of M^h on $\mathbf{J}$, we invert the characteristic function with the Hilbert transform method in a precomputation and record the cumulative distribution function on a large and fine enough grid, see [12].

We are now in the position to introduce the approximate solution. We let $\hat{X}_0 = x_0$ and for $n = 1, 2, \dots$

$$\begin{aligned}\hat{X}_{T_n} = \hat{X}_{T_{n-1}} + a(\hat{X}_{T_{n-1}})\,(b(T_n - T_{n-1}) + \sigma(W_{T_n} - W_{T_{n-1}}) + N^h_{T_n} - N^h_{T_{n-1}}) \\ + \mathbb{1}_{\mathbf{J}}(T_n)\, a(\hat{X}_{T_n - \varepsilon'})\,(M^h_{T_n} - M^h_{T_n - \varepsilon'}).\end{aligned}$$

There are two canonical ways to extend the approximate solution to a process on $[0,T]$: the *piecewise constant approximation* extends the process such that it is constant on each interval $[T_{n-1}, T_n)$ and the *continuous approximation*, which uses

$$\hat{X}_t = \hat{X}_{T_{n-1}} + a(\hat{X}_{T_{n-1}})\,(b(t - T_{n-1}) + \sigma(W_t - W_{T_{n-1}}) + N^h_t - N^h_{T_{n-1}})$$

for $t \in [T_{n-1}, T_n)$.

We work with multilevel Monte Carlo schemes as introduced by Giles [22] and Heinrich [24]. Instead of working with one approximate solution that is parametrised by $(h, \varepsilon, \varepsilon')$ we now work with a hierarchical family of approximate solutions parametrised by $(h_k, \varepsilon_k, \varepsilon'_k : k \in \mathbb{N})$ with all individual sequences decreasing to zero as $k \to \infty$. For each $k \in \mathbb{N}$, we denote by $\hat{X}^k = (\hat{X}^k_t)_{t\in[0,T]}$ the piecewise constant approximation for the triple $(h_k, \varepsilon_k, \varepsilon'_k)$. Formally, a multilevel Monte Carlo scheme $\hat{S}$ is an element of the $\mathbb{N}$-valued vectors $(n_1, \dots, n_L)$ of arbitrary finite length: for a measurable function $f : D[0,T] \to \mathbb{R}$ we approximate $S(f)$ by

$$\mathbf{E}[f(\hat{X}^1)] + \mathbf{E}[f(\hat{X}^2) - f(\hat{X}^1)] + \dots + \mathbf{E}[f(\hat{X}^L) - f(\hat{X}^{L-1})]$$

and denote by $\hat{S}(f)$ the random output that is obtained when estimating the individual expectations $\mathbf{E}[f(\hat{X}^1)]$, $\mathbf{E}[f(\hat{X}^2) - f(\hat{X}^1)]$, $\dots$, $\mathbf{E}[f(\hat{X}^L) - f(\hat{X}^{L-1})]$ by classical Monte-Carlo with $n_1, \dots, n_L$ iterations and summing up the individual

estimates. A natural notion of the cost of such an algorithm is the expected number of Euler steps used in the simulation, that is

$$\mathrm{cost}(\hat{S}) = n_1 \mathbf{E}[\#\mathbf{I}_1] + \sum_{k=2}^{L} n_k \, \mathbf{E}[\#\mathbf{I}_{k-1} + \#\mathbf{I}_k],$$

where $\mathbf{I}_k$ is as the $\mathbf{I}$ above for the k-th approximate solution.

6.4.2 Error Estimates in the Path-Dependent Case

In this section we consider the path-dependent setting. Let $D[0,T]$ denote the space of real-valued càdlàg functions on $[0,T]$ endowed with *supremum norm*. The quadrature problem for Lipschitz continuous functionals $f : D[0,T] \to \mathbb{R}$ was already treated in the first period, see [15] and [16]. To explain the findings we make use of the Blumenthal-Getoor index which is given by

$$\beta := \inf\Big\{ p \in [0,2] : \int_{0<|x|<1} |x|^p \, \mathrm{d}\nu(x) < \infty \Big\} \in [0,2].$$

The bottleneck in the simulation of Lévy-driven SDEs is the simulation of increments of the process $(\int_{(0,t]\times(B(0,h)\setminus\{0\})} x \, \mathrm{d}\bar{N}(s,x))_{t\in[0,T]}$. In high dimensions it is typically not feasible to do direct simulation. Two canonical approaches, typically referred to as shot noise approximations, are

1. To simulate jumps larger than a threshold and disregard small jumps or
2. To simulate jumps larger than a threshold and add an appropriate Brownian motion for the non-simulated small jumps, see [7].

In the case where the Blumenthal-Getoor index is smaller or equal to one, the first approach leads to multilevel Monte Carlo schemes whose root mean squared error decays of order $N^{-1/2+o(1)}$ when the runtime N of the algorithm tends to infinity. However, if the Blumenthal-Getoor index is larger than one efficiency is drastically fading, see [16] . A remedy to get better rates for $\beta \in [1,2]$ is approach two. Here significantly better rates can be proved for large β. However the order is still decreasing with increasing β and clearly lower than the order $1/2$ obtained for small Blumenthal-Getoor indices, see [15]. The marginal setting was considered in [32] and similar effects where observed.

For the multilevel scheme introduced in Sect. 6.4.1 the Blumenthal-Getoor index has no severe impact on the order of convergence. One often obtains multilevel Monte Carlo schemes with errors of order $N^{-\frac{1}{2}+o(1)}$, when the runtime N of the algorithm tends to infinity:

Theorem 6.6 ([18]). *Let Y be a square integrable Lévy process with Blumenthal-Getoor index smaller than two and $a : \mathbb{R} \to \mathbb{R}$ be a Lipschitz*

continuous coefficient. For a measurable and Lipschitz continuous (w.r.t. supremum norm) function $f : D[0,T] \to \mathbb{R}$*, there exist multilevel Monte Carlo algorithms* $(\hat{S}_N : N \in \mathbb{N})$ *with* $\mathrm{cost}(\hat{S}_N) \leq N$ *such that the following property holds for a positive constant* κ*:*

1. *If the Lévy process* Y *has no Gaussian component, then one has*

$$\mathbf{E}\left[\left|\hat{S}_N(f) - S(f)\right|^2\right]^{1/2} \leq \kappa\, N^{-\frac{1}{2}}, \text{ for large } N.$$

2. *If the Lévy process* Y *has a Gaussian component, then one has*

$$\mathbf{E}\left[\left|\hat{S}_N(f) - S(f)\right|^2\right]^{1/2} \leq \kappa\, N^{-\frac{1}{2}} (\log N)^{\frac{3}{2}}, \text{ for large } N.$$

Remark 6.6. The algorithms $(\hat{S}_N : N \in \mathbb{N})$ and the constant κ can be uniformly chosen for all f with Lipschitz-seminorm less than a constant. For the particular choice of the parameters we refer the reader to [18].

6.4.3 Stable Convergence and Central Limit Theorems

Central limit theorems for classical Monte Carlo were derived by Duffie and Glynn in [20] for various schemes for classical and Lévy-driven SDEs. These results are typically the consequence of stable convergence of the error process between approximate and genuine solution. Unfortunately the stable convergence results cannot be immediately used in the context of multilevel Monte Carlo and we mention Ben Alaya and Kebaier [10] for a central limit theorem for multilevel Monte Carlo for classical diffusions.

Motivated by Theorem 6.6 we analyse the jump adapted Euler approximations introduced above which are typically not covered by standard theory even for classical Monte Carlo. We will proceed as follows. First we give the necessary assumptions and introduce the main theorem on stable convergence of the involved error processes. In a second step, we deduce the central limit theorem for the multilevel scheme.

Stable convergence, first introduced by Rényi [41] and then studied by Aldous and Eagleson [2], is a property of sequences of random variables. To be more precise, given a σ-field $\mathscr{F}$ and a sequence $(Z^k)_{k\in\mathbb{N}}$ of $\mathscr{F}$-measurable random variables taking values in a Polish space E, we say (Z^k) converges stably to Z or briefly $Z^k \overset{\text{stable}}{\Longrightarrow} Z$, if for every $A \in \mathscr{F}$ and continuous and bounded $f : E \to \mathbb{R}$

$$\lim_{k\to\infty} \mathbf{E}[\mathbb{1}_A f(Z^k)] = \mathbf{E}[\mathbb{1}_A f(Z)].$$

We refer the reader to [25] and [26] for basic results concerning stable convergence.

Assumptions. *We assume that* $\sigma \neq 0$ *and suppose that the Blumenthal-Getoor index is strictly smaller than two. The parameters* $(h_k, \varepsilon_k, \varepsilon'_k : k \in \mathbb{N})$ *are componentwise monotonically decreasing sequences satisfying*

- $\varepsilon_k = M^{-k}T$, *where* $M \in \{2, 3, \dots\}$ *is fixed,*
- $\nu(B(0,h_k)^c)\,\varepsilon_k \to \theta$ *as* $k \to \infty$, *where* $\theta \in [0,\infty)$, *and*
- $\varepsilon'_k \in \varepsilon_k \mathbb{N} \cap (0,T]$ *with* $\varepsilon'_k \int_{B(0,h_k)} x^2\, \nu(\mathrm{d}x) \log^2(1 + 1/\varepsilon'_k) = o(\varepsilon_k)$ *and* $h_k^2 \log^2(e + 1/\varepsilon'_k) = o(\varepsilon_k)$.

Such parameters exist, if the Blumenthal-Getoor index is strictly smaller than two.

To state the main theorem concerning stable convergence, we need some further notation. We let $\mathscr{F}$ *denote the* σ*-field generated by the Lévy process* Y. *Further, we denote by* $(B_t)_{t\in[0,T]}$ *an* $\mathscr{F}$*-independent standard Brownian motion and equip the points of the point process* Π *with independent marks by denoting for every point* $(s,x) \in \Pi$

- *By* χ_s, *a standard normal distributed variable,*
- *By* ζ_s, *an independent uniform random variable on* $[0,1]$, *and*
- *By* $\xi_{s,1}$ *and* $\xi_{s,2}$ *independent* $Exp(\theta)$ *and* $Exp((M-1)\theta)$*-distributed random variables respectively.*

Theorem 6.7 ([17]). *Assume that* $a : \mathbb{R} \to \mathbb{R}$ *is continuously differentiable with uniformly bounded derivative* a'. *We denote by* $(\hat{X}^k : k \in \mathbb{N})$ *the* continuous *approximate solutions, as introduced in Sect. 6.4.1. Under the assumptions from above, one has*

$$\varepsilon_k^{-\frac{1}{2}}(\hat{X}^{k+1} - \hat{X}^k) \overset{\text{stable}}{\Longrightarrow} U, \quad \text{in the Skorokhod topology},$$

where $U = (U_t)_{t\in[0,T]}$ *is the solution of*

$$\begin{aligned} U_t = {} & \int_0^t a'(X_{s-})\, U_{s-}\, \mathrm{d}Y_s + \sigma^2 \Upsilon(\theta) \int_0^t (aa')(X_{s-})\, \mathrm{d}B_s \\ & + \sigma \sum_{s\in(0,t]:\Delta Y_s \neq 0} \sqrt{\varphi_s}\, \chi_s\, (aa')(X_{s-})\, \Delta Y_s, \end{aligned} \tag{6.12}$$

where $\Upsilon^2 = \left[\frac{1}{\theta} - (1-e^{-\theta})\frac{1}{\theta^2}\right](1-\frac{1}{M})$, *if* $\theta > 0$, $\Upsilon^2 = \frac{1}{2}(1-\frac{1}{M})$, *if* $\theta = 0$, *and*

$$\varphi_s = \sum_{1\le m\le M} \mathbb{1}_{\{\frac{m-1}{M} \le \zeta_s < \frac{m}{M}\}} \Big[\min(\xi_{s,1}, \zeta_s) - \min\big(\xi_{s,1}, \xi_{s,2}, \zeta_s - \frac{m-1}{M}\big)\Big].$$

Here the infinite sum in (6.12) has to be understood as an appropriate martingale limit.

Theorem 6.8 ([17]). *Let* $f : \mathbb{R} \to \mathbb{R}$ *be Lipschitz continuous and differentiable in* $\mathbf{P}_{X_T} = \mathbf{P} \circ X_T^{-1}$*-almost every point. Assume that* $\alpha \ge \frac{1}{2}$ *is such that the limit*

$$\kappa := \lim_{n\to\infty} \varepsilon_n^{-\alpha} \mathbf{E}\Big[f(\hat{X}_T^n) - f(X_T) \Big]$$

exists. For $N \geq 1$ *we denote by* $\hat{S}_N$ *the multilevel scheme with parameters*

$$(n_1(N), \ldots, n_{L(N)}(N))$$

with

1. $L(N) = \left\lceil \frac{\log N}{2\alpha \log M} \right\rceil$ *and*
2. $n_k(N) = \lceil \varepsilon_{k-1}\, N\, L(N) \rceil$, *for* $k = 1, 2, \ldots, L(N)$.

Then $\mathrm{cost}(\hat{S}_N) = \mathcal{O}(N(\log N)^2)$ *and one has*

$$\sqrt{N}\Big(\hat{S}_N(f) - S(f) \Big) \Longrightarrow \mathcal{N}(\kappa, \rho^2), \text{ as } N \to \infty,$$

where $\mathcal{N}(\kappa, \rho^2)$ *is the normal distribution with expectation* κ *and variance* $\rho^2 = \mathrm{Var}(f'(X_T)\, U_T)$.

References

1. Aït-Sahalia, Y., Kimmel, R.: Maximum likelihood estimation of stochastic volatility models. J. Financ. Econ. **83**, 413–452 (2007)
2. Aldous, D.J., Eagleson, G.K.: On mixing and stability of limit theorems. Ann. Probab. **6**(2), 325–331 (1978)
3. Alfonsi, A.: Strong order one convergence of a drift implicit Euler scheme: application to the CIR process. Stat. Probab. Lett. **83**(2), 602–607 (2013)
4. Altmayer, M., Neuenkirch, A.: Multilevel Monte Carlo quadrature of discontinuous payoffs in the generalized Heston model using Malliavin integration by parts. DFG SPP 1324 Preprint 144 (2013)
5. Andersen, L.B.G., Piterbarg, V.V.: Moment explosions in stochastic volatility models. Financ. Stoch. **11**, 29–50 (2006)
6. Applebaum, D.: Lévy processes and stochastic calculus. In: Cambridge Studies in Advanced Mathematics, vol. 116. Cambridge University Press, Cambridge/New York (2009)
7. Asmussen, S., Rosiński, J.: Approximations of small jumps of Lévy processes with a view towards simulation. J. Appl. Probab. **38**(2), 482–493 (2001)
8. Avikainen, R.: On irregular functionals of SDEs and the Euler scheme. Financ. Stoch. **13**, 381–401 (2009)
9. Bakhvalov, N.S.: On approximate computation of integrals. Vestnik MGU, Ser. Math. Mech. Astron. Phys. Chem. **4**, 3–18 (1959)
10. Ben Alaya, M., Kebaier, A.: Central limit theorem for the multilevel Monte Carlo Euler method to appear in Ann. Appl. Probab. (2013, Preprint)
11. Bertoin, J.: Lévy Processes. Cambridge University Press, Cambridge (1996)
12. Chen, Z.S., Feng, L.M., Lin, X.: Simulating Lévy processes from their characteristic functions and financial applications. ACM Trans. Model. Comput. Simul. **22**(3), 14 (2012)
13. Davis, P.: A construction of nonnegative approximate quadratures. Math. Comp. **21**, 578–582 (1967)

14. Deck, T., Kruse, S.: Parabolic differential equations with unbounded coefficents – a generalization of the parametrix method. Acta Appl. Math. **74**, 71–91 (2002)
15. Dereich, S.: Multilevel Monte Carlo algorithms for Lévy-driven SDEs with Gaussian correction. Ann. Appl. Probab. **21**(1), 283–311 (2011)
16. Dereich, S., Heidenreich, F.: A multilevel Monte Carlo algorithm for Lévy-driven stochastic differential equations. Stoch. Process. Appl. **121**(7), 1565–1587 (2011)
17. Dereich, S., Li, S.: Multilevel Monte Carlo for Lévy-driven SDEs: central limit theorems for adaptive Euler schemes, DFG SPP 1324 Preprint 161 (2014)
18. Dereich, S., Li, S.: Multilevel Monte Carlo for Lévy-driven SDEs with direct simulation of increments (2014, work in progress)
19. Dimitroff, G., Lorenz, S., Szimayer, A.: A parsimonious multi-asset Heston model: Calibration and derivative pricing. Int. J. Theor. Appl. Financ. **14**(08), 1299–1333 (2011)
20. Duffie, D., Glynn, P.: Efficient Monte Carlo simulation of security prices. Ann. Appl. Probab. **5**(4), 897–905 (1995)
21. Giles, M., Nagapetyan, T., Ritter, K.: Multi-level Monte Carlo approximation of distribution functions and densities. DFG SPP 1324 Preprint 157 (2014)
22. Giles, M.B.: Multilevel Monte Carlo path simulation. Oper. Res **56**(3), 607–617 (2008)
23. Giles, M.B., Higham, D.J., Mao, X.: Analysing multilevel Monte Carlo for options with non-globally Lipschitz payoff. Financ. Stoch. **13**(1), 403–413 (2009)
24. Heinrich, S.: Multilevel Monte Carlo methods. In: Large-Scale Scientific Computing. Lecture Notes in Computer Science, vol. 2179, pp. 58–67. Springer, Berlin/Heidelberg (2001)
25. Jacod, J.: On continuous conditional Gaussian martingale and stable convergence in law. In: Séminaire de Probabilités, XXXI, Lecture Notes in Mathematics, vol. 1655, pp. 232–246. Springer, Berlin/Heidelberg (1997)
26. Jacod, J., Shiryaev, A.N.: Limit Theorems for Stochastic Processes. Springer, Berlin/New York (1987)
27. Kloeden, P., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin (1999)
28. Kusuoka, S.: Approximation of expectation of diffusion processes and mathematical finance. Adv. Stud. Pure Math. **31**, 147–165 (2001)
29. Kusuoka, S.: Approximation of expectation of diffusion processes based on Lie algebra and Malliavin calculus. In: Advances in Mathematical Economics, vol. 6, pp. 69–83. Springer, Tokyo (2004)
30. Litterer, C., Lyons, T.: High order recombination and an application to cubature on Wiener space. Ann. Appl. Prob. **22**, 1301–1327 (2012)
31. Lyons, T., Victoir, N.: Cubature on Wiener space. Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci. **460**, 169–198 (2004)
32. Marxen, H.: The multilevel monte carlo method used on a Lévy driven SDE. Mt. Carlo Methods Appl. **16**(2), 167–190 (2010)
33. Müller-Gronbach, T., Ritter, K.: A local refinement strategy for constructive quantization of scalar SDEs. Found. Comput. Math. **13**, 1005–1033 (2013)
34. Müller-Gronbach, T., Ritter, K., Yaroslavtseva, L.: Derandomization of the Euler scheme for scalar stochastic differential equations. J. Complex. **28**(2), 139–153 (2012)
35. Müller-Gronbach, T., Ritter, K., Yaroslavtseva, L.: On the complexity of computing quadrature formulas for marginal distributions of SDEs. J. Complex. (2014, to appear)
36. Müller-Gronbach, T., Yaroslavtseva, L.: Deterministic quadrature formulas for SDEs based on simplified weak Itô-Taylor steps. DFG SPP 1324 Preprint 167 (2014)
37. Neuenkirch, A., Szpruch, L.: First order strong approximation of scalar SDEs defined in a domain. Numer. Math. (2014, to appear)
38. Novak, E.: Deterministic and Stochastic Error Bounds in Numerical Analysis. Lecture Notes in Mathematics, vol. 1349. Springer, Berlin (1988)
39. Nualart, D.: The Malliavin calculus and related topics. In: Probability and its Applications, 2nd edn. Springer, Berlin (2006)

40. Petras, K., Ritter, K.: On the complexity of parabolic initial-value problems with variable drift. J. Complex. **22**(1), 118–145 (2006)
41. Rényi, A.: On stable sequences of events. Sankhyā Ser. A **25**, 293–302 (1963)
42. Sato, K.: Lévy Processes and Infinitely Divisible Distributions. Cambridge Studies in Advanced Mathematics, vol. 68. Cambridge University Press, Cambridge (1999)
43. Talay, D.: Efficient numerical schemes for the approximation of expectations of functionals of the solution of a SDE and applications. In: Filtering and Control of Random Processes (Paris, 1983). Lecture Notes in Control and Information Science, vol. 61, pp. 294–313. Springer, Berlin (1984)

Chapter 7
Bayesian Inverse Problems and Kalman Filters

Oliver G. Ernst, Björn Sprungk, and Hans-Jörg Starkloff

Abstract We provide a brief introduction to Bayesian inverse problems and Bayesian estimators emphasizing their similarities and differences to the classical regularized least-squares approach to inverse problems. We then analyze Kalman filtering techniques for nonlinear systems, specifically the well-known Ensemble Kalman Filter (EnKF) and the recently proposed Polynomial Chaos Expansion Kalman Filter (PCE-KF), in this Bayesian framework and show how they relate to the solution of Bayesian inverse problems.

7.1 Introduction

In recent years the interest and research activity in uncertainty quantification (UQ) for complex systems modelled by partial differential equations (PDEs) has increased significantly. This is due both to growing available computing resources as well as new efficient numerical methods for high-dimensional problems, which together make the solution of UQ problems associated with PDEs feasible. The motivation driving UQ is the simple fact that, in practical applications, we usually do not know parameters, coefficients or even boundary conditions for the PDE model under consideration exactly. A typical example are material properties such as conductivity. At the same time, we may still have some knowledge about possible values for these uncertain input data, e.g., the hydraulic conductivity of layered clay may be between 10^{-6} and 10^{-4} cm/s. A careful simulation would take into account the uncertainty in the input data and quantify the resulting uncertainty in the output of the physical or PDE model. Although there are also other mathematical techniques for modeling uncertainty such as fuzzy set theory or interval arithmetic, we focus here on the probabilistic approach.

O.G. Ernst (✉) • B. Sprungk
Technical University of Chemnitz, Reichenhainer Str. 41, 09126 Chemnitz, Germany
e-mail: oliver.ernst@mathematik.tu-chemnitz.de; bjoern.sprungk@mathematik.tu-chemnitz.de

H.-J. Starkloff
University of Applied Sciences Zwickau, Postfach 201037, 08012 Zwickau, Germany
e-mail: hans.joerg.starkloff@fh-zwickau.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_7

Initially, the main interest has been in solving the *forward problem*, in which one is given the probability law of uncertain data $u \sim \mu$ with the goal of computing the corresponding law of a quantity of interest $\phi = F(u)$, where F represents the composition of solving a PDE and evaluating a functional of its solution. Current numerical methods for this task include, e.g., multilevel Monte Carlo, stochastic Galerkin and stochastic collocation methods, proper orthogonal decomposition, and Gaussian process emulators.

Within UQ, the more fundamental task is to develop a good probability law for the unknown quantity u reflecting our (possibly subjective) knowledge of u, since this determines the outcome. In general, transforming expert knowledge and physical reasoning into a probability distribution is a subtle and quite difficult task. Moreover, incorporating any available information about the unknown into the probability law is desirable, since this will, in general, reduce uncertainty and lead to improved models. For this reason the *inverse problem* has received increased attention in the UQ community.

Specifically, given noisy data $z = G(u)+\varepsilon$, the task is to either identify u or make inferences, i.e., refine an initial model of u. Here we want to distinguish between identification, i.e., determining a value u which best explains the data, and inference, i.e., updating our understanding or belief about u based on the new information z.

The latter is more interesting for UQ purposes, since adjusting prior probability models of the unknown according to indirect data yields an improved uncertainty model for u, whereas identification would merely provide a certain best estimate with no indication of how well this estimate is determined.

In the probabilistic setting, merging new information with a given prior model (i.e., a prior random variable or probability measure), is performed by *conditioning* this model on the available information, resulting in a conditional measure. The procedure of conditioning, and thus also the conditional measure or distribution, are rooted in Kolmogorov's fundamental concept of *conditional expectation*. In particular, *Bayes' rule* provides an analytic expression for the conditional measure in terms of the prior measure and provides the main tool in Bayesian inference as well as Bayesian inverse problems.

Since Bayesian inverse problems have gained much attention in the scientific computing community in the last few years, numerous algorithms and numerical methods have been proposed for their solution. We provide a short overview of existing methods and focus on the *Kalman Filter* and two of its variants, namely the *Ensemble Kalman Filter* [15] and the *Polynomial Chaos Expansion Kalman Filter* [35], which have been recently proposed for UQ in association with inverse problems. In particular, we investigate what these Kalman filtering methods actually compute and how they relate to Bayesian inverse problems and Bayes estimators. Thus, our main purpose is to clarify which quantities Kalman filters can and cannot approximate.

The remainder of this paper is organized as follows: Sect. 7.2 briefly recalls the deterministic and Bayesian approaches to inverse problems and provides a short overview of computational methods. In Sect. 7.3 we consider Kalman filtering methods and analyze these in the light of Bayes estimators. In particular, we show

that these filtering methods approximate a random variable which is, in general, not distributed according to the desired posterior measure. Moreover, we illustrate the performance of Kalman filters and the difference between their output and the solution of the Bayesian inverse problem for a simple 1D boundary value problem in Sect. 7.4. A summary and conclusions are given in Sect. 7.5.

7.2 Bayesian Approach to Inverse Problems

In this section we introduce the setting and notation for the inverse problem and recall the basic concepts of the classical regularized least-squares and the Bayesian approaches.

Throughout the article, $|\cdot|$ shall denote the Euclidean norm on $\mathbb{R}^k$, $\|\cdot\|$ the norm on a general separable Banach space $(\mathscr{X}, \|\cdot\|)$, $\mathscr{X}^*$ the topological dual of $\mathscr{X}$ and $\mathscr{Y}$ a second separable Banach space.

We consider the abstract inverse problem of identifying an unknown $u \in \mathscr{X}$ given finite-dimensional but noisy observations $z \in \mathbb{R}^k$ according to the model

$$z = G(u) + \varepsilon \tag{7.1}$$

containing an observation operator $G : \mathscr{X} \to \mathbb{R}^k$ and measurement noise $\varepsilon \in \mathbb{R}^k$.

Example 7.1 (Elliptic PDE). Consider the problem of determining the logarithm $\kappa \in C(D)$ of the conductivity $\exp(\kappa)$ of an incompletely known porous medium occupying a bounded domain $D \subset \mathbb{R}^d$ given observations of the pressure head p at several locations in the domain of a fluid in steady flow through the medium. The relation between κ and p can be modelled by, e.g.,

$$-\nabla \cdot (\mathrm{e}^{\kappa} \nabla p) = f \text{ on } D, \qquad p|_{\partial D} = 0. \tag{7.2}$$

Here the unknown is $u = \kappa$ and the observation operator G is the mapping $\kappa \mapsto (p(x_1), \ldots, p(x_k))$ for given measurement locations $x_i \in D$, $i = 1, \ldots, k$.

Example 7.2 (Discrete dynamics). Consider a discrete-time dynamical system $\{y_n\}_{n \in \mathbb{N}_0}$ with state evolution equation

$$y_{n+1} = h_n(y_n), \qquad y_0 = x \in \mathbb{R}^N,$$

where $h_n : \mathbb{R}^N \to \mathbb{R}^N$ governs the (deterministic) dynamics driving the system at step n. Suppose we observe J noisy states

$$z_{n_j} = y_{n_j} + \varepsilon_j, \qquad j = 1, \ldots, J, \quad 0 < n_1 < \cdots < n_J,$$

and wish to infer from these the unknown initial state $u = x$. Setting $G_j = h_0 \circ \cdots \circ h_{n_j-1}$ and $G := (G_1, \ldots, G_J)$, we arrive at a problem of the form (7.1). By extending the unknown u to the vector $(y_0, y_{n_1}, \ldots, y_{n_J})$ one arrives at the problem of inferring the J states at $n_1, \ldots, n_J$.

Remark 7.1. Identification problems for dynamical systems with sequentially arriving data call for special, efficient sequential methods for solving (7.1). These are methods for computing the solution for $z = (z_{n_1}, \ldots, z_{n_J})^\top$ based only on the solution for $(z_{n_1}, \ldots, z_{n_{J-1}})^\top$ and the current observation z_{n_J}. To limit the scope of this paper, we omit considerations of sequentiality in this work.

7.2.1 Deterministic Identification for Inverse Problems

Solving (7.1) by determining $u = G^{-1}(z)$ is usually not an option since $\varepsilon \neq 0$ generally results in $z \notin G(\mathscr{X})$. Moreover, the more general least-squares formulation $u = \operatorname{argmin}_{v\in\mathscr{X}} |z - G(v)|^2$ is typically ill-posed, as u may depend discontinuously on z and is often heavily underdetermined. Making (7.1) mathematically tractable is usually achieved by some form of *regularization*, which, generally speaking, involves the incorporation of additional *prior* information on u and ε. A comprehensive introduction to the regularized least-squares approach to inverse problems is given in [10]. We briefly summarize this approach for nonlinear G here.

The conceptual starting point for the deterministic approach is the noise-free model $z^\dagger = G(u)$, i.e., $z = z^\dagger + \varepsilon$. Since we want to identify the element $u \in \mathscr{X}$ which led to the observations z, it is reasonable to assume that the "true", unpolluted data $z^\dagger$ lies in the range of G. Thus we assume the existence of $u^\dagger \in \mathscr{X}$ such that $G(u^\dagger) = z^\dagger$. This is sometimes called the *attainability assumption* [11]. Next, we introduce a penalty or regularizing functional $\mathrm{R} : \mathscr{X} \to [0, \infty]$ and define an R-minimizing solution to $z^\dagger = G(u)$ to be any element $u^* \in \mathscr{X}$ which satisfies

$$\mathrm{R}(u^*) = \min\left\{\mathrm{R}(u) : u \in \mathscr{X},\ G(u) = z^\dagger\right\}. \tag{7.3}$$

Note that u^* need not be unique. Furthermore, the choice of R is significant and reflects prior assumptions about u. Often R is taken to be convex. A common choice for R is, e.g., $\mathrm{R}(u) = \|u - u^{\mathrm{ref}}\|^2$, where $u^{\mathrm{ref}} \in \mathscr{X}$ is a given reference state known to lie in the vicinity of the solution. For a broader discussion of different penalty functionals we refer to [36].

However, since only polluted data $z = z^\dagger + \varepsilon$ is available, we can only ask for an approximation of u^* which should improve with diminishing noise ε. This approximation is the regularized solution $\hat{u}_\alpha$ given by

$$\hat{u}_\alpha = \operatorname*{argmin}_{u\in\mathscr{X}} |z - G(u)|^2 + \alpha \mathrm{R}(u), \tag{7.4}$$

where $\alpha \in [0, \infty)$ serves as a regularization parameter to be chosen wisely. If further smoothness assumptions on u^* and G are satisfied and if α is chosen as a suitable function $\alpha = \alpha(\delta)$ of the noise level $|\varepsilon| \leq \delta$, then convergence rate bounds such as

$$\|\hat{u}_{\alpha(\delta)} - u^*\| = O(\sqrt{\delta}) \qquad \text{as } \delta \to 0$$

can be obtained [11]. These rates are typically based on explicit error estimates such as $\|\hat{u}_{\alpha(\delta)} - u^*\| \leq C(\alpha)\sqrt{\delta}$ for the above result. For further analysis of the smoothness requirements on u^* and related convergence rates see, e.g., [21] and, for appropriate choices $\alpha = \alpha(\delta)$, see, e.g., [1] and the references therein.

7.2.2 The Bayesian Inverse Problem

Recall that, in order to regularize the usually ill-posed least-squares formulation of the inverse problem (7.1), we incorporated additional prior information about the desired u into the (deterministic) identification problem by way of the regularization functional R. A further possibility for regularization is to restrict u to a subset or subspace $\tilde{\mathscr{X}} \subset \mathscr{X}$, e.g., by using a stronger norm of $u - u^{\text{ref}}$ as the regularization functional. Speaking very broadly, the Bayesian approach stems from yet another way of modelling prior information on u and adding it to the inverse problem. In this case we express our prior belief about u through a probability distribution μ_0 on the Banach space $\mathscr{X}$, by which a quantitative preference of some solutions u over others may be given by assigning higher and lower probabilities. However, the goal in the Bayesian approach is not the identification of a particular $u \in \mathscr{X}$, but rather inference on u, i.e., we would like to *learn from the data* in a statistical or probabilistic fashion by *adjusting our prior belief* μ_0 about u in accordance with the newly available data z. The task of identification may also be achieved within the Bayesian framework through *Bayes estimates* and *Bayes estimators*, which are discussed in Sect. 7.2.3.

The Bayesian approach to the inverse problem (7.1) thus differs conceptually from the regularized least-squares approach as summarized above in that its objective is inference rather than identification. As stated in [24], the Bayesian approach[1] is based on the following four principles:

1. All quantities occurring in (7.1) are modelled as random variables.
2. The randomness describes our degree of information concerning their realizations.
3. This degree of information concerning these values is encoded in probability distributions.
4. The solution of the inverse problem is the posterior probability distribution.

[1] This is referred to in [24] as the *statistical inversion approach*.

In the Bayesian setting we therefore replace our model (7.1) in the following with

$$Z = G(U) + \varepsilon, \tag{7.5}$$

where ε and hence Z are both random variables on $\mathbb{R}^k$ while U is a random variable on $\mathscr{X}$ whose posterior probability distribution given the available observations $Z = z$ is to be determined. Before giving a precise definition of the posterior distribution we require some basic concepts from probability theory.

7.2.2.1 Probability Measures and Random Variables

Let $(\Omega, \mathscr{F}, \mathbb{P})$ denote a probability space. We denote by $\mathscr{B}(\mathscr{X})$ the Borel σ-algebra of $\mathscr{X}$ generated by the open sets in $\mathscr{X}$ w.r.t. $\|\cdot\|$. A measurable mapping $X : (\Omega, \mathscr{F}) \to (\mathscr{X}, \mathscr{B}(\mathscr{X}))$ is called a random variable (RV) and the measure $\mathbb{P}_X := \mathbb{P} \circ X^{-1}$, i.e., $\mathbb{P}_X(A) = \mathbb{P}(X^{-1}(A))$ for all $A \in \mathscr{B}(\mathscr{X})$, defines the distribution of X as the push-forward measure of $\mathbb{P}$ under X. Conversely, given a probability measure μ on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$, then $X \sim \mu$ means $\mathbb{P}_X = \mu$. By $\sigma(X) \subset \mathscr{F}$ we denote the σ-algebra generated by X, i.e., $\sigma(X) = \{X^{-1}(A) : A \in \mathscr{B}(\mathscr{X})\}$.

The Bochner space of p-integrable $\mathscr{X}$-valued RVs, i.e., the space of RVs $X : \Omega \to \mathscr{X}$ such that $\int_\Omega \|X(\omega)\|^p \, \mathbb{P}(\mathrm{d}\omega) < \infty$, is denoted by $L^p(\Omega, \mathscr{F}, \mathbb{P}; \mathscr{X})$ or simply $L^p(\mathscr{X})$ when the context is clear.

An element $m \in \mathscr{X}$ is called the *mean* of a RV X if for any $f \in \mathscr{X}^*$ there holds $f(m) = \mathbb{E}[f(X)]$. Here and in the following $\mathbb{E}$ denotes the expectation operator w.r.t. $\mathbb{P}$. If $X \in L^1(\Omega, \mathscr{F}, \mathbb{P}; \mathscr{X})$ then its mean is given by $m = \mathbb{E}[X] = \int_\Omega X(\omega) \, \mathbb{P}(\mathrm{d}\omega)$. An operator $C : \mathscr{Y}^* \to \mathscr{X}$ is called the *covariance* of two RVs $X : \Omega \to \mathscr{X}$ and $Y : \Omega \to \mathscr{Y}$ if it satisfies $f(Cg) = \mathbb{E}\big[f(X - \mathbb{E}[X])\, g(Y - \mathbb{E}[Y])\big]$ for all $f \in \mathscr{X}^*$ and $g \in \mathscr{Y}^*$. We denote the covariance of X and Y by $\mathrm{Cov}(X, Y)$ and, if $X = Y$, simply by $\mathrm{Cov}(X)$.

Besides normed vector spaces of RVs we will also work with metric spaces of probability measures. One notion of distance between measures is the *Hellinger metric* d_H: given two probability measures μ_1 and μ_2 on the Banach space $\mathscr{X}$, it is defined as

$$d_H(\mu_1, \mu_2) := \left[\int_{\mathscr{X}} \left(\sqrt{\frac{\mathrm{d}\mu_1}{\mathrm{d}\nu}(u)} - \sqrt{\frac{\mathrm{d}\mu_2}{\mathrm{d}\nu}(u)} \right)^2 \nu(\mathrm{d}u) \right]^{1/2},$$

where ν is a dominating measure of μ_1 and μ_2, e.g., $\nu = (\mu_1 + \mu_2)/2$. Note that the definition of the Hellinger metric is independent of the dominating measure. For relations of the Hellinger metric to other probability metrics such as total variation distance or the Wasserstein metric, we refer to [18].

In the following, we will use upper case Latin letters such as X, Y, Z, U to denote RVs on Banach spaces and lower case Latin letters like x, y, z, u for elements in these Banach spaces or realizations of the associated RVs, respectively. Greek letters such as ε will be used to denote RVs on $\mathbb{R}^k$ as well as their realizations.

7.2.2.2 Conditioning

Bayesian inference consists in updating the probability distribution encoding our prior knowledge on the unknown U to a new probability distribution reflecting a gain in knowledge due to new observations. There are certain subleties associated with the probabilistic formulation of this transition from prior to posterior measure, and we take some care in this section to point these out.

The distribution of the RV U, characterized by the probabilities $\mathbb{P}(U \in B)$ for $B \in \mathscr{B}(\mathscr{X})$, quantifies in stochastic terms our knowledge about the uncertainty associated with U. When new information becomes available, such as knowing that the event $Z = z$ has occurred, this is reflected in our quantitative description as the "conditional distribution of U given $\{Z = z\}$", denoted $\mathbb{P}(U \in B|Z = z)$. Unfortunately, $\mathbb{P}(U \in B|Z = z)$ cannot be defined in an elementary fashion when $\mathbb{P}(Z = z) = 0$, in which case the conditional distribution is defined by an integral relation. The key concept here is that of *conditional expectation*.

Given RVs $X \in L^1(\Omega, \mathscr{F}, \mathbb{P}; \mathscr{X})$ and $Y : \Omega \to \mathscr{Y}$, we define the conditional expectation $\mathbb{E}[X|Y]$ of X given Y as any mapping $\mathbb{E}[X|Y] : \Omega \to \mathscr{X}$ with the following two properties:

1. $\mathbb{E}[X|Y]$ is $\sigma(Y)$-measurable.
2. For any $A \in \sigma(Y)$ there holds

$$\int_A \mathbb{E}[X|Y]\, \mathbb{P}(\mathrm{d}\omega) = \int_A X\, \mathbb{P}(\mathrm{d}\omega).$$

Note that, since it is defined by an integral relation, the RV $\mathbb{E}[X|Y]$ is determined only up to sets of $\mathbb{P}$-measure zero and is thus understood as an equivalence class of such mappings. By the Doob-Dynkin Lemma (cf. [25, Lemma 1.13]) there exists a measurable function $\phi : \mathscr{Y} \to \mathscr{X}$ such that $\mathbb{E}[X|Y] = \phi(Y)$ $\mathbb{P}$-almost surely. Again, we note that this does not determine a unique function ϕ but an equivalence class of measurable functions, where $\phi_1 \sim \phi_2$ iff $\mathbb{P}(Y \in \{y \in \mathscr{Y} : \phi_1(y) \neq \phi_2(y)\}) = 0$. For a specific realization y of Y (and a specific ϕ), we also denote the function value by

$$\mathbb{E}[X|Y = y] := \phi(y) \in \mathscr{X}.$$

Setting $X = \mathbf{1}_{\{U \in B\}}$, one can, for each fixed $B \in \mathscr{B}(\mathscr{X})$, define

$$\mathbb{E}[\mathbf{1}_{\{U \in B\}}|Z = z] =: \mathbb{P}(U \in B|Z = z) \tag{7.6}$$

as an equivalence class of measurable functions $\mathbb{R}^k \to [0, 1]$. One would like to view this, conversely, as a family of probability measures with the realization z as a parameter, giving the posterior distribution of U resulting from having made the observation $Z = z$. Unfortunately, this construction need not, in general, yield a

probability measure for each fixed value of z (cf. [33]). In case $\mathscr{X}$ is a separable Banach space, a function

$$Q : \mathscr{B}(\mathscr{X}) \times \mathbb{R}^k \to \mathbb{R}$$

can be shown to exist (cf. [33]) such that

(a) For each $z \in \mathbb{R}^k$, $Q(\cdot, z)$ is a probability measure on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$.
(b) For each $B \in \mathscr{B}(\mathscr{X})$ the function

$$\mathbb{R}^k \ni z \mapsto Q(B, z)$$

is a representative of the equivalence class (7.6), i.e., it is measurable and there holds

$$\mathbb{P}(U \in B, Z \in A) = \int_A Q(B, z)\, \mathbb{P}_Z(\mathrm{d}z) \qquad \forall A \in \mathscr{B}(\mathbb{R}^k).$$

Such a function Q, also denoted by $\mu_{U|Z}$, is called the *regular conditional distribution* of U given Z and is defined uniquely up to sets of z-values of $\mathbb{P}_Z$-measure zero. We have thus arrived at a consistent definition of the posterior probability $\mathbb{P}(U \in B | Z = z)$ as $\mu_{U|Z}(B, z)$.

It is helpful to maintain a clear distinction between *conditional* and *posterior* quantities: the former contain the – as yet unrealized – observation as a parameter, while in the latter the observation has been made. Specifically, $\mu_{U|Z}$ is the conditional measure of U conditioned on Z, whereas $\mu_{U|Z}(\cdot, z)$ denotes the posterior measure of U for the observation $Z = z$.

7.2.2.3 Bayes' Rule and the Posterior Measure

We make the following assumptions for the model (7.5).

Assumption 7.1. *1. $U \sim \mu_0$, $\varepsilon \sim \mu_\varepsilon$ and $(U, \varepsilon) \sim \mu_0 \otimes \mu_\varepsilon$, i.e., U and ε are independent.*
2. $\mu_\varepsilon = \rho(\varepsilon)\, \mathrm{d}\varepsilon$ where $\rho(\varepsilon) = C \mathrm{e}^{-\ell(\varepsilon)}$ with $C > 0$ and $\ell : \mathbb{R}^k \to \mathbb{R}_0^+$ measurable and nonnegative. Here $\mathrm{d}\varepsilon$ denotes Lebesgue measure on $\mathbb{R}^k$.
3. $G : \mathscr{X} \to \mathbb{R}^k$ is continuous.

Throughout we assume $\mu_0(\mathscr{X}) = 1$ and $\mu_\varepsilon(\mathbb{R}^k) = 1$. By Assumption 7.1, the distribution μ_Z of Z in (7.5) is determined as $\mu_Z = C\pi(z)\mathrm{d}z$ where $C > 0$ and

$$\pi(z) := \int_{\mathscr{X}} \mathrm{e}^{-\ell(z - G(u))}\, \mu_0(\mathrm{d}u).$$

Note that $\pi(z)$ is well-defined since $|e^{-\ell(z-G(u))}| \leq 1$ and $\pi \in L^1(\mathbb{R}^k)$ due to Fubini's theorem [25, Theorem 1.27]. In particular, we have that $(U, Z) \sim \mu$ with $\mu(du, dz) = C e^{-\ell(z-G(u))}\, \mu_0(du) \otimes dz$ where dz again denotes Lebesgue measure on $\mathbb{R}^k$. Further, we define the *potential*

$$\Phi(u; z) := \ell(z - G(u))$$

and assume the following to be satisfied.

Assumption 7.2. *1. The potential Φ is continuous w.r.t. z in mean-square sense w.r.t. μ_0, i.e, there exists an increasing function $\psi : [0, \infty) \to [0, \infty)$ with $\lim_{s\to 0} \psi(s) = \psi(0) = 0$ such that*

$$\int_{\mathcal{X}} |\Phi(u; z) - \Phi(u; z')|^2\, \mu_0(du) \leq \psi(|z - z'|).$$

For instance, there may exist a function $\theta \in L^2(\mathcal{X}, \mathcal{B}(\mathcal{X}), \mu_0; \mathbb{R})$ such that

$$|\Phi(u; z) - \Phi(u; z')| \leq \theta(u)\, \psi(|z - z'|).$$

Before stating the abstract version of Bayes' Rule in Theorem 7.1, we recall the finite-dimensional case $\mathcal{X} \simeq \mathbb{R}^n$ where it can be stated in terms of densities: here $\mu_0(du) = \pi_0(u)du$ and *Bayes' rule* takes the form

$$\pi^z(u) = \frac{1}{\pi(z)} \exp(-\Phi(u; z))\, \pi_0(u)$$

where $e^{-\Phi(u;z)} = e^{-\ell(z-G(u))}$ represents the *likelihood* of observing z when fixing u. The denominator $\pi(z)$ can be interpreted as a *normalizing constant* such that $\int_{\mathcal{X}} \pi^z(u)\, du = 1$. We now show that, in the general setting, Bayes' rule yields (a version of) the (regular) conditional measure $\mu_{U|Z}$ of U w.r.t. Z.

Theorem 7.1 (cf. [42, Theorems 4.2 and 6.31]). *Let Assumptions 7.1 and 7.2 be satisfied and define for each $z \in \mathbb{R}^k$ a probability measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ by*

$$\mu^z(du) := \frac{1}{\pi(z)} \exp(-\Phi(u; z))\, \mu_0(du). \tag{7.7}$$

Then the mapping $Q : \mathcal{B}(\mathcal{X}) \times \mathbb{R}^k$ given by

$$Q(B, z) := \mu^z(B) \qquad \forall B \in \mathcal{B}(\mathcal{X})$$

is a regular conditional distribution of U given Z. We call μ^z the posterior measure (of U given $Z = z$). *Moreover, μ^z depends continuously on z w.r.t. the Hellinger metric, i.e., for any $z_1, z_2 \in \mathbb{R}^k$ with $|z_1 - z_2| \le r$ there holds*

$$d_H(\mu^{z_1}, \mu^{z_2}) \le C_r(z_1)\, \psi(|z_1 - z_2|),$$

where $C_r(z_1) = C(1 + \min\{\pi(z') : |z_1 - z'| \le r\}^3)^{-1} < +\infty$.

Proof. Continuity with respect to the Hellinger metric is a slight generalization of [42, Theorem 4.2] and may be proved in the same way with obvious modifications. To show that Q is a regular conditional distribution we verify the two properties (a) and (b) given in Sect. 7.2.2.2. The first follows from the construction of μ^z. For the second property, note that measurability follows from continuity. The continuity of μ^z w.r.t. z in the Hellinger metric implies also that $\mu^z(B)$ depends continuously on z due to the relations between Hellinger metric and total variation distance (see [18]). Finally, we have for any $A \in \mathscr{B}(\mathbb{R}^k)$ and $B \in \mathscr{B}(\mathscr{X})$ that

$$\begin{aligned}\mathbb{P}(U \in B, Z \in A) &= \int_{A\times B} \mu(\mathrm{d}u, \mathrm{d}z) = \int_A \int_B C\, \mathrm{e}^{-\ell(z-G(u))}\, \mu_0(\mathrm{d}u)\, \mathrm{d}z \\ &= \int_A C\pi(z) Q(B, z)\, \mathrm{d}z = \int_A Q(B, z)\, \mathbb{P}_Z(\mathrm{d}z)\end{aligned}$$

which completes the proof. □

Remark 7.2. We wish to emphasize that Theorem 7.1 and Assumption 7.2 show in detail the connection between the smoothness of the potential $\Phi(u; z) = \ell(z - G(u))$ and the continuity of the posterior μ^z w.r.t. z for a general prior μ_0 and an additive error ε with Lebesgue density proportional to $\mathrm{e}^{-\ell(\varepsilon)}$. Roughly speaking, the negative log-likelihood ℓ and the posterior μ^z share the same local modulus of continuity. This generalizes the results in [42] in that we allow for non-Gaussian priors μ_0 and errors ε.

Thus, under mild conditions, the Bayesian inverse problem is well-posed. It is also possible to prove continuity of μ^z w.r.t. to the forward map G, see [42, Section 4.4], which is crucial when the forward map G is realized by numerical approximation.

To give meaning to the mean and covariance of $U \sim \mu_0$ and $Z = G(U) + \varepsilon$, we make the further assumption that all second moments exist:

Assumption 7.3. *There holds*

$$\int_{\mathscr{X}} (\, \|u\|^2 + |G(u)|^2\,)\, \mu_0(\mathrm{d}u) < +\infty \quad \textit{and} \quad \int_{\mathbb{R}^k} |\varepsilon|^2\, \mu_\varepsilon(\mathrm{d}\varepsilon) < +\infty.$$

7.2.3 Bayes Estimators

Although the posterior measure μ^z is by definition the solution to the Bayesian inverse problem, it is, in general, by no means easy to compute in practice. In special cases, e.g., when G is linear and μ_0 and μ_ε are Gaussian measures, closed-form expressions for μ^z are available, but in general μ^z can only be computed in an approximate sense, see also Sect. 7.2.4. Moreover, when the dimension of $\mathscr{X}$ is large or infinite, visualizing, exploring or using μ^z for postprocessing are demanding tasks.

Other, more accessible quantities from Bayesian statistics, [3] which are also more similar to the result of deterministic parameter identification procedures than the posterior measure, are point estimates for the unknown u. In the Bayesian setting a point estimate is a "best guess" $\hat{u}$ of u based on posterior knowledge. Here "best" is determined by a *cost function* $c : \mathscr{X} \to \mathbb{R}_+$ satisfying $c(0) = 0$ and $c(u) \leq c(\lambda u)$ for any $u \in \mathscr{X}$ and $\lambda \geq 1$. This cost function describes the loss or costs $c(u - \hat{u})$ incurred when $\hat{u}$ is substituted for (the true) u for post processing or decision making. Note that also more general forms of a cost function are possible, see, e.g., [2,3].

For any realization $z \in \mathbb{R}^k$ of the observation RV Z we introduce the *(posterior) Bayes cost* of the estimate $\hat{u}$ w.r.t. c as

$$\mathrm{B}_c(\hat{u}; z) := \int_{\mathscr{X}} c(u - \hat{u})\, \mu^z(\mathrm{d}u),$$

and define the *Bayes estimate* $\hat{u}$ as a minimizer of this cost, i.e.,

$$\hat{u} := \operatorname*{argmin}_{u' \in \mathscr{X}} \mathrm{B}_c(u'; z),$$

assuming that such a minimizer exists. The *Bayes estimator* $\hat{\phi} : \mathbb{R}^k \to \mathscr{X}$ is then the mapping which assigns to an observation z the associated Bayes estimate $\hat{u}$, i.e.,

$$\hat{\phi} : z \mapsto \operatorname*{argmin}_{u' \in \mathscr{X}} \mathrm{B}_c(u'; z)\ .$$

We assume measurability of $\hat{\phi}$ in the following. Note that $\hat{\phi}$ is then also the minimizer of the *(prior) Bayes cost*

$$\mathrm{B}_c(\hat{\phi}) := \int_{\mathbb{R}^k} \mathrm{B}_c(\hat{\phi}(z); z)\, \mu_Z(\mathrm{d}z) = \mathbb{E}\left[\mathrm{B}_c(\hat{\phi}(Z); Z)\right],$$

i.e., there holds

$$\mathbb{E}\left[\mathrm{B}_c(\hat{\phi}(Z); Z)\right] \leq \mathbb{E}\left[\mathrm{B}_c(\phi(Z); Z)\right]$$

for any other measurable $\phi : \mathbb{R}^k \to \mathscr{X}$.

Remark 7.3. Since $\hat{\phi} = \operatorname{argmin}_{\phi} \mathrm{B}_c(\phi)$ it is possible to determine the estimator $\hat{\phi}$ and thereby also the estimate $\hat{u} = \hat{\phi}(z)$ for a given z without actually computing the posterior measure μ^z, as the integrals in $\mathrm{B}_c(\hat{\phi})$ are w.r.t. the prior measure. Therefore, Bayes estimators are typically easier to approximate than μ^z.

We now introduce two very common Bayes estimators: the *posterior mean estimator* and the *maximum a posteriori estimator*. For the remainder of the discussion we assume that $\mathscr{X}$ is a separable Hilbert space.

7.2.3.1 Posterior Mean Estimator

For the cost function $c(u) = \|u\|^2$ the posterior Bayes cost

$$\mathrm{B}_c(\hat{u}; z) = \int_{\mathscr{X}} \|u - \hat{u}\|^2 \, \mu^z(\mathrm{d}u)$$

is minimized by the posterior mean $\hat{u} = u_{\mathrm{CM}} := \int_{\mathscr{X}} u \, \mu^z(\mathrm{d}u)$. The corresponding Bayes estimator for $c(u) = \|u\|^2$ is then given by

$$\hat{\phi}_{\mathrm{CM}}(z) := \int_{\mathscr{X}} u \, \mu^z(\mathrm{d}u).$$

There holds in particular $\hat{\phi}_{\mathrm{CM}}(Z) = \mathbb{E}[U|Z]$ $\mathbb{P}$-almost surely.

Recall that, $\mathbb{E}[U|Z]$ is the best approximation of U in $L^2(\Omega, \sigma(Z), \mathbb{P}; \mathscr{X})$ w.r.t. the norm in $L^2(\Omega, \mathscr{F}, \mathbb{P}; \mathscr{X})$. Hence, the Bayes estimator $\hat{\phi}_{\mathrm{CM}}(Z) = \mathbb{E}[U|Z]$ represents the best L^2-approximation to U w.r.t. the information $\sigma(Z)$ available from the observation process Z.

7.2.3.2 Maximum A Posteriori Estimator

Another common estimator in Bayesian statistics is the *maximum a posteriori (MAP)* estimator $\hat{\phi}_{\mathrm{MAP}}$. For finite-dimensional $\mathscr{X} \simeq \mathbb{R}^n$ and absolutely continuous prior μ_0, i.e., $\mu_0(\mathrm{d}u) = \pi_0(u)\mathrm{d}u$, the MAP estimate is defined as

$$\hat{\phi}_{\mathrm{MAP}}(z) = \operatorname*{argmin}_{u \in \mathbb{R}^n} \Phi(u, z) - \log \pi_0(u)$$

provided the minimum exists for all $z \in \mathbb{R}^k$. For the definition of the MAP estimate via a cost function and the Bayes cost, we refer to the literature, e.g., [28, Section 16.2] or the recent work [5] fora novel approach; for MAP estimates in infinite dimensions, we refer to [8].

There is an interesting link between the Bayes estimator $\hat{\phi}_{\mathrm{MAP}}$ and the solution of the associated regularized least-squares problem: If $\mathrm{R} : \mathbb{R}^n \to [0, \infty)$ is a

regularizing functional which satisfies $\int_{\mathbb{R}^n} \exp(-\frac{\alpha}{\sigma^2} \mathrm{R}(u))\ \mathrm{d}u < +\infty$, then the solution $\hat{u}_\alpha = \operatorname{argmin} |z - G(u)|^2 + \alpha \mathrm{R}(u)$ corresponds to the MAP estimate $\hat{\phi}_{\mathrm{MAP}}(z)$ for $\varepsilon \sim N(0, \sigma^2 I)$ and $\mu_0(\mathrm{d}u) \propto \exp(-\frac{\alpha}{\sigma^2} \mathrm{R}(u))\ \mathrm{d}u$.

7.2.4 Computational Methods for Bayesian Inverse Problems

We summarize the most common methods for computing the posterior measure and Bayes estimators, referring to the cited literature for details.

In finite dimensions $\mathscr{X} \simeq \mathbb{R}^n$ and in the case of *conjugate priors*, see, e.g., [20], the posterior density is available in closed form since in this case the product of the prior density and the likelihood function belongs to the same class of probability densities as the prior. Therefore only the parameters of the posterior need to be computed, and for these analytical formulas are often available.

Aside from these special cases μ^z can only be approximated – but how may a probability distribution, possibly on an infinite-dimensional space, be approximated computationally? Perhaps the simplest and most natural idea is to generate samples distributed according to the posterior measure. A well-known method for this purpose is the *Markov Chain Monte Carlo* method (MCMC). The idea here is to construct a Markov chain with the posterior measure as its stationary resp. limiting distribution. If such a chain is run sufficiently long, it will yield (correlated) samples which are asymptotically distributed according to the posterior measure. For details we refer to [17] and, for the underlying theory of Markov chains, to [30]. The computational efficiency of the chain mainly depends on its transition kernel. Recently, much research has been devoted towards constructing good kernels. We mention [7] for MCMC suited for very high and even infinite dimensions, [19] for the idea of adapting the kernel to geometrical features of the posterior and [29], where this idea is realized by a transition kernel derived from the Gauss-Newton method.

Besides MCMC another common Bayesian method are *particle filters* [24, Section 4.3]. Here samples are generated according to the prior and all samples are assigned initially equal weights. Then, in an updating step, the weights are modified according to the posterior distribution. A further extension, *Gaussian mixture filters* [41], approximate the posterior density by a weighted mean of Gaussian kernels located at samples/particles. Here, in addition to the weights, also the location of the particles are modified according to the posterior.

A further technique for sampling from the posterior is presented in [9]: here a mapping $F : \mathscr{X} \to \mathscr{X}$ is constructed in such a way that $F(U) \sim \mu^z$ for a random variable $U \sim \mu_0$. Given F, which is obtained by solving an optimal transport problem, samples according to μ^z can then easily be generated by evaluating F for samples from the prior.

For the posterior mean, the immediate computational method is numerical integration w.r.t. $\mu^z(\mathrm{d}u)$ or $\mathrm{e}^{-\Phi(u;z)}\mu_0(\mathrm{d}u)$. A Monte Carlo integration is again

performed by averaging samples generated by a suitable Markov chain. Recently, sparse quadrature methods based on known quadrature rules for μ_0 have been investigated, see [37, 38]. Due to assumed smoothness of the likelihood $\mathrm{e}^{-\Phi(u;z)}$ w.r.t. u, these methods can yield faster convergence rates than Monte Carlo/MCMC integration and are also suited to infinite dimensions.

Alternatively, the corresponding Bayes estimator ϕ_{CM} could be approximated, e.g., by linear functions, and simply evaluated for the observational data. We return to this approach in Sect. 7.3.3 and show that Kalman filters may be viewed as approximation methods of this type.

Computing the MAP estimate is, by construction, a minimization problem for the posterior density and related to classical Tikhonov regularization. Therefore, methods from numerical optimization and computational inverse problems, respectively, can be applied here [10, 44]. Note that in numerical weather prediction the popular methods *3DVar* and *4DVar* are precisely computations of the MAP estimate. The difference between both is that 3DVar treats the typically sequential data recursively, while 4DVar performs the optimization w.r.t. the entire data set at once, see also [28].

7.3 Analysis of Kalman Filters for Bayesian Inverse Problems

In this section we consider Kalman filters and their application to the nonlinear Bayesian inverse problem (7.5). We begin with the classical Kalman filter for state estimation in linear dynamics and then consider two generalizations to the nonlinear setting which have been recently proposed for UQ in inverse problems. We show that both methods can be understood as discretizations of the same updating scheme for a certain RV and analyze the properties of this updated variable, thereby characterizing the properties of the approximations provided by the two filtering methods. In particular, we show that Kalman filters do not solve the nonlinear Bayesian inverse problem, nor can they be justified as approximations to its solution. They are, rather, related to a linear approximation of the Bayes estimator ϕ_{CM} and its estimation error.

7.3.1 The Kalman Filter

The Kalman filter [26] is a well-known method for sequential state estimation for incompletely observable, linear discrete-time dynamics, see, e.g., [6, 39] for a broader introduction and discussion. Thus, the Kalman filter may be applied to systems of the form

$$U_n = A_n U_{n-1} + \eta_n, \qquad Z_n = G_n U_n + \varepsilon_n, \qquad n = 1, 2, \ldots \tag{7.8}$$

where U_n denotes the unknown, unobservable state and Z_n the observable process at time n, and where U_0, η_n and $\boldsymbol{\varepsilon}_n$ are mutually independent RVs. The operators A_n and G_n are linear mappings in state space and from state to observation space, respectively. For the noises η_n and ε_n, zero mean and given covariances Γ_n and Σ_n, respectively, are assumed. Then, given observations $Z_1 = z_1, \ldots, Z_n = z_n$ of the process Z, the state U_n is to be inferred. Assume an initial guess $\hat{u}_0$ of the unknown U_0 with minimal variance trace(E_0) where $E_0 := \mathrm{Cov}(U_0 - \hat{u}_0)$ denotes the error covariance of the estimate $\hat{u}_0$. Then the Kalman filter results in recursive equations for the minimum variance estimates $\hat{u}_n$ of U_n and their error covariances $E_n := \mathrm{Cov}(U_n - \hat{u}_n)$.

Although the main advantage of the Kalman filter is its sequential structure which allows for a significant reduction of computational work (see [42, Section 5.3] for a nice discussion on this topic) we will apply the Kalman filter to our stationary inverse problem

$$Z = GU + \varepsilon, \qquad U \sim N(m_0, C_0), \quad \varepsilon \sim N(0, \Sigma), \tag{7.9}$$

which is, of course, only a special case of the system (7.8) in that there are no dynamics, $A_n \equiv I$, $\eta_n \equiv 0$ and only a single update $n = 1$. If we take $\hat{u}_0 = m_0$ as the initial guess this yields $E_0 = C_0$ and the Kalman filter yields the updates

$$\hat{u}_1 = \hat{u}_0 + K(z - G\hat{u}_0), \qquad E_1 = E_0 - KGE_0$$

where $K = E_0 G^*(GE_0 G^* + \Sigma)^{-1}$ is the well-known *Kalman gain*.

In the Gaussian case (7.9), for which (U, Z) is a jointly Gaussian RV, the posterior measure μ^z is again Gaussian, i.e., $\mu^z = N(m^z, C^z)$. Moreover, the posterior mean m^z and the posterior covariance C^z are given by

$$m^z = m_0 + K(z - Gm_0), \qquad C^z = C_0 - KGC_0,$$

where $K = C_0 G^*(GC_0 G^* + \Sigma)^{-1}$. Thus, for (7.9) the Kalman filter is seen to yield the solution of the Bayesian inverse problem by providing the posterior mean and covariance. However, we emphasize that the Kalman filter does not directly approximate the posterior measure. The filter provides estimates and error covariances which, in the Gaussian case, coincide with the posterior mean and covariance which, in turn, uniquely determine a Gaussian posterior measure. Whenever the linearity of G or Gaussianity of the prior $U \sim \mu_0$ or noise $\varepsilon \sim N(0, \Sigma)$ do not hold, then neither does the Kalman filter yield the first two posterior moments nor is the posterior measure necessarily Gaussian. We will return to the interpretation of the Kalman filter for linear G but non-Gaussian U or ε in Sect. 7.3.3.

7.3.2 *Kalman Filter Extensions for Nonlinear Inverse Problems*

Besides the extended Kalman filter (EKF), which is based on linearizations of the nonlinear forward map G but which we shall not consider here, a widely used method for nonlinear systems is the Ensemble Kalman Filter (EnKF) introduced by Evensen [13]. In addition, a more recent development, the Polynomial Chaos Expansion Kalman Filter (PCE-KF) developed by Matthies et al. [32, 34, 35] can also be applied to the nonlinear inverse problem (7.5).

7.3.2.1 The Ensemble Kalman Filter

Since its introduction in 1994, the EnKF has been investigated and evaluated in many publications [4, 14–16, 31]. However, the focus is usually on its application to state or parameter estimation rather than solving Bayesian inverse problems. Recently, the interest in the EnKF for UQ in inverse problems has increased, see, e.g., [22, 23, 27].

If we consider $Z = G(U) + \varepsilon$ with $U \sim \mu_0$ and $\varepsilon \sim \mu_\varepsilon$ and given observations $z \in \mathbb{R}^k$, the EnKF algorithm proceeds as follows:

1. **Initial ensemble:** Generate samples $u_1, \ldots, u_M$ of U according to μ_0.
2. **Forecast:** Generate samples $z_1, \ldots, z_M$ of Z by

$$z_j = G(u_j) + \varepsilon_j, \qquad j = 1, \ldots, M,$$

   where $\varepsilon_1, \ldots, \varepsilon_M$ are samples of ε according to μ_ε.
3. **Analysis:** Update the inital ensemble $\boldsymbol{u} = (u_1, \ldots, u_M)$ member by member via

$$u_j^a = u_j + \tilde{K}(z - z_j), \qquad j = 1, \ldots, M, \tag{7.10}$$

   where $\tilde{K} = \mathrm{Cov}(\boldsymbol{u}, \boldsymbol{z})\mathrm{Cov}(\boldsymbol{z})^{-1}$ and $\mathrm{Cov}(\boldsymbol{u}, \boldsymbol{z})$ and $\mathrm{Cov}(\boldsymbol{z}) = \mathrm{Cov}(\boldsymbol{z}, \boldsymbol{z})$ are the empirical covariances of the samples $\boldsymbol{u}$ and $\boldsymbol{z} = (z_1, \ldots, z_M)$. This yields an *analysis ensemble* $\boldsymbol{u}^a = (u_1^a, \ldots, u_M^a)$.
   The empirical mean of $\boldsymbol{u}^a$ serves as estimate $\hat{u}$ for the unknown u and the empirical covariance of $\boldsymbol{u}^a$ as an indicator for the accuracy of the estimate.

Note that for dynamical systems such as (7.8), the analysis ensemble $A_n(\boldsymbol{u}^a)$ serves as the initial ensemble for the next step n.

7.3.2.2 The Polynomial Chaos Expansion Kalman Filter

In [32, 34, 35] the authors propose a sampling-free Kalman filtering scheme for nonlinear systems. Rather than updating samples of the unknown, this is carried

out for the coefficient vector of a polynomial chaos expansion of the unknown. This necessitates the construction of a polynomial chaos expansion distributed according to the prior measure μ_0: we assume there exist countably many independent real-valued random variables $\boldsymbol{\xi} = (\xi_m)_{m\in\mathbb{N}}$, and *chaos coefficients* $u_{\boldsymbol{\alpha}} \in \mathscr{X}$, $\varepsilon_{\boldsymbol{\alpha}} \in \mathbb{R}^k$ for each

$$\boldsymbol{\alpha} \in \mathbb{J} := \{\boldsymbol{\alpha} \in \mathbb{N}_0^{\mathbb{N}} : \alpha_j \neq 0 \text{ for only finitely many } j\},$$

such that

$$\sum_{\boldsymbol{\alpha}\in\mathbb{J}} \|u_{\boldsymbol{\alpha}}\|^2 < +\infty \quad \text{and} \quad \sum_{\boldsymbol{\alpha}\in\mathbb{J}} |\varepsilon_{\boldsymbol{\alpha}}|^2 < +\infty,$$

and

$$\Big(\sum_{\boldsymbol{\alpha}\in\mathbb{J}} u_{\boldsymbol{\alpha}} P_{\boldsymbol{\alpha}}(\boldsymbol{\xi}), \quad \sum_{\boldsymbol{\alpha}\in\mathbb{J}} \varepsilon_{\boldsymbol{\alpha}} P_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\Big) \sim \mu_0 \otimes \mu_\varepsilon.$$

Here, $P_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = \prod_{m\geq 1} P_{\alpha_m}^{(m)}(\xi_m)$ denotes the product of univariate orthogonal polynomials $P_{\alpha_m}^{(m)}$ where we require $\{P_\alpha^{(m)}\}_{\alpha\in\mathbb{N}}$ to be a CONS in $L^2(\Gamma_m, \mathscr{B}(\Gamma_m), \mathbb{P}_{\xi_m})$, $\Gamma_m = \xi_m(\Omega) \subseteq \mathbb{R}$. Note, that the completeness of orthogonal polynomials will depend in general on properties of the measure $\mathbb{P}_{\xi_m}$, see [12] for a complete characterization.

We then define $U := \sum_{\boldsymbol{\alpha}\in\mathbb{J}} u_{\boldsymbol{\alpha}} P_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$ and $\varepsilon := \sum_{\boldsymbol{\alpha}\in\mathbb{J}} \varepsilon_{\boldsymbol{\alpha}} P_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$, denoting their PCE vectors $(u_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{J}}$ and $(\varepsilon_{\boldsymbol{\alpha}})_{\boldsymbol{\alpha}\in\mathbb{J}}$ by $[U]$ and $[\varepsilon]$. For the same problem considered for the EnKF, the PCE-KF algorithm is as follows.

1. **Initialization:** Compute a PCE with coefficient vector $[U]$ such that $U \sim \mu_0$.
2. **Forecast:** Compute the PC vector $[G(U)]$ of $G(U)$ and set

$$[Z] := [G(U)] + [\varepsilon],$$

   where $[\varepsilon]$ is a PC vector such that then $\varepsilon \sim \mu_\varepsilon$. Note that, by linearity, $[Z]$ is the PC vector of the RV defined by $Z := G(U) + \varepsilon$.
3. **Analysis:** Update the inital PC vector by

$$[U]^a = [U] + K \otimes I_{\mathbb{J}} \,([z] - [Z]), \tag{7.11}$$

   where $[z] = (z, 0, \ldots)$ is the PC vector of the observed data $z \in \mathbb{R}^k$ and $K := \mathrm{Cov}(U, Z)\mathrm{Cov}(Z)^{-1}$. The action of the covariances as operators can be described, e.g. in the case of $\mathrm{Cov}(U, Z) : \mathbb{R}^k \to \mathscr{X}$, by

$$\mathrm{Cov}(U, Z)z = \sum_{\boldsymbol{\alpha}\in\mathbb{J}} \sum_{\boldsymbol{\beta}\in\mathbb{J}} z_{\boldsymbol{\beta}}^\top z \, u_{\boldsymbol{\alpha}}.$$

The result of one step of the PCE-KF algorithm is an *analysis PC vector* $[U]^a$.

Remark 7.4. Neither the independence of the $\{\xi_m\}_{m\in\mathbb{N}}$ nor an expansion in polynomials $\{P_{\boldsymbol{\alpha}}(\boldsymbol{\xi})\}$ is crucial for the PCE-KF. In principle, only a countable CONS $\{\Psi_\alpha\}_{\alpha\in\mathbb{N}}$ for the space $L^2(\boldsymbol{\Gamma},\mathscr{B}(\boldsymbol{\Gamma}),\mathbb{P}_{\boldsymbol{\xi}})$, $\boldsymbol{\Gamma} = \boldsymbol{\xi}(\Omega) \subseteq \mathbb{R}^{\mathbb{N}}$, is required such that $\big(\sum_\alpha u_\alpha\Psi_\alpha(\boldsymbol{\xi}), \sum_\alpha \varepsilon_\alpha\Psi_\alpha(\boldsymbol{\xi})\big) \sim \mu_0 \otimes \mu_\varepsilon$. However, the independence structure of $\mu_0 \otimes \mu_\varepsilon$ requires at least two independent random vectors $\boldsymbol{\eta} = (\eta_1,\ldots,\eta_M)$, $\boldsymbol{\zeta} = (\zeta_1,\ldots,\zeta_N)$, $\boldsymbol{\xi} = (\boldsymbol{\eta},\boldsymbol{\zeta})$, and expansions of the form $\sum_\alpha u_\alpha\Psi_\alpha(\eta_1,\ldots,\eta_M)$ and $\sum_\alpha \varepsilon_\alpha\Psi_\alpha(\zeta_1,\ldots,\zeta_N)$.

7.3.2.3 The Analysis Variable

Note that the analysis PC vector $[U]^a$ defines an *analysis variable* $U^a := \sum_{\boldsymbol{\alpha}\in\mathbb{J}} u^a_{\boldsymbol{\alpha}} P_{\boldsymbol{\alpha}}(\boldsymbol{\xi})$. Indeed, both EnKF and PCE-KF perform discretized versions of an update for RVs, namely,

$$U^a = U + K(z - Z), \qquad K = \mathrm{Cov}(U,Z)\mathrm{Cov}(Z)^{-1},$$

where $Z := G(U) + \varepsilon$, and $(U,\varepsilon) \sim \mu_0 \otimes \mu_\varepsilon$, providing samples $\boldsymbol{u}^a$ and PCE vectors $[U]^a = [U^a]$ of U^a, respectively. This raises the question of how the analysis variable U^a is to be understood in context of Bayesian inverse problems?

7.3.3 The Linear Conditional Mean

To relate the results produced by the EnKF or PCE-KF to the Bayesian setting, we introduce a new Bayes estimator, or, more precisely, a linear approximation to the Bayes estimator $\hat{\phi}_{\mathrm{CM}}$ resp. the conditional mean $\mathbb{E}[U\,|\,Z]$. The *linear posterior mean estimator* $\hat{\phi}_{\mathrm{LCM}}$ is given by

$$\hat{\phi}_{\mathrm{LCM}} = \operatorname*{argmin}_{\phi\in\mathrm{span}\{1,z\}} \mathbb{E}\big[\|U - \phi(Z)\|^2\big], \tag{7.12}$$

here

$$\mathrm{span}\{1,z\} = \{\phi : \phi(z) = b + Az \text{ with } b \in \mathscr{X},\, A : \mathbb{R}^k \to \mathscr{X} \text{ linear and bounded}\}$$

Moreover, we refer to the RV $\hat{\phi}_{\mathrm{LCM}}(Z)$ as the *linear conditional mean*. Thus, $\hat{\phi}_{\mathrm{LCM}}(Z)$ is the best $L^2(\Omega,\mathscr{F},\mathbb{P};\mathscr{X})$-approximation to $U \sim \mu_0$ in the subspace $\mathrm{span}\{1,Z\} \subset L^2(\Omega,\sigma(Z),\mathbb{P};\mathscr{X})$. Or, alternatively, ϕ_{LCM} is the linear estimator with minimal prior Bayes cost for $c(u) = \|u\|^2$. Furthermore, there holds

$$\hat{\phi}_{\mathrm{LCM}}(z) = \mathbb{E}[U] + K(z - \mathbb{E}[Z]),$$

with the usual Kalman gain $K = \mathrm{Cov}(U,Z)\mathrm{Cov}(Z)^{-1}$, and we immediately obtain the following result.

Theorem 7.2. *Consider* (7.5) *and let Assumptions 7.1–7.3 be satisfied. Then for any* $z \in \mathbb{R}^k$ *the* analysis variable $U^a = U + K(z - Z)$, $K = \mathrm{Cov}(U, Z)\mathrm{Cov}(Z)^{-1}$, *coincides with*

$$U^a = \hat{\phi}_{\mathrm{LCM}}(z) + (U - \hat{\phi}_{\mathrm{LCM}}(Z)).$$

In particular, there holds

$$\mathbb{E}[U^a] = \hat{\phi}_{\mathrm{LCM}}(z) \quad \textit{and} \quad \mathrm{Cov}(U^a) = \mathrm{Cov}(U) - K\mathrm{Cov}(Z, U).$$

We summarize the consequences of Theorem 7.2 as follows:

- The analysis variable U^a, to which the EnKF and the PCE-KF provide approximations, is the sum of a Bayes estimate $\hat{\phi}_{\mathrm{LCM}}(z)$ and the prior error $U - \hat{\phi}_{\mathrm{LCM}}(Z)$ of the corresponding Bayes estimator $\hat{\phi}_{\mathrm{LCM}}$.
- The resulting mean of the EnKF analysis ensemble or the PCE-KF analysis vector corresponds to the linear posterior mean estimate and therefore provides an approximation to the true posterior mean.
- The covariance approximated by the empirical covariance of the EnKF analysis ensemble, as well as that of the PCE-KF analysis vector, is independent of the actual observational data $z \in \mathbb{R}^k$. It therefore constitutes a prior rather than a posterior measure of uncertainty.
- In particular, the randomness in U^a is entirely determined by the prior measures μ_0 and μ_ε. Only the location, i.e., the mean, of U^a is influenced by the observation data z; the randomness of U^a is independent of z and determined only by the projection error $U - \hat{\phi}_{\mathrm{LCM}}(Z)$ w.r.t. the prior measures.
- By the last two items, the analysis variable U^a, and therefore the EnKF analysis ensemble or the result of the PCE-KF, are in general not distributed according to the posterior measure μ^z. Moreover, the difference between μ^z and the distribution of U^a depends on the data z and can become quite large for nonlinear problems, see Example 7.3.

Remark 7.5. Note that in particular the second and third item above explain the observations made in [27], i.e., that "[...] (i) with appropriate parameter choices, approximate filters can perform well in reproducing the mean of the desired probability distribution, (ii) they do not perform as well in reproducing the covariance [...] ".

We illustrate the conceptual difference between the distribution of the analysis variable U^a and the posterior measure μ^z with a simple yet striking example.

Example 7.3. We consider $U \sim N(0, 1)$, $\varepsilon \sim N(0, \sigma^2)$ and $G(u) \equiv u^2$. Given data $z \in \mathbb{R}$, the posterior measure, obtained from Bayes' rule for the densities, is

$$\mu^z(\mathrm{d}u) = C \exp\left(-\frac{\sigma^2 u^2 + (z - u^2)^2}{2\sigma^2}\right) \mathrm{d}u.$$

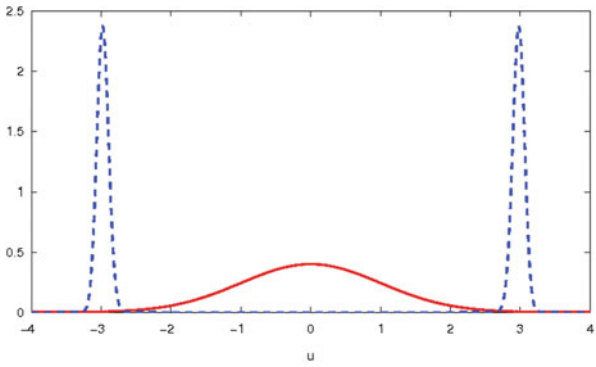


Fig. 7.1 Density of the posterior μ^z (*dashed, blue line*) and the probability density of the analysis variable U^a (*solid, red line*) for $z = 9$ and $\sigma = 0.5$

Due to the symmetry of μ^z we have $\hat{u}_{\mathrm{CM}} = \int_{\mathscr{X}} u\, \mu^z(\mathrm{d}u) = 0$ for any $z \in \mathbb{R}^k$. Thus, $\mathbb{E}[U|Z] \equiv 0$ and $\hat{\phi}_{\mathrm{LCM}} \equiv \hat{\phi}_{\mathrm{CM}}$. In particular, we have $K = 0$ due to

$$\mathrm{Cov}(U, Z) = \mathrm{Cov}(U, U^2) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} u(u^2 - 1)\mathrm{e}^{-u^2/2}\mathrm{d}u = 0,$$

which in turn yields $U^a = U \sim N(0, 1)$. Hence the analysis variable is distributed according to the prior measure. This is not surprising as, by definition, its mean is the best linear approximation to the posterior mean according to μ^z and its fluctuation is simply the prior estimation error $U - \hat{\phi}_{\mathrm{LCM}}(Z) = U - 0 = U$. This illustrates that U^a is suited for approximating the posterior mean, but not appropriate as a method for uncertainty quantification for the nonlinear inverse problem. As displayed in Fig. 7.1, the distribution of U^a can be markedly different from the true posterior distribution.

7.4 Numerical Example: 1D Elliptic Boundary Value Problem

To illustrate the application of the EnKF and PCE-KF to a simple Bayesian inverse problems, we consider the following PDE model on $D = [0, 1]$:

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(\exp(u_1)\frac{\mathrm{d}}{\mathrm{d}x}p(x)\right) = f(x), \qquad p(0) = p_0, \quad p(1) = u_2. \tag{7.13}$$

Here $u = (u_1, u_2)$ are the unknown parameters to be identified. The solution of (7.13) is given by

$$p(x) = p_0 + (u_2 - p_0)x + \exp(-u_1)\left(S_x(F) - S_1(F)\, x\right), \tag{7.14}$$

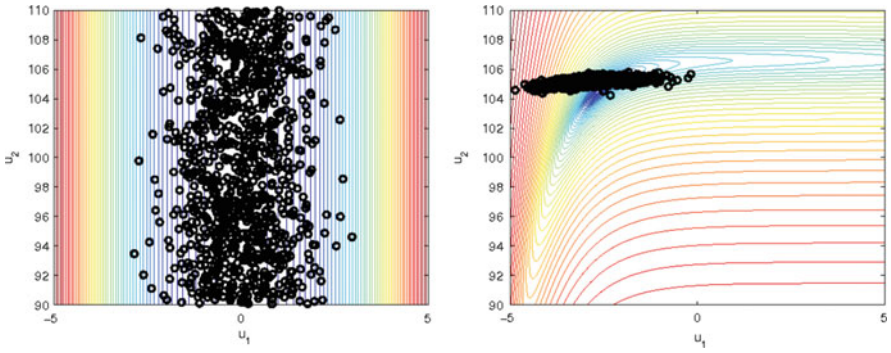


Fig. 7.2 *Left*: Contour plot of the negative logarithm of the prior density and the locations of 1,000 ensemble members of the initial EnKF-ensemble. *Right*: Contour plot of the logarithm of the negative logarithm of the posterior density and the locations of the updated 1,000 ensemble members in the analysis EnKF-ensemble

where $S_x(g) := \int_0^x g(y)\,\mathrm{d}y$ and $F(x) = S_x(f) = \int_0^x f(y)\,\mathrm{d}y$. For simplicity we choose $f \equiv 1$ and $p_0 = 0$ in the following.

Assume now that noisy measurements of p are available at $x_1 = 0.25$ and $x_2 = 0.75$, namely $z = (27.5, 79.7)$. We wish to infer u based on this data and on a priori information modelled by the prior distributions of the independent random variables

$$u_1 \sim N(0,1), \quad \text{and} \quad u_2 \sim \mathrm{Uni}(90,110).$$

Here $\mathrm{Uni}(90,110)$ denotes the uniform distribution on the interval $[90,110]$. Thus, the forward map here is $G(u) = (p(x_1), p(x_2))$ with p according to (7.14) for $f \equiv 1$, and the model for the measurement noise is $\varepsilon \sim N(0, 0.01 \cdot I_2)$.

In Fig. 7.2 we show the prior and the posterior densities as well as 1,000 ensemble members of the initial and analysis ensemble obtained by the EnKF. A total ensemble size of $M = 10^5$ was chosen in order to reduce the sampling error to a negligible level. It can be seen, however, that the analysis EnKF-ensemble does not follow the posterior distribution, although its mean $(-2.92, 105.14)$ is quite close to the true posterior mean $(-2.65, 104.5)$ (computed by quadrature). To illustrate the difference between the distribution of the analysis ensemble resp. variable and the true posterior distribution, we present the marginal posterior distributions of u_1 and u_2 in Fig. 7.3. For the posterior the marginals were evaluated by quadrature, whereas for the analysis ensemble we show a relative frequency plot.

We note that slightly changing the observational data to $\tilde{z} = (23.8, 71.3)$ moves the analysis ensemble resp. variable much closer to the true posterior, see Fig. 7.4. Also, the mean of the analysis ensemble $(0.33, 94.94)$ provides a better fit to the true posterior mean $(0.33, 94.94)$ here.

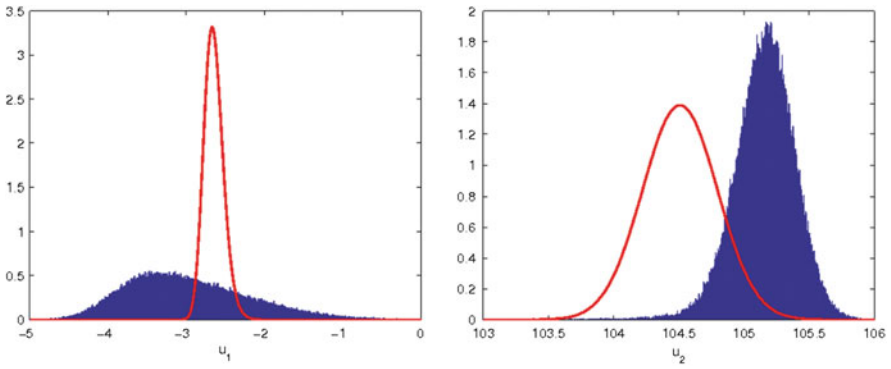


Fig. 7.3 *Left*: Posterior marginal and relative frequencies in the analysis ensemble for u_1. *Right*: The same for u_2

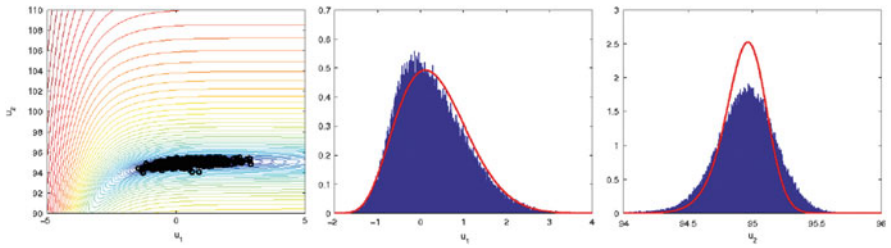


Fig. 7.4 *Left*: Contours of the logarithm of the negative log posterior density and locations of 1,000 members of the analysis EnKF-ensemble. *Middle*: Posterior marginal and relative frequencies in the analysis ensemble for u_1. *Right*: The same for u_2

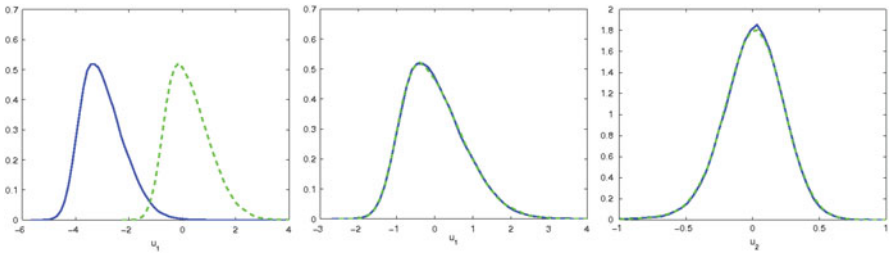


Fig. 7.5 *Left*: Kernel density estimates for $\boldsymbol{u}_1^a$ (*blue, solid line*) and $\tilde{\boldsymbol{u}}_1^a$ (*green, dashed line*). *Middle*: Kernel density estimates for $\boldsymbol{u}_1^a - \mathbb{E}[\boldsymbol{u}_1^a]$ (*blue, solid*) and $\tilde{\boldsymbol{u}}_1^a - \mathbb{E}[\tilde{\boldsymbol{u}}_1^a]$ (*green, dashed*). *Right*: Kernel density estimates for $\boldsymbol{u}_2^a - \mathbb{E}[\boldsymbol{u}_2^a]$ (*blue, solid*) and $\tilde{\boldsymbol{u}}_2^a - \mathbb{E}[\tilde{\boldsymbol{u}}_2^a]$ (*green, dashed*)

To reaffirm the fact that only the mean of analysis variable U^a depends on the actual data, we show density estimates for the marginals of u_1 and u_2 of U^a in Fig. 7.5. Here we have used once the data $z = (27.5, 79.7)$ (blue lines) and once $\tilde{z} = (23.8, 71.3)$ (green lines). The density estimates were obtained by normal kernel density estimation (KDE, in this case MATLAB's `ksdensity` routine) based on

the resulting analysis ensembles $(\boldsymbol{u}_1^a, \boldsymbol{u}_2^a)$ and $(\tilde{\boldsymbol{u}}_1^a, \tilde{\boldsymbol{u}}_2^a)$, respectively, of the EnKF for these two data sets $z, \tilde{z}$. In the left picture we show the KDE for $\boldsymbol{u}_1^a$ and $\tilde{\boldsymbol{u}}_1^a$ and in the middle picture we display the KDE for the corresponding centered ensembles $\boldsymbol{u}_1^a - \mathbb{E}[\boldsymbol{u}_1^a]$ and $\tilde{\boldsymbol{u}}_1^a - \mathbb{E}[\tilde{\boldsymbol{u}}_1^a]$. In the right picture we provide the KDEs for the centered ensembles of u_2. Note that the marginal distributions of the centered ensembles coincide, in agreement with Theorem 7.2.

However, note that, particularly in this example where the prior, and thus posterior, support for u_2 is bounded, the EnKF may yield members in the analysis ensemble which are outside this support. This is a further consequence of Theorem 7.2: Since the analysis ensemble of the EnKF follows the distribution of the analysis variable rather than that of the true posterior distribution, ensemble members lying outside the posterior support can always occur whenever the support of the analysis variable is not a subset of the support of the posterior.

In addition, we would like to stress that, whether or not the distribution of the analysis variable is a good fit to the true posterior distribution depends entirely on the observed data – which can neither be controlled nor are known a priori.

Applying the PCE-KF to this simple example problem can be done analytically. We require four basic independent random variables $\xi_1 \sim N(0,1)$, $\xi_2 \sim \mathrm{Uni}(0,1)$, $\xi_3 \sim N(0,1)$ and $\xi_4 \sim N(0,1)$ to define PCEs which yield random variables distributed according to the prior and error distributions:

$$U := (\xi_1,\ 90 + 20\xi_2)^\top \sim \mu_0, \qquad \varepsilon := (0.1\xi_3,\ 0.1\xi_4)^\top \sim \mu_\varepsilon.$$

Moreover, due to (7.14), $G(U)$ is also available in closed form as

$$G(U) = \begin{pmatrix} c_{11}(90 + 20\xi_2) + c_{12} \sum_{n=0}^{\infty} (-1)^n \frac{\sqrt{\mathrm{e}}}{\sqrt{n!}} H_n(\xi_1) \\ c_{21}(90 + 20\xi_2) + c_{22} \sum_{n=0}^{\infty} (-1)^n \frac{\sqrt{\mathrm{e}}}{\sqrt{n!}} H_n(\xi_1) \end{pmatrix},$$

where H_n denotes the nth normalized Hermite polynomial and $c_{11}, c_{12}, c_{21}, c_{22}$ can be deduced from inserting $x = 0.25$ and $x = 0.75$ into (7.14). Here, we have used the expansion of $\exp(-\xi)$ in Hermite polynomials, see also [43, Example 2.2.7]. Thus, the PCE coefficient vectors $[U]$ and $[G(U) + \varepsilon]$ w.r.t. the polynomials

$$P_{\boldsymbol{\alpha}}(\boldsymbol{\xi}) = H_{\alpha_1}(\xi_1)\, L_{\alpha_2}(\xi_2)\, H_{\alpha_3}(\xi_3)\, H_{\alpha_4}(\xi_4), \qquad \alpha \in \mathbb{N}_0^4,$$

can be obtained explicitly. Here H_α and L_α denote the αth normalized Hermite and Legendre polynomials, respectively. In particular, the nonvanishing chaos coefficients involve only the basis polynomials

$$P_0(\boldsymbol{\xi}) \equiv 1, \quad P_1(\boldsymbol{\xi}) = L_1(\xi_2), \quad P_2(\boldsymbol{\xi}) = H_1(\xi_3), \quad P_3(\boldsymbol{\xi}) = H_1(\xi_4)$$

and $P_\alpha(\boldsymbol{\xi}) = H_{\alpha-3}(\xi_1)$ for $\alpha \geq 4$. Arranging the two-dimensional chaos coefficients of U and $G(U)$ as the column vectors $[U], [G(U) + \varepsilon] \in \mathbb{R}^{2\times\mathbb{N}_0}$, and denoting by $[\dot{U}]$ the matrix $(u_1, u_2, \ldots) \in \mathbb{R}^{2\times\mathbb{N}}$ we get

$$K = [\dot{U}][G(\dot{U})]^\top \left([G(\dot{U})][G(\dot{U})]^\top + 0.01 I_2\right)^{-1}.$$

Thus, the only numerical error for applying the PCE-KF to the example is the truncation of the PCE. We have carried out this calculation using a truncated PCE of length $J = 4 + 50$ according to the reduced basis above, evaluated the approximation to K by using the truncated vector $[G(U)]$ in the formula above and then performed the update of the PCE vectors according to (7.11). We then sampled the resulting random variable U^a again $M = 10^5$ times. The resulting empirical distributions were essentially indistinguishable from the results obtained by the EnKF described previously and are therefore omitted.

Remark 7.6. Although a detailed complexity analysis of these methods is beyond the scope of this contribution, we would like to mention that the EnKF calls for M evaluations of the forward map $G(u_j)$, $j = 1, \ldots, M$, whereas the PCE-KF requires computing the chaos coefficients of $G(U)$ by, e.g., the Galerkin method. Thus the former yields, in general, many small systems to solve, whereas the latter typically requires the solution of a large coupled system. Moreover, we emphasize the computational savings by applying Kalman filters compared to a "full Bayesian update", i.e., sampling from the posterior measure by MCMC methods. In particular, each MCMC run one may require calculating many hundreds of thousands forward maps $G(u)$, e.g., for each iteration u_j of the Markov chain as in the case of Metropolis-Hastings MCMC. Hence, if one is interested in only the posterior mean as a Bayes estimate, then EnKF and PCE-KF provide substantially less expensive alternatives to MCMC for its approximation by means of the linear posterior mean.

7.5 Conclusions

We have contrasted the deterministic and Bayesian formulations of nonlinear inverse problems such as arise in parameter estimation and data assimilation settings. An important distinction lies in the objectives of the two approaches: the identification of a particular value of the unknown quantity in the deterministic case versus the updating of a prior to a posterior probability measure encoding the uncertainty associated with the unknown quantity due to new observations. Moreover, we have also pointed out the relation between regularized least-squares solutions and the concept of Bayesian (point) estimators. Among the computational methods for Bayesian inverse problems we have focused on Kalman filters such as the EnKF and PCE-KF and presented a precise characterization of these methods in the Bayesian setting. A summary of the contrasting features of Bayesian inversion, Bayes estimators and Kalman filter-based methods is given in Table 7.1.

Most important, the RVs approximated by the Kalman filter-based methods, will not, in general, be distributed according to the posterior distribution in the Bayes' sense. They are rather related to a common Bayes estimator – the linear conditional

Table 7.1 Distinguishing features of Bayesian inverse problems, Bayes estimators and Kalman filters

	Bayesian Inversion	Bayes Estimators	Kalman Filters
Goal	Merge prior belief with new observational data	Compute best guess w.r.t. posterior belief	Compute best linear guess and associated error
Result	Measure μ^z on $\mathscr{X}$	Estimate $\hat{u} \in \mathscr{X}$	Estimate $\hat{u} \in \mathscr{X}$ and estimation error $U - \hat{\phi}_{\mathrm{LCM}}(Z)$
Allows for	Rigorous UQ in post-processing	Deterministic post-processing with $\hat{u}$	Deterministic post-processing with $\hat{u}$ and certain UQ

mean – and its estimation error RV, and therefore represent a different uncertainty model than the posterior measure. Some carefully chosen numerical examples were given to illustrate these basic differences.

References

1. Anzengruber, S., Hofmann, B., Mathé, P.: Regularization properties of the sequential discrepancy principle for Tikhonov regularization in Banach spaces. Appl. Anal. **93**(7), 1382–1400 (2014)
2. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis, 2nd edn. Springer, New York (1985)
3. Bernardo, J.M.: Bayesian statistics. In: Viertl, R. (ed.) Probability and Statistics. Encyclopedia of Life Support Systems (EOLSS). UNESCO, Oxford (2003)
4. Burgers, G., van Leeuwen, P.J., Evensen, G.: Analysis scheme in the ensemble Kalman filter. Mon. Weather Rev. **126**, 1719–1724 (1998)
5. Burger, M., Lucka, F.: Maximum-A-Posteriori estimates in linear inverse problems with log-concave priors are proper Bayes estimators (2014). arXiv:1402.5297
6. Catlin, D.E.: Estimation, Control, and the Discrete Kalman Filter. Springer, New York (1989)
7. Cotter, S.L., Roberts, G.O., Stuart, A.M., White, D.: MCMC methods for functions: modifying old algorithms to make them faster. Stat. Sci. **28**(3), 283–464 (2013)
8. Dashti, M., Law, K.J.H., Stuart, A.M., Voss, J.: MAP estimators and their consistency in Bayesian nonparametric inverse problems. Inverse Probl. **29**(9), 095017:1–27 (2013)
9. El Moselhy, T.A., Marzouk, Y.M.: Bayesian inference with optimal maps. J. Comput. Phys. **231**(23), 7815–7850 (2012)
10. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (2000)
11. Engl, H.W., Kunisch, K., Neubauer, A.: Convergence rates for Tikhonov regularisation of non-linear ill-posed problems. Inverse Probl. **5**(4), 523–540 (1989)
12. Ernst, O.G., Mugler, A., Starkloff, H.J., Ullmann, E.: On the convergence of generalized polynomial chaos expansions. ESAIM: Math. Model. Numer. Anal. **46**(2), 317–339 (2012)
13. Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrohpic model using Monte Carlo methods to forecast error statistics. J. Geophys. Res. **99**(C5), 10143–10162 (1994)
14. Evensen, G.: The ensemble Kalman filter: theoretical formulation and practical implementation. Ocean Dyn. **53**, 343–367 (2003)

15. Evensen, G.: Data Assimilation: The Ensemble Kalman Filter, 2nd edn. Springer, New York (2009)
16. Evensen, G.: The ensemble Kalman filter for combined state and parameter estimation. Control Syst. Mag. **29**(3), 83–104 (2009)
17. Geyer, C.J.: Introduction to Markov Chain Monte Carlo. In: Brooks, S., Gelman, A., Jones, G.J., Meng, X.L. (eds.) Handbook of Markov Chain Monte Carlo. Handbooks of Modern Statistical Methods, pp. 3–48. CRC, Boca Raton (2011)
18. Gibbs, A.L., Su, F.E.: On choosing and bounding probability metrics. Int. Stat. Rev. **70**(3), 419–435 (2001)
19. Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. R. Stat. Soc.: Ser. B **73**(2), 123–214 (2010)
20. Hoff, P.D.: A First Course in Bayesian Statistical Methods. Springer, New York (2009)
21. Hofmann, B., Kaltenbacher, B., Pöschl, C., Scherzer, O.: A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. Inverse Probl. **23**(3), 987–1010 (2007)
22. Iglesias, M.A., Law, K.J.H., Stuart, A.M.: Ensemble Kalman methods for inverse problems. Inverse Probl. **29**(4), 045001:1–20 (2013)
23. Iglesias, M.A., Law, K.J.H., Stuart, A.M.: Evaluation of Gaussian approximations for data assimilation in reservoir models. Comput. Geosci. **17**(5), 851–885 (2013)
24. Kaipio, J., Somersalo, E.: Statistical and Computational Inverse Problems. Springer, New York (2005)
25. Kallenberg, O.: Foundations of Modern Probaility. Springer, New York (2002)
26. Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans. AMSE – J. Basic. Eng. **82**(Series D), 35–45 (1960)
27. Law, K.J.H., Stuart, A.M.: Evaluating data assimilation algorithms. Mon. Weather Rev. **140**, 3757–3782 (2012)
28. Lewis, J.M., Lakshmivarahan, S., Dhall, S.: Dynamic Data Assimilation – A Least Squares Appoach. Cambridge University Press, Cambridge (2006)
29. Martin, J., Wilcox, L.C., Burstedde, C., Ghattas, O.: A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. SIAM J. Sci. Comput. **34**(3), A1460–A1487 (2012)
30. Meyn, S., Tweedie, R.L.: Markov Chains and Stochastic Stability, 2nd edn. Cambridge University Press, Cambridge (2009)
31. Myrseth, I., Omre, H.: The ensemble Kalman filter and related filters. In: Biegler, L. (ed.) Large-Scale Inverse Problems and Quantification of Uncertainty. Wiley Series in Computational Statistics, pp. 217–246. Wiley, Chichester (2011)
32. Pajonk, O., Rosić, B.V., Litvinenko, A., Matthies, H.G.: A deterministic filter for non-gaussian bayesian estimation — applications to dynamical system estimation with noisy measurements. Phys. D: Nonlin. Phenom. **241**(7), 775–788 (2012)
33. Rao, M.M.: Conditional Measures and Applications. Chapman and Hall/CRC, Boca Raton (2010)
34. Rosić, B.V., Kučerová, A., Sýkora, J., Pajonk, O., Litvinenko, A., Matthies, H.G.: Parameter identification in a probabilistic setting. Eng. Struct. **60**, 179–196 (2013)
35. Rosić, B.V., Litvinenko, A., Pajonk, O., Matthies, H.G.: Sampling-free linear Bayesian update of polynomial chaos representations. J. Comput. Phys. **231**(17), 5761–5787 (2012)
36. Scherzer, O., Grasmair, M., Grossauer, H., Haltmeier, M., Lenzen, F.: Variational Methods in Imaging. Springer, New York (2009)
37. Schillings, C., Schwab, C.: Sparse, adaptive Smolyak quadratures for Bayesian inverse problems. Inverse Probl. **29**(6), 065011:1–28 (2013)
38. Schwab, C., Stuart, A.M.: Sparse deterministic approximation of Bayesian inverse problems. Inverse Probl. **28**(4), 045003:1–32 (2012)
39. Simon, D.: Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches. Wiley, Hoboken (2006)

40. Speyer, J.L., Chung, W.H.: Stochastic Processes, Estimation, and Control. SIAM, Philadelphia (2008)
41. Stordal, A.S., Karlsen, H.A., Nævdal, G., Skaug, H.J., Vallès, B.: Bridging the ensemble Kalman filter and particle filters: the adaptive Gaussian mixture filter. Comput. Geosci. **15**(2), 293–305 (2011)
42. Stuart, A.M.: Inverse problems: a Bayesian perspective. Acta Numer. **19**, 451–559 (2010)
43. Ullmann, E.: Solution strategies for stochastic finite element discretizations. Ph.D. thesis, TU Bergakademie Freiberg (2008)
44. Vogel, C.R.: Computational Methods for Inverse Problems. SIAM, Philadelphia (2002)

Chapter 8
Robustness in Stochastic Filtering and Maximum Likelihood Estimation for SDEs

Joscha Diehl, Peter K. Friz, Hilmar Mai, Harald Oberhauser, Sebastian Riedel, and Wilhelm Stannat

Abstract We consider complex stochastic systems in continuous time and space where the objects of interest are modelled via stochastic differential equations, in general high dimensional and with nonlinear coefficients. The extraction of quantifiable information from such systems has a long history and many aspects. We shall focus here on the perhaps most classical problems in this context: the filtering problem for nonlinear diffusions and the problem of parameter estimation, also for nonlinear and multidimensional diffusions. More specifically, we return to the question of robustness, first raised in the filtering community in the mid-1970s: will it be true that the conditional expectation of some observable of the signal process, given an observation (sample) path, depends continuously on the latter? Sadly, the answer here is no, as simple counterexamples show. Clearly, this is an unhappy state of affairs for users who effectively face an ill-posed situation: close observations may lead to vastly different predictions. A similar question can be asked in the context of (maximum likelihood) parameter estimation for diffusions. Some (apparently novel) counter examples show that, here again, the answer is no. Our contribution (Crisan et al., Ann Appl Probab **23**(5):2139–2160, 2013);

J. Diehl (✉) • S. Riedel • W. Stannat
Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: diehl@math.tu-berlin.de; riedel@math.tu-berlin.de; stannat@math.tu-berlin.de

P.K. Friz
Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany

Weierstrass Institute, Mohrenstrae 39, 10117 Berlin, Germany
e-mail: friz@math.tu-berlin.de

H. Mai
Weierstrass Institute, Mohrenstrae 39, 10117 Berlin, Germany
e-mail: Hilmar.Mai@wias-berlin.de

H. Oberhauser
University of Oxford, Eagle House, Walton Well Road, OX2 6ED Oxford, UK
e-mail: h.oberhauser@gmail.com

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_8

Diehl et al., A Levy-area between Brownian motion and rough paths with applications to robust non-linear filtering and RPDEs (2013, arXiv:1301.3799; Diehl et al., Pathwise stability of likelihood estimators for diffusions via rough paths (2013, arXiv:1311.1061) changed to yes, in other words: well-posedness is restored, provided one is willing or able to regard observations as rough paths in the sense of T. Lyons.

8.1 Introduction

The first part of this paper concerns the problem of stochastic filtering. Let a filtered probability space $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\geq 0}, \mathbb{P})$ be given, on which a two component diffusion process (X, Y) solving a stochastic differential equation is given, driven by a multidimensional Brownian motion. One assumes that the first component X is unobservable and the second component Y is observed. The filtering problem consists in computing the conditional distribution of the unobserved component, called the *signal* process, given the *observation* process Y. Equivalently, one is interested in computing

$$\pi_t(f) = \mathbb{E}[f(X_t, Y_t)|\mathscr{Y}_t],$$

where $\mathscr{Y} = \{\mathscr{Y}_t, t \geq 0\}$ is the observation filtration and f is a suitably chosen test function. An elementary measure theoretic result tells us[1] that there exists a Borel-measurable map $\theta_t^f : C([0,t], \mathbb{R}^{d_Y}) \to \mathbb{R}$, such that

$$\pi_t(f) = \theta_t^f(Y_\cdot) \qquad \mathbb{P}\text{-a.s.}, \tag{8.1}$$

where d_Y is the dimension of the observation state space and $Y_\cdot$ is the path-valued random variable

$$Y_\cdot : \Omega \to C([0,t], \mathbb{R}^{d_Y}),\ Y_\cdot(\omega) = (Y_s(\omega),\ 0 \leq s \leq t).$$

Of course, θ_t^f is not unique. Any other function $\bar{\theta}_t^f$ such that

$$\mathbb{P} \circ Y_\cdot^{-1}\left(\bar{\theta}_t^f \neq \theta_t^f\right) = 0,$$

where $\mathbb{P} \circ Y_\cdot^{-1}$ is the distribution of $Y_\cdot$ on the path space $C([0,t], \mathbb{R}^{d_Y})$ can replace θ_t^f in (8.1). It would be desirable to solve this ambiguity by choosing a suitable representative from the class of functions that satisfy (8.1). A *continuous* version, if it exists, would enjoy the following uniqueness property: if the law of the

[1] See, for example, Proposition 4.9 page 69 in [3].

observation $\mathbb{P} \circ Y_\cdot^{-1}$ positively charges all non-empty open sets in $C([0,t],\mathbb{R}^{d_Y})$ then there exists a unique continuous function θ_t^f that satisfies (8.1). In this case, we call $\theta_t^f(Y_\cdot)$ the *robust version* of $\pi_t(f)$ and Eq. (8.1) is the robust representation formula for the solution of the stochastic filtering problem.

The need for this type of representation arises when the filtering framework is used to model and solve "real-life" problems. As explained in a substantial number of papers (e.g. [5, 6, 9–13, 26]) the model chosen for the "real-life" observation process $\bar{Y}$ may not be a perfect one. However, if θ^f is continuous (or even locally Lipschitz, as in the setting of [6]), and as long as the distribution of $\bar{Y}_\cdot$ is close in a weak sense to that of $Y_\cdot$ (and some integrability assumptions hold), the estimate $\theta_t^f(\bar{Y}_\cdot)$ computed on the actual observation will still be reasonable, as $\mathbb{E}[(f(X_t,Y_t)-\theta_t^f(\bar{Y}_\cdot))^2]$ is close to the idealized error $\mathbb{E}[(f(X_t,Y_t)-\theta_t^f(Y_\cdot))^2]$.

Moreover, even when Y and $\bar{Y}$ actually coincide, one is never able to obtain and exploit a continuous stream of data as modelled by the continuous path $Y_\cdot(\omega)$. Instead the observation arrives and is processed at discrete moments in time

$$0 = t_0 < t_1 < t_2 < \cdots < t_n = t.$$

However the continuous path $\hat{Y}_\cdot(\omega)$ obtained by linear interpolation of the discrete observations $(Y_{t_i}(\omega))_{i=1}^n$ is close to $Y_\cdot(\omega)$ (with respect to the supremum norm on $C([0,t],\mathbb{R}^{d_Y})$; hence, by the same argument, $\theta_t^f(\hat{Y}_\cdot)$ will be a sensible approximation to $\pi_t(f)$. In the *uncorrelated framework*, that is the case where the Brownian motions driving signal and observation are uncorrelated, continuity of the filter (in supremum norm) has first been established in [5] and technical details on measurability questions were provided in [6].

To our knowledge, the general correlated noise and multidimensional observation case has not been studied and it is the subject of the current work. In this case it turns out that we cannot hope to have robustness in the sense advocated by Clark. More precisely, there may not exists a continuous map $\theta_t^f : C([0,t],\mathbb{R}^{d_Y}) \to \mathbb{R}$, such that the representation (8.1) holds almost surely. The following is a simple example that illustrates this.

Example 8.1. Consider the filtering problem where the signal and the observation process solve the following pair of equations

$$X_t = X_0 + \int_0^t X_r d\left[Y_r^1 + Y_r^2\right] + \int_0^t X_r dr$$
$$Y_t = \int_0^t h(X_r)dr + W_t,$$

where Y is 2-dimensional and $\mathbb{P}(X_0 = 0) = \mathbb{P}(X_0 = 1) = \frac{1}{2}$. Then with f, h such that $f(0) = h_1(0) = h_2(0) = 0$ one can explicitly compute

$$\mathbb{E}[f(X_t)|\mathcal{Y}_t] = \frac{f(\exp(Y_t^1 + Y_t^2))}{1 + \exp\Big(-\sum_{k=1,2}\int_0^t h^k(\exp(Y_r^1 + Y_r^2))dY_r^k + \frac{1}{2}\int_0^t ||h(\exp(Y_r^1 + Y_r^2))||^2 dr\Big)}. \tag{8.2}$$

Following the findings of rough path theory (see, eg, [23, 27–29]) the expression on the right hand side of (8.2) is not continuous in supremum norm (nor in any other metric on path space) because of the stochastic integral. Explicitly, this follows, for example, from Theorem 1.1.1 in [29] by rewriting the exponential term as the solution to a stochastic differential equation driven by Y.

As detailed in Sect. 8.2, continuity is established when considering the observation as a *rough path*. This amounts to taking the rough path lift $\mathbf{Y}$ of Y as input to the problem. In addition to the path Y itself, $\mathbf{Y}$ consists of its iterated (Stratonovich) integrals $\int Y^i \circ dY^j$. Distance on the space of rough paths is measured using a Hölder metric, that also takes into account the iterated integrals. Our main result can then be stated as follows.

Theorem. *Under appropriate assumptions on the vector fields defining the diffusion processes X and Y, there exists a locally Lipschitz continuous map θ on the space of rough paths, such that $\theta(\mathbf{Y})$ is a version of the conditional expectation $\pi_t(f)$.*

We prove this result in Sect. 8.2.1 by directly working with the conditional expectation (8.1). One can also characterize the measure valued process π in terms of a stochastic partial differential equation, the Zakai equation. Showing continuity in the noise for this type of equations hence also entails robustness of the filter. The convergence of Wong-Zakai type approximations (in probability) has already been shown in [24, 25]. The true (rough) pathwise formulation and continuity of such measure valued rough differential equations will be laid out in our forthcoming work [15].

Regarding the second part of this paper, consider now, given an observation $\hat{X}$ of the SDE

$$dX = A \cdot h(X)dt + \Sigma(X)dW,$$

the maximum likelihood estimator for $A \in L(\mathbb{R}^d, \mathbb{R}^d)$, that is the maximizer of the likelihood $\mathbb{P}_A[\hat{X}]$. For X of dimension one, the estimator depends continuously on the path of a realization of X. To wit, in the case $h(x) = x$, $\Sigma(x) \equiv \sigma > 0$,

$$\hat{A}_T(X) = \frac{X_T^2 - x_0^2 - \sigma^2 T}{2\int_0^T X_s^2 ds},$$

which is, away from the singularity at the zero path $X_t \equiv 0$, immediately seen to be continuous in X. Note that the estimator is also continuous in σ, even though pathspace measure associated to different values of σ are actually mutually singular. Now, the numerator here is (two times) the stochastic Stratonovich integral $\int_0^T X_s \circ dX_s$. For X a multidimensional diffusion, continuous dependence of the estimator on the path itself does not hold anymore.

Example 8.2. For X of dimension 2, $h(x) = x$, it is a straightforward calculation (see for example [14]) that the $(1,2)$ component of the estimator $\hat{A}_T$ is given by

$$\hat{A}_T^{1,2} = \frac{\int_0^T X_r^{(2)} X_r^{(2)} dr \int_0^T X_r^{(1)} dX_r^{(1)} - \int_0^T X_r^{(1)} X_r^{(2)} dr \int_0^T X_r^{(2)} dX_r^{(1)}}{\int_0^T X_r^{(1)} X_r^{(1)} dr \int_0^T X_r^{(2)} X_r^{(2)} dr - (\int_0^T X_r^{(1)} X_r^{(2)} dr)^2 dr}.$$

As in the previous example, because of the stochastic integral $\int_0^T X_r^{(2)} dX_r^{(1)}$, this expression is not continuous on pathspace.

Using rough path metrics, continuity can be established though:

Theorem. *Under appropriate assumptions on the vector fields h and Σ, there exists a continuous map $\hat{\mathbf{A}}_T$ on a subset of the space of rough paths, such that $\hat{\mathbf{A}}(\mathbf{X})$ is a version of the maximum likelihood estimator $\hat{A}$.*

The following is the outline of the paper: In Sect. 8.2 we give our results on the filtering problem. In Sect. 8.2.1 we introduce the notion of a stochastic differential equation with rough drift, which is necessary for our main result there. We present it separately of the filtering problem, since this notion is of independent interest. Section 8.2.2 contains the main result of this section. In Sect. 8.3 we report on the problem of robustness of the maximum likelihood estimation for SDEs. Finally, in Sect. 8.4 we give some concluding remarks on the practical implications of our results.

8.2 Robustness of the Stochastic Filter

In the following, we will assume that the pair of processes (X, Y) satisfy the equation

$$dX_t = l_0(X_t, Y_t)dt + \sum_k Z_k(X_t, Y_t)dW_t^k + \sum_j L_j(X_t, Y_t)dB_t^j, \tag{8.3}$$

$$dY_t = h(X_t, Y_t)dt + dW_t, \tag{8.4}$$

with X_0 being a bounded random variable and $Y_0 = 0$. In (8.3) and (8.4), the process X is the d_X-dimensional signal, Y is the d_Y-dimensional observation, B and W are independent d_B-dimensional, respectively, d_Y-dimensional Brownian motions

independent of X_0. Suitable assumptions on the coefficients will be introduced later on. This framework covers a wide variety of applications of stochastic filtering (see, for example, [8] and the references therein) and has the added advantage that, within it, $\pi_t(f)$ admits an alternative representation that is crucial for the construction of its robust version. Let us detail this representation first.

Let $u = \{u_t, t > 0\}$ be the process defined by

$$u_t = \exp\left[-\sum_{i=1}^{d_Y} \left(\int_0^t h^i(X_s, Y_s) dW_s^i - \frac{1}{2} \int_0^t (h^i(X_s, Y_s))^2 ds \right) \right]. \tag{8.5}$$

Then, under suitable assumptions,[2] u is a martingale which is used to construct the probability measure $\mathbb{P}_0$ equivalent to $\mathbb{P}$ on $\mathscr{F}_t$ whose Radon–Nikodym derivative with respect to $\mathbb{P}$ is given by u, viz

$$\left. \frac{\mathrm{d}\mathbb{P}_0}{\mathrm{d}\mathbb{P}} \right|_{\mathscr{F}_t} = u_t.$$

Under $\mathbb{P}_0$, Y is a Brownian motion independent of B. Moreover the equation for the signal process X becomes

$$dX_t = \bar{l}_0(X_t, Y_t)dt + \sum_k Z_k(X_t, Y_t)dY_t^k + \sum_j L_j(X_t, Y_t)dB_t^j. \tag{8.6}$$

Observe that Eq. (8.6) is now written in terms of the pair of Brownian motions (Y, B) and the coefficient $\bar{l}_0$ is given by $\bar{l}_0 = l_0 + \sum_k Z_k h_k$. Moreover, for any measurable, bounded function $f : \mathbb{R}^{d_X + d_Y} \to \mathbb{R}$, we have the following formula called the Kallianpur-Striebel formula,

$$\pi_t(f) = \frac{p_t(f)}{p_t(1)}, \quad p_t(f) := \mathbb{E}_0[f(X_t, Y_t)v_t | \mathscr{Y}_t] \tag{8.7}$$

where $v = \{v_t, t > 0\}$ is the process defined as $v_t := \exp(I_t)$, $t \geq 0$ and

$$I_t := \sum_{i=1}^{d_Y} \left(\int_0^t h^i(X_r, Y_r) dY_r^i - \frac{1}{2} \int_0^t (h^i(X_r, Y_r))^2 dr \right), \quad t \geq 0. \tag{8.8}$$

The representation (8.7) suggests the following three-step methodology to construct a robust representation formula for π_t^f:

[2] For example, if Novikov's condition is satisfied, that is, if $\mathbb{E}\left[\exp\left(\frac{1}{2}\int_0^t \|h^i(X_s, Y_s)\|^2 ds\right)\right] < \infty$ for all $t > 0$, then u is a martingale. In particular it will be satisfied in our setting, in which h is bounded.

Step 1 We construct the triplet of processes (X^y, Y^y, I^y) [3] corresponding to the pair (y, B) where y is now a *fixed* observation path $y_\cdot = \{y_s, s \in [0,t]\}$ belonging to a suitable class of continuous functions and prove that the random variable $f(X^y, Y^y)\exp(I^y)$ is $\mathbb{P}_0$-integrable.

Step 2 We prove that the function $y_\cdot \to g_t^f(y_\cdot)$ defined as

$$g_t^f(y_\cdot) = \mathbb{E}_0\left[f(X_t^y, Y_t^y)\exp(I_t^y)\right] \tag{8.9}$$

is continuous.

Step 3 We prove that $g_t^f(Y_\cdot)$ is a version of $p_t(f)$. Then, following (8.7), the function, $y_\cdot \to \theta_t^f(y_\cdot)$ defined as

$$\theta_t^f = \frac{g_t^f}{g_t^1} \tag{8.10}$$

provides the robust version of $\pi_t(f)$.

We emphasize that **Step 3** cannot be omitted from the methodology. Indeed one has to prove that $g_t^f(Y_\cdot)$ is a version of $p_t(f)$ as this fact is not immediate from the definition of g_t^f.

Step 1 is immediate in the particular case when only the Brownian motion B drives X (i.e. the coefficient $Z = 0$) and X is itself a diffusion, i.e., it satisfies an equation of the form

$$dX_t = l_0(X_t)dt + \sum_j L_j(X_t)dB_t^j, \tag{8.11}$$

and h does only depend on X. In this case the process (X^y, Y^y) can be taken to be the pair (X, y). Moreover, we can define I^y by the formula

$$I_t^y := \sum_{i=1}^{d_Y}\left(h^i(X_t)y_t^i - \int_0^t y_r^i dh^i(X_r) - \frac{1}{2}\int_0^t (h^i(X_r, Y_r))^2 dr\right), \quad t \geq 0. \tag{8.12}$$

provided the processes $h^i(X)$ are semi-martingales. In (8.12), the integral $\int_0^t y_r^i dh^i(X_r)$ is the Itô integral of the non-random process y^i with respect to $h^i(X)$. Note that the formula for I_t^y is obtained by applying integration by parts to the stochastic integral in (8.8)

$$\int_0^t h^i(X_r)dY_r^i = h^i(X_t)Y_r^i - \int_0^t Y_r^i dh^i(X_r), \tag{8.13}$$

[3] As we shall see momentarily, in the uncorrelated case the choice of Y^y will trivially be y. In the correlated case we make it part of the *SDE with rough drift*, for (notational) convenience.

and replacing the process Y by the fixed path y in (8.13). This approach has been successfully used to study the robustness property for the filtering problem for the above case in a number of papers [5,6,26].

The construction of the process (X^y, Y^y, I^y) is no longer immediate in the case when $Z \neq 0$, i.e. when the signal is driven by both B and W (the correlated noise case). In the case when the observation is one dimensional one can solve this problem by using a method akin with Doss-Sussmann's "pathwise solution" of a stochastic differential equation (see [17, 30]). This approach has been employed by Davis to extend the robustness result to the correlated noise case with scalar observation (see, [10–13]). In this case one constructs first a diffeomorphism which is a pathwise solution of the equation[4]

$$\phi(t,x) = x + \int_0^t Z(\phi(s,x)) \circ dY_t. \tag{8.14}$$

The diffeomorphism is used to express the solution X of Eq. (8.6) as a composition between the diffeomorphism ϕ and the solution of a stochastic differential equation driven by B only and whose coefficients depend continuously on Y. As a result, we can make sense of X^y. I^y is then defined by a suitable (formal) integration by parts that produces a pathwise interpretation of the stochastic integral appearing in (8.8) and Y^y is chosen to be y, as before. The robust representation formula is then introduced as per (8.10). Additional results for the correlated noise case with scalar observation can be found in [19]. The extension of the robustness result to special cases of the correlated noise and multidimensional observation has been tackled in several works. Robustness results in the correlated setting have been obtained by Davis in [10,11] and Elliott and Kohlmann in [18], under a commutativity condition on the signal vector fields. Florchinger and Nappo [20] do not have correlated noise, but allow the coefficients to depend on the signal and the observation To sum up, all previous works on the robust representation problem either treat the uncorrelated case, the case with one-dimensional observation or the case where the Lie brackets of the relevant vector fields vanish. In parallel, Bagchi and Karandikar treat in [1] a different model with "finitely additive" state white noise and "finitely additive" observation noise. Robustness there is virtually built into the problem.

Nevertheless, we can show that a variation of the robustness representation formula still exists in this case. For this we need to "enhance" the original process Y by adding a second component to it which consists of its iterated integrals (that, knowing the path, is in a one-to-one correspondence with the Levỳ area process). Explicitly we consider the process $\mathbf{Y} = \{\mathbf{Y}_t, t \geq 0\}$ defined as

$$\mathbf{Y}_t = \left(Y_t, \begin{pmatrix} \int_0^t Y_r^1 \circ dY_r^1 & \cdots & \int_0^t Y_r^1 \circ dY_r^{d^Y} \\ \cdots & \cdots & \cdots \\ \int_0^t Y_r^{d^Y} \circ dY_r^1 & \cdots & \int_0^t Y_r^{d^Y} \circ dY_r^{d^Y} \end{pmatrix} \right), \quad t \geq 0. \tag{8.15}$$

[4]Here $d^Y = 1$ and Y is scalar.

The stochastic integrals in (8.15) are Stratonovich integrals. The state space of $\mathbf{Y}$ is $G^2(\mathbb{R}^{d_Y}) \cong \mathbb{R}^{d_Y} \oplus \mathrm{so}(d_Y)$, where $\mathrm{so}(d_Y)$ is the set of anti-symmetric matrices of dimension d_Y.[5] Over this state space we consider not the space of continuous function, but a subspace $\mathscr{C}^{0,\alpha}$ that contains paths $\eta : [0,t] \to G^2(\mathbb{R}^{d_Y})$ that are α-Hölder in the $\mathbb{R}^{d_Y}$-component and somewhat "2α-Hölder" in the $\mathrm{so}(d_Y)$-component, where α is a suitably chosen constant $\alpha < 1/2$. Note that there exists a modification of $\mathbf{Y}$ such that $\mathbf{Y}(\omega) \in \mathscr{C}^{0,\alpha}$ for all ω (Corollary 13.14 in [23]).

The space $\mathscr{C}^{0,\alpha}$ is endowed with the α-Hölder rough path metric under which $\mathscr{C}^{0,\alpha}$ becomes a complete metric space. The main result of the paper (captured in Theorem 8.2) is that there exists a continuous map $\theta_t^f : \mathscr{C}^{0,\alpha} \to \mathbb{R}$, such that

$$\pi_t(f) = \theta_t^f(\mathbf{Y}.) \qquad \mathbb{P}\text{-a.s.} \tag{8.16}$$

Even though the map is defined on a slightly more abstract space, it nonetheless enjoys the desirable properties described above for the case of a continuous version on $C([0,t],\mathbb{R}^d)$. Since $\mathbb{P} \circ \mathbf{Y}^{-1}$ positively charges all non-empty open sets of $\mathscr{C}^{0,\alpha}$,[6] the continuous version we construct will be unique. Also, it provides a certain model robustness, in the sense that $\mathbb{E}[(f(X_t) - \theta_t^f(\bar{\mathbf{Y}}.))^2]$ is well approximated by the idealized error $\mathbb{E}[(f(X_t) - \theta_t^f(\mathbf{Y}.))^2]$, if $\bar{\mathbf{Y}}.$ is close in distribution to $\mathbf{Y}.$. The problem of discrete observation is a little more delicate. One one hand, it is true that the rough path lift $\hat{\mathbf{Y}}$ calculated from the linearly interpolated Brownian motion $\hat{Y}$ will converge to the true rough path $\mathbf{Y}$ in probability as the mesh goes to zero (Corollary 13.21 in [23]), which implies that $\theta_t^f(\hat{\mathbf{Y}})$ is close in probability to $\theta_t^f(\mathbf{Y})$ (we provide local Lipschitz estimates for θ^f). Actually, most sensible approximations will do, as is for example shown in Chapter 13 in [23] (although, contrary to the uncorrelated case, not all interpolations that converge in uniform topology will work, see e.g. Theorem 13.24 ibid.). But these are probabilistic statements, that somehow miss the pathwise robustness that one wants to provide with θ_t^f. If, on the other hand, one is able to observe at discrete time points not only the process itself, but also its second level, i.e. the area, one can construct an interpolating rough path using geodesics (see e.g. Chapter 13.3.1 in [23]) which is close to the true (lifted) observation path $\mathbf{Y}$ in the relevant metric *for all realizations* $\mathbf{Y} \in \mathscr{C}^{0,\alpha}$.

Nomenclature Lip^γ is the set of γ-Lipschitz [7] functions $a : \mathbb{R}^m \to \mathbb{R}^n$ where m and n are chosen according to the context.

[5] More generally, $G^{[1/\alpha]}(\mathbb{R}^d)$ is the "correct" state space for a geometric α-Hölder rough path; the space of such paths subject to α-Hölder regularity (in rough path sense) yields a complete metric space under α-Hölder rough path metric. Technical details of geometric rough path spaces (as found e.g. in section 9 of [23]) are not required for understanding the results of the present paper.

[6] This fact is a consequence of the support theorem of Brownian motion in Hölder rough path topology [21], see also Chapter 13 in [23].

[7] In the sense of E. Stein, i.e. bounded k-th derivative for $k = 0, \ldots, \lfloor \gamma \rfloor$ and $\gamma - \lfloor \gamma \rfloor$-Hölder continuous $\lfloor \gamma \rfloor$-th derivative.

$G^2(\mathbb{R}^{d_Y}) \cong \mathbb{R}^d \oplus \mathfrak{so}(d_Y)$ is the state space for a d_Y-dimensional Brownian motion (or, in general for an arbitrary semi-martingale) and its corresponding Lévy area.

$\mathscr{C}^{0,\alpha} := C_0^{0,\alpha-\text{Höl}}([0,t], G^2(\mathbb{R}^{d_Y}))$ is the set of geometric α-Hölder rough paths $\eta : [0,t] \to G^2(\mathbb{R}^{d_Y})$ starting at 0. We shall use the non-homogenous metric $\rho_{\alpha-\text{Höl}}$ on this space, with associated "norm" $||.||_{\alpha-\text{Höl}}$.

In the following we will make use of an auxiliary filtered probability space $(\bar{\Omega}, \bar{\mathscr{F}}, (\bar{F}_t)_{t\geq 0}, \bar{\mathbb{P}})$ carrying a d_B-dimensional Brownian motion $\bar{B}$.[8]

Let $\mathscr{S}^0 = \mathscr{S}^0(\bar{\Omega})$ denote the space of adapted, continuous processes in $\mathbb{R}^{d_S}$, with the topology of uniform convergence in probability.

For $q \geq 1$ we denote by $\mathscr{S}^q = \mathscr{S}^q(\bar{\Omega})$ the space of processes $X \in \mathscr{S}^0$ such that

$$||X||_{\mathscr{S}^q} := \left(\bar{\mathbb{E}}[\sup_{s\leq t} |X_t|^q]\right)^{1/q} < \infty.$$

8.2.1 SDE with Rough Drift

For the statement and proof of the main result we shall use the notion (and the properties) of an *SDE with rough drift* captured in the following theorem.

As defined above, let $(\bar{\Omega}, \bar{\mathscr{F}}, (\bar{F}_t)_{t\geq 0}, \bar{\mathbb{P}})$ be a filtered probability space carrying a d_B-dimensional Brownian motion $\bar{B}$ and a bounded d_S-dimensional random vector S_0 independent of $\bar{B}$. We recall that $(\Omega, \mathscr{F}, \mathbb{P}_0)$ carries, as above, the d_Y-dimensional Brownian motion Y and let $\hat{\Omega} = \Omega \times \bar{\Omega}$ be the product space, with product measure $\hat{\mathbb{P}} := \mathbb{P}_0 \otimes \bar{\mathbb{P}}$. Let S be the unique solution on this probability space to the SDE

$$S_t = S_0 + \int_0^t a(S_r)dr + \int_0^t b(S_r) \circ d\bar{B}_r + \int_0^t c(S_r) \circ dY_r. \tag{8.17}$$

Denote by $\mathbf{Y}$ the rough path lift of Y (i.e. the enhanced Brownian Motion over Y). In the following, we fix $\varepsilon \in (0,1)$ and $\alpha \in (\frac{1}{2+\varepsilon}, \frac{1}{2})$.

We will use one of the following assumptions.

(a1) $a, b \in Lip^1$ and $c \in Lip^{4+\varepsilon}$
(a1') $a, b \in Lip^1$ and $c \in Lip^{5+\varepsilon}$
(a2) $a \in Lip^{1+\varepsilon}$ and $b, c \in Lip^{2+\varepsilon}$

[8] We introduce this auxiliary probability space, since in the proof of Theorem 8.2 it will be easier to work on a product space separating the randomness coming from Y and B. A similar approach was followed in the proof of Theorem 1 in [2].

Theorem 8.1. *Under assumption* **(a1), (a1′)** *or* **(a2)**, *for every* $\boldsymbol{\eta} \in \mathscr{C}^{0,\alpha}$, *there exists a* d_S*-dimensional process* $S^{\boldsymbol{\eta}} \in \mathscr{S}^0$ *such that*

- *It possesses exponential integrability: for every* $R > 0, q \geq 1$

$$\sup_{||\boldsymbol{\eta}||_{\alpha-H\ddot{o}l} < R} \mathbb{E}[\exp(q|S^{\boldsymbol{\eta}}|_{\infty;[0,t]})] < \infty. \tag{8.18}$$

- *The mapping* $\boldsymbol{\eta} \mapsto S^{\boldsymbol{\eta}}$, *from* $\mathscr{C}^{0,\alpha}$ *to* $\mathscr{S}^q$, *is locally uniformly continuous for all* $q \geq 1$*; under assumption* **(a1′)** *or* **(a2)** *it is even locally Lipschitz.*
- $S^{\boldsymbol{\eta}}$ *is consistent with the Stratonovich SDE solution: for* $\mathbb{P}_0 - a.e.\ \omega$

$$\bar{\mathbb{P}}[S_s(\omega, \cdot) = S^{\mathbf{Y}}(\omega)_s(\cdot), \quad s \leq t] = 1. \tag{8.19}$$

We write that formally $S^{\boldsymbol{\eta}}$ *solves the following* stochastic differential equation with rough drift

$$S_t^{\boldsymbol{\eta}} = S_0^{\boldsymbol{\eta}} + \int_0^t a(S_r^{\boldsymbol{\eta}})dr + \int_0^t b(S_r^{\boldsymbol{\eta}})d\bar{B}_r + \int_0^t c(S_r^{\boldsymbol{\eta}})d\boldsymbol{\eta}_r. \tag{8.20}$$

The preceding theorem was proven in [7] under assumptions **(a1)** and **(a1′)**, using a flow decomposition. Although (8.20) does at first sight not have a rigorous meaning, if one (formally) computes the equation for $\tilde{S}_t^{\boldsymbol{\eta}} := \phi^{-1}(t, S_t^{\boldsymbol{\eta}})$, where $\phi = \phi^{\boldsymbol{\eta}}$ is the flow corresponding to $dx = c(x)d\boldsymbol{\eta}$, one is left with a classical SDE (the $d\boldsymbol{\eta}$ term disappears). In the equation for $\tilde{S}$ the dependence on the rough path is harmless and only comes in through the flow and its space-derivatives (but not its time-derivative!). Hence $\tilde{S}^{\boldsymbol{\eta}}$ can be solved classically and we *define* now $S_t^{\boldsymbol{\eta}} := \phi(t, \tilde{S}_t^{\boldsymbol{\eta}})$. The claimed continuity and integrability properties then follow from estimates on the rough flow ϕ and classical robustness and integrability properties of SDEs. It is well-known in rough path theory (for example Chapter 13 in [23]), that $\phi^{\boldsymbol{\eta}}$ evaluated at the lifted Brownian motion $\mathbf{Y}$ coincides with the Stratonovich flow to $dX = c(X) \circ dY$. It then is easy to show property (8.19).

In [16] under assumption **(a2)** another approach was chosen. We realized that there is a canonical way of defining a joint rough path lift Λ of $\boldsymbol{\eta}$ and $\bar{B}$. In fact the only term that is not immediately well-defined is the cross-term $\int \bar{B}d\boldsymbol{\eta}$. But using formal integration by parts one can rewrite as $B\boldsymbol{\eta} - \int \boldsymbol{\eta} d\bar{B}$, which is well-defined via Itô-integration. It is then straightforward to check that this indeed defines a rough path (almost surely, with the null-set depending on $\boldsymbol{\eta}$). To show (8.18) we modify recent integrability results for Gaussian rough paths [4, 22] using the Borell inequality on Gaussian spaces.

8.2.2 Assumptions and Main Result

In the following we will make use of the Stratonovich version of Eq. (8.6)

$$\begin{aligned} X_t &= X_0 + \int_0^t L_0(X_r, Y_r)dr + \sum_k \int_0^t Z_k(X_r, Y_r) \circ dY_r^k + \sum_j \int_0^t L_j(X_r, Y_r) \circ dB_r^j, \\ Y_t &= \int_0^t h(X_r, Y_r)dr + W_t. \end{aligned} \tag{8.21}$$

where

$$L_0^j(x,y) = \bar{l}_0^j(x,y) - \frac{1}{2}\partial_{x_i} Z_k^j(x,y) Z_k^i(x,y) - \frac{1}{2}\partial_{y_k} Z_k^j(x,y) - \frac{1}{2}\partial_{x_i} L_k^j(x,y) L_k^i(x,y).$$

We remind that under $\mathbb{P}_0$ the observation Y is a Brownian motion independent of B.

We will assume that f is a bounded Lipschitz function and we fix $\varepsilon \in (0,1)$ $\alpha \in (\frac{1}{2+\varepsilon}, \frac{1}{2})$, $t > 0$ and X_0 is a bounded random vector independent of B and Y. We will use one of the following assumptions.

(A1) $Z_1, \dots, Z_{d_Y} \in \mathrm{Lip}^{4+\varepsilon}$, $h^1, \dots, h^{d_Y} \in \mathrm{Lip}^{4+\varepsilon}$ and $L_0, L_1, \dots, L_{d_B} \in \mathrm{Lip}^1$
(A1′) $Z_1, \dots, Z_{d_Y} \in \mathrm{Lip}^{5+\varepsilon}$, $h^1, \dots, h^{d_Y} \in \mathrm{Lip}^{5+\varepsilon}$ and $L_0, L_1, \dots, L_{d_B} \in \mathrm{Lip}^1$
(A2) $Z_1, \dots, Z_{d_Y} \in \mathrm{Lip}^{2+\varepsilon}$, $h^1, \dots, h^{d_Y} \in \mathrm{Lip}^{2+\varepsilon}$ and $L_0, L_1, \dots, L_{d_B} \in \mathrm{Lip}^{1+\varepsilon}$

Assume **(A1)**, **(A1′)** or **(A2)**. For $\boldsymbol{\eta} \in \mathscr{C}^{0,\alpha}$ there exists, by Theorem 8.1, a solution $(X^{\boldsymbol{\eta}}, I^{\boldsymbol{\eta}})$ to the following *SDE with rough drift*

$$\begin{aligned} X_t^{\boldsymbol{\eta}} &= X_0 + \int_0^t L_0(X_r^{\boldsymbol{\eta}}, Y_r^{\boldsymbol{\eta}})dr + \int_0^t Z(X_r^{\boldsymbol{\eta}}, Y_r^{\boldsymbol{\eta}}) d\boldsymbol{\eta}_r + \sum_j \int_0^t L_j(X_r^{\boldsymbol{\eta}}, Y_r^{\boldsymbol{\eta}}) d\bar{B}_r^j, \\ Y_t^{\boldsymbol{\eta}} &= \int_0^t d\boldsymbol{\eta}_r, \\ I_t^{\boldsymbol{\eta}} &= \int_0^t h(X_r^{\boldsymbol{\eta}}, Y_r^{\boldsymbol{\eta}}) d\boldsymbol{\eta}_r - \frac{1}{2}\sum_k \int_0^t D_k h^k(X_r^{\boldsymbol{\eta}}, Y_r^{\boldsymbol{\eta}})dr. \end{aligned} \tag{8.22}$$

Remark 8.1. Note that formally (!) when replacing the rough path $\boldsymbol{\eta}$ with the process Y, $X^{\boldsymbol{\eta}}, Y^{\boldsymbol{\eta}}$ yields the solution to the SDE (8.21) and $\exp(I_t^{\boldsymbol{\eta}})$ yields the (Girsanov) multiplicator in (8.7). This observation is made precise in the statement of Theorem 8.1.

We introduce the functions $g^f, g^1, \theta : \mathscr{C}^{0,\alpha} \to \mathbb{R}$ defined as

$$g^f(\boldsymbol{\eta}) := \bar{\mathbb{E}}[f(X_t^{\boldsymbol{\eta}}, Y_t^{\boldsymbol{\eta}}) \exp(I_t^{\boldsymbol{\eta}})], \quad g^1(\boldsymbol{\eta}) := \bar{\mathbb{E}}[\exp(I_t^{\boldsymbol{\eta}})], \quad \theta(\boldsymbol{\eta}) := \frac{g^f(\boldsymbol{\eta})}{g^1(\boldsymbol{\eta})}.$$

Denote by $\mathbf{Y}.$, as before, the canonical rough path lift of Y to $\mathscr{C}^{0,\alpha}$.

Theorem 8.2. *1. Assume that* **(A1)** *holds then θ is locally uniformly continuous function of the "observation" rough path $\boldsymbol{\eta}$. If* **(A1′)** *or* **(A2)** *holds, then θ is locally Lipschitz.*
2. Under any of the conditions **(A1′)**, **(A1′)** *or* **(A2)** *we have $\theta(\mathbf{Y}.) = \pi_t(f)$, $\mathbb{P}-a.s.$*

Proof. **(i)** From Theorem 8.1 we know that for $\boldsymbol{\eta} \in \mathscr{C}^{0,\alpha}$ the SDE with rough drift (8.22) has a unique solution $(X^{\boldsymbol{\eta}}, Y^{\boldsymbol{\eta}}, I^{\boldsymbol{\eta}})$ belonging to $\mathscr{S}^2$.

Let now $\boldsymbol{\eta}, \boldsymbol{\eta}' \in \mathscr{C}^{0,\alpha}$. Denote $X = X^{\boldsymbol{\eta}}, Y = Y^{\boldsymbol{\eta}}, I = I^{\boldsymbol{\eta}}$ and analogously for $\boldsymbol{\eta}'$.

Then

$$\begin{aligned}
&|g^f(\boldsymbol{\eta}) - g^f(\boldsymbol{\eta})| \\
&\le \mathbb{E}[|f(X_t, Y_t)\exp(I_t) - f(X_t', Y"_t)\exp(I_t')|] \\
&\le \mathbb{E}[|f(X_t, Y_t)||\exp(I_t) - \exp(I_t')|] + \mathbb{E}[|f(X_t, Y_t) - f(X_t', Y_t')|\exp(I_t')] \\
&\le |f|_\infty \mathbb{E}[|\exp(I_t) - \exp(I_t')|] + \mathbb{E}[|f(X_t, Y_t) - f(X_t', Y_t')|^2]^{1/2}\mathbb{E}[|\exp(I_t')|^2]^{1/2} \\
&\le |f|_\infty \mathbb{E}[|\exp(I_t) + \exp(I_t')|^2]^{1/2}\mathbb{E}[|I_t - I_t'|]^{1/2} \\
&\qquad + \mathbb{E}[|f(X_t, Y_t) - f(X_t', Y_t')|^2]^{1/2}\mathbb{E}[|\exp(I_t')|^2]^{1/2}
\end{aligned}$$

Hence, using from Theorem 8.1 the continuity statement as well as the boundedness of exponential moments, we see that g^f is locally uniformly continuous under (A1) and it is locally Lipschitz under (A1') or (A2).

The same then holds true for g^1 and moreover $g^1(\boldsymbol{\eta}) > 0$. Hence θ is locally uniformly continuous under (A1) and locally Lipschitz under (A1') or (A2).

(ii) To prove the statement it is enough to show that

$$g^f(\mathbf{Y}.) = p_t(f) \qquad \mathbb{P}-a.s.$$

which is equivalent to

$$g^f(\mathbf{Y}.) = p_t(f) \qquad \mathbb{P}_0-a.s.$$

For that, it suffices to show that

$$\mathbb{E}_0[p_t(f)\Upsilon(Y.)] = \mathbb{E}_0[g^f(\mathbf{Y}.)\Upsilon(Y.)], \tag{8.23}$$

for an arbitrary continuous bounded function $\Upsilon : C([0,t], \mathbb{R}^{d_Y}) \to \mathbb{R}$.

Let $(\bar{\Omega}, \bar{\mathscr{F}}, \bar{\mathbb{P}})$ be the auxiliary probability space from before, carrying an d_B-dimensional Brownian motion $\bar{B}$. Let $(\hat{\Omega}, \hat{\mathscr{F}}, \hat{\mathbb{P}}) := (\Omega \times \bar{\Omega}, \mathscr{F} \otimes \bar{\mathscr{F}}, \mathbb{P}_0 \otimes \bar{\mathbb{P}})$. By Y and X_0 we denote also the 'lift' of Y to $\hat{\Omega}$, i.e. $Y(\omega, \bar{\omega}) = Y(\omega)$, $X_0(\omega, \bar{\omega}) = X_0(\omega)$. Then (Y, B) (on Ω under $\mathbb{P}_0$) has the same distribution as $(Y, \bar{B})$ (on $\hat{\Omega}$ under $\hat{\mathbb{P}}$).

Denote by $(\hat{X}, \hat{I})$ the solution on $(\hat{\Omega}, \hat{F}, \hat{\mathbb{P}})$ to the SDE

$$\hat{X}_t = X_0 + \int_0^t L_0(\hat{X}_r, Y_r)dr + \sum_k \int_0^t Z_k(\hat{X}_r, Y_r) \circ dY_r^k + \sum_j \int_0^t L_j(\hat{X}_r, Y_r) d\bar{B}_r^j,$$

$$\hat{I}_t = \sum_k \int_0^t h^k(\hat{X}_r, Y_r) \circ dY_r^k - \frac{1}{2} \sum_k \int_0^t D_k h^k(\hat{X}_r, Y_r)dr.$$

Then

$$(Y, \hat{X}, \hat{I})_{\hat{\mathbb{P}}} \sim \left(Y, X, \sum_k \int_0^{\cdot} h^k(X_r, Y_r) \circ dY_r^k - \frac{1}{2} \sum_k \int_0^{\cdot} D_k h^k(X_r, Y_r)dr \right)_{\mathbb{P}_0}.$$

Hence, for the left hand side of (8.23),

$$\begin{aligned}
&\mathbb{E}_0[p_t(f)\Upsilon(Y.)] \\
&\quad = \mathbb{E}_0[f(X_t, Y_t) \exp\left(\sum_k \int_0^t h^k(X_r, Y_r) \circ dY_r^k - \frac{1}{2} \sum_k \int_0^t D_k h^k(X_r, Y_r)dr \right) \Upsilon(Y.)] \\
&\quad = \hat{\mathbb{E}}[f(\hat{X}_t, Y_t) \exp\left(\hat{I}_t \right) \Upsilon(Y.)]
\end{aligned}$$

On the other hand, from Theorem 8.1 we know that for $\mathbb{P}_0 - a.e.\ \omega$, for $\bar{\mathbb{P}} - a.e.\ \bar{\omega}$

$$X^{\mathbf{Y}.(\omega)}(\bar{\omega})_t = \hat{X}_t(\omega, \bar{\omega}),\ Y^{\mathbf{Y}.(\omega)}(\bar{\omega})_t = \hat{Y}_t(\omega, \bar{\omega}),\ I^{\mathbf{Y}.(\omega)}(\bar{\omega})_t = \hat{I}_t(\omega, \bar{\omega}).$$

Hence, for the right hand side of (8.23) we get (using Fubini for the last equality)

$$\begin{aligned}
\mathbb{E}_0[g^f(\mathbf{Y}.)\Upsilon(Y.)] &= \mathbb{E}_0[\bar{\mathbb{E}}[f(X_t^{\mathbf{Y}.}, Y_t^{\mathbf{Y}.}) \exp\left(I_t^{\mathbf{Y}.} \right)]\Upsilon(Y.)] \\
&= \mathbb{E}_0[\bar{\mathbb{E}}[f(\hat{X}_t, Y_t) \exp\left(\hat{I}_t \right)]\Upsilon(Y.)] \\
&= \hat{\mathbb{E}}[f(\hat{X}_t, Y_t) \exp\left(\hat{I}_t \right) \Upsilon(Y.)],
\end{aligned}$$

which yields (8.23).

8.3 Maximum Likelihood Estimation for SDEs

Let W be a d-dimensional Wiener process and $A \in \mathbb{V} := L\left(\mathbb{R}^d, \mathbb{R}^d\right)$. Consider sufficiently regular $h : \mathbb{R}^d \to \mathbb{R}^d$ and $\Sigma : \mathbb{R}^d \to L\left(\mathbb{R}^d, \mathbb{R}^d\right)$ so that

$$dX_t = A\, h\left(X_t\right) dt + \Sigma\left(X_t\right) dW_t \tag{8.24}$$

has a unique solution, started from $X_0 = x_0$. The important example of multidimensional Ornstein-Uhlenbeck dynamics, for instance, falls in the class of diffusions considered here (take $h(x) = x, g = 0$ and constant, non-degenerate diffusion matrix Σ). We are interested in estimating the drift parameter A, given some observation sample path $\{X_t(\omega) = \omega_t : t \in [0,T]\}$. More precisely, we are looking for a Maximum Likelihood Estimator (MLE) of the form

$$\hat{A}_T = \hat{A}_T(\omega) = \hat{A}_T(X) \in \mathbb{V}$$

relative to the reference measure given by the law of X, viewed as measure on pathspace, in the case $A \equiv 0$.

Theorem 8.3. *(i) Define*

$$R_h := \left\{X \in C\left([0,T], \mathbb{R}^d\right) : \text{span}\left\{h\left(X_t\right) : t \in [0,T]\right\} = \mathbb{R}^d\right\}. \tag{8.25}$$

Assume that the set of critical points of h has no accumulation points (i.e. on every bounded set, there is only a finite set of points at which $\det Dh(x) = 0$). Then, for every fixed, non-degenerate volatility function Σ

$$\mathbb{P}^{0,\Sigma}\left(R_h\right) = 1.$$

Hence, $I_T = I_T(\omega)$ is $\mathbb{P}^{0,\Sigma}$-almost surely invertible so that $A_T(\omega) := I_T^{-1} S_T(\omega)$ is $\mathbb{P}^{0,\Sigma}$-almost surely well-defined.

(ii) Fix $\alpha \in (1/3, 1/2)$. Then, $\mathbb{P}^{0,\Sigma}$-almost surely, $X(\omega)$ lifts to a (random) geometric α-Hölder rough path, i.e. a random element in the rough path space $\mathscr{C}^{0,\alpha}$, via the (existing) limit in probability

$$\mathbf{X}(\omega) := \left(X(\omega), \mathbb{X}(\omega)\right) := \lim_n \left(X^n, \int X^n \otimes dX^n\right)$$

where X^n denotes dyadic piecewise linear approximations to X.

(iii) Define $\mathbb{D} \subset \mathscr{D}_g^\alpha\left([0,T], \mathbb{R}^d\right)$ by

$$\mathbb{D} = \left\{(X, \mathbb{X}) \in \mathscr{C}^{0,\alpha} : X \in R_h\right\}.$$

Then, under the assumption of (i), for every fixed, non-degenerate volatility function Σ,

$$\mathbb{P}^{0,\Sigma}\left(\mathbf{X}(\omega) \in \mathbb{D}\right) = 1.$$

(iv) There exists a deterministic, continuous [with respect to α-Hölder rough path metric] map

$$\hat{\mathbf{A}}_T : \begin{cases} \mathbb{D} & \to \mathbb{R}^{d\times d} \\ \mathbf{X} & \mapsto \hat{\mathbf{A}}_T(\mathbf{X}) \end{cases}$$

so that, for every fixed, non-degenerate volatility function Σ,

$$\mathbb{P}^{0,\Sigma}\left[\hat{\mathbf{A}}_T(\mathbf{X}(\omega)) = A_T(\omega)\right] = 1. \tag{8.26}$$

In fact, $\hat{\mathbf{A}}_T$ is explicitly given, for $(X,\mathbb{X}) \in \mathbb{D} \subset \mathscr{C}^{0,\alpha}$, by

$$\hat{\mathbf{A}}(X,\mathbb{X}) := \mathbf{I}_T^{-1}(X)\mathbf{S}_T(X,\mathbb{X}),$$

where

$$\begin{aligned} \mathbf{I}_T(X) &:= \int_0^T h(X_s) \otimes C^{-1}(X_s) \otimes h(X_s)\, ds, \\ \mathbf{S}_T(X,\mathbb{X})_{i,j} &:= \int_0^T h_i(X_s) \otimes C_{j\cdot}^{-1}(X_s) \circ dX_s \\ &\quad - \frac{1}{2}\int_0^T \mathrm{Tr}[D(h_i C_{j\cdot}^{-1})(X_s)\Sigma(X_s)\Sigma(X_s)^T]ds \end{aligned}$$

and the $\circ dX$ integral[9] *is understood as a (deterministic) rough integration against $\mathbf{X} = (X,\mathbb{X})$.*

(v) The map $\hat{\mathbf{A}}_T$ is also continuous with respect to the volatility specification. Indeed, fix $c > 0$ and set

$$\Xi := \left\{\Sigma \in \mathrm{Lip}^2 : c^{-1}I \le \Sigma\Sigma^T \le cI\right\}.$$

Then $\hat{\mathbf{A}}_T$ viewed as map from $\mathbb{D} \times \Xi \to \mathbb{R}^d$ is also continuous.

[9] ... often written as d$\mathbf{X}$ integral in the literature on rough integration ...

8.4 Practical Implications

Both the optimal filter for partially observed diffusion processes and the maximum likelihood estimator for the drift of an SDE are functionals on pathspace. In dimension strictly larger then one (and with non-zero correlation in the filtering problem) there are simple countexamples (Examples 8.1 and 8.2) that demonstrate that continuity of these functionals in the path itself (in supremum norm, say) is not possible. Our main results, Theorems 8.2 and 8.3, re-establish continuity by measuring distance with a rough path metric.

As already briefly touched upon, this requires the user to understand observations as a rough path. In our setting (Hölder regularity just below $\frac{1}{2}$) this amounts to recording the second order integrals of the stochastic process.

A possible approach that we are currently investigating, is observing the effect that the observation has on an (auxiliary) differential equations (which would model a physical measuring device). Under certain assumption on this differential equation, this forces the rough path to reveal its higher order elements (up to a certain order of error).

A different interpretation is the view that (discrete) observation data always gives rise to (piecewise) smooth paths by suitable interpolation (piecewise linear, say). However, with increasing frequency of observation the Lipschitz norm of these continuous paths will blow up. As a result one has no control over the modulus of continuity (of the filter resp. the MLE) with respect to the discrete data as the observation frequency becomes large. Our point of view here leads to a uniform (in the frequency of observation) modulus of continuity.

References

1. Bagchi, A., Karandikar, R.: White noise theory of robust nonlinear filtering with correlated state and observation noises. Syst. Control Lett. **23**(2), 137–148 (1994)
2. Bain, A., Crisan, D.: Fundamentals of Stochastic Filtering. Springer, New York/London (2008)
3. Breiman, L.: Probability. Classics in Applied Mathematics. SIAM, Philadelphia (1992)
4. Cass, T., Lyons, T.: Evolving communities with individual preferences (2013, preprint). arXiv:1303.4243
5. Clark, J.M.C.: The design of robust approximations to the stochastic differential equations of nonlinear filtering. In: Communication Systems and Random Process Theory: Proceedings of 2nd NATO Advanced Study Institute, Darlington, 1977, pp. 721–734. NATO Advanced Study Institute Series, Series E, Applied sciences, vol. 25. Sijthoff & Noordhoff, Alphen aan den Rijn (1978)
6. Clark, J.M.C., Crisan, D.: On a robust version of the integral representation formula of nonlinear filtering. Probab. Theory Relat. F. **133**(1), 43–56 (2005). doi:10.1007/s00440-004-0412-5
7. Crisan, D., Diehl, J., Friz, P., Oberhauser, H.: Robust filtering: multdimensional noise and multidimensional observation. Ann. Appl. Probab. **23**(5), 2139–2160 (2013)
8. Crisan, D., Rozovskiĭ, B.(eds.): The Oxford Handbook of Nonlinear Filtering, xiv, p. 1063. Oxford University Press, Oxford (2011)

9. Davis, M.: Pathwise nonlinear filtering. In: Stochastic Systems: The Mathematics of Filtering and Identification and Applications: Proceedings of the NATO Advanced Study Institute, Les Arcs, 1980, pp. 505–528 (1981)
10. Davis, M.: Pathwise nonlinear filtering with correlated noise. In: Crisan, D., Rozovskiĭ, B.L. (eds.) The Oxford Handbook of Nonlinear Filtering, pp. 403–424. Oxford University Press, Oxford (2011)
11. Davis, M.H.A.: On a multiplicative functional transformation arising in nonlinear filtering theory. Z. Wahrsch. Verw. Gebiete **54**(2), 125–139 (1980). doi:10.1007/BF00531444. http://dx.doi.org/10.1007/BF00531444
12. Davis, M.H.A.: A pathwise solution of the equations of nonlinear filtering. Teor. Veroyatnost. i Primenen. **27**(1), 160–167 (1982)
13. Davis, M.H.A., Spathopoulos, M.P.: Pathwise nonlinear filtering for nondegenerate diffusions with noise correlation. SIAM J. Control Optim. **25**(2), 260–278 (1987)
14. Diehl, J., Friz, P., Mai, H.: Pathwise stability of likelihood estimators for diffusions via rough paths (2013, preprint). arXiv:1311.1061
15. Diehl, J., Friz, P., Stannat, W.: Measure valued rough differential equations (2014, in preparation)
16. Diehl, J., Oberhauser, H., Riedel, S.: A Levy-area between Brownian motion and rough paths with applications to robust non-linear filtering and RPDEs (2013, preprint). arXiv:1301.3799
17. Doss, H.: Liens entre équations différentielles stochastiques et ordinaires. Ann. Inst. H. Poincaré Sect. B (N.S.) **13**(2), 99–125 (1977)
18. Elliott, R., Kohlmann, M.: Robust filtering for correlated multidimensional observations. Math. Z. **178**(4), 559–578 (1981)
19. Florchinger, P.: Zakai equation of nonlinear filtering with unbounded coefficients. the case of dependent noises. Syst. Control Lett.**21**(5), 413–422 (1993)
20. Florchinger, P., Nappo, G.: Continuity of the filter with unbounded observation coefficients. Stoch. Anal. Appl. **29**(4), 612–630 (2011)
21. Friz, P.: Continuity of the Ito-map for Hoelder rough paths with applications to the support theorem in Hoelder norm. In: Probability and Partial Differential Equations in Modern Applied Mathematics. IMA Volumes in Mathematics and its Applications, vol. 140, pp. 117–135. Springer, New York (2005)
22. Friz, P., Riedel, S.: Integrability of (non-) linear rough differential equations and integrals. Stoch. Anal. Appl. **31**(2), 336–358 (2013)
23. Friz, P.K., Victoir, N.B.: Multidimensional Stochastic Processes as Rough Paths: Theory and Applications. Cambridge Studies in Advanced Mathematics, vol. 120. Cambridge University Press, Cambridge (2010)
24. Gyöngy, I.: On the approximation of stochastic partial differential equations i. Stoch. Stoch. Rep. **25**(2), 59–85 (1988)
25. Gyöngy, I.: On the approximation of stochastic partial differential equations ii. Stoch. Stoch. Rep. **26**(3), 129–164 (1989)
26. Kushner, H.J.: A robust discrete state approximation to the optimal nonlinear filter for a diffusion. Stochastics **3**(2), 75–83 (1979)
27. Lyons, T.J.: Differential equations driven by rough signals. I. An extension of an inequality of L. C. Young. Math. Res. Lett. **1**(4), 451–464 (1994)
28. Lyons, T.J., Caruana, M., Lévy, T.: Differential Equations Driven by Rough Paths. Lecture Notes in Mathematics, vol. 1908. Springer, Berlin (2007). Lectures from the 34th Summer School on Probability Theory, Saint-Flour, 6–24 July 2004
29. Lyons, T.J., Qian, Z.: System Control and Rough Paths. Oxford Mathematical Monographs. Oxford University Press, Oxford (2002)
30. Sussmann, H.J.: On the gap between deterministic and stochastic ordinary differential equations. Ann. Probab. **6**(1), 19–41 (1978)

Chapter 9
Adaptive Sparse Grids in Reinforcement Learning

Jochen Garcke and Irene Klompmaker

Abstract We propose a model-based online reinforcement learning approach for continuous domains with deterministic transitions using a spatially adaptive sparse grid in the planning stage. The model learning employs Gaussian processes regression and allows a low sample complexity. The adaptive sparse grid is introduced to allow the representation of the value function in the planning stage in higher dimensional state spaces. This work gives numerical evidence that adaptive sparse grids are applicable in the case of reinforcement learning.

9.1 Introduction

We consider function approximation techniques for reinforcement learning (RL). Reinforcement learning is a computational approach to learning, where an agent tries to maximise the total amount of reward it receives when interacting with a complex, uncertain environment [33]. The setting is very closely related to solving optimal control problems using Hamilton-Jacobi Bellman (HJB) equations, but in contrast to that only a partial amount of the data describing the system is known. For example the state dynamics describing the evolution of a system are unknown and can only be observed by performing actions.

Formally the evolution of the problem in the control space is determined by the differential equation

$$\frac{\partial x(t)}{\partial t} = f(x(t), \beta(t)),$$

where $x(t)$ is the *state*, $\beta(t)$ the *action* and f is called *state dynamics*. The latter describes the effect of an action β taken in a particular state x, and gives the new

J. Garcke (✉)
University of Bonn, Wegelerstr. 6, 53115 Bonn, Germany
e-mail: garcke@ins.uni-bonn.de

I. Klompmaker

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_9

state $f(x, \beta)$ after the action is taken. Although we consider deterministic dynamics in this work, they could also be stochastic, to which situation our approach can be extended. For an initial state x_0 the choice of actions β therefore leads to a unique trajectory $x(t)$. Further, there is the *reinforcement* or *reward* function $r(x, \beta)$, which assigns each state (or state-action pair) a numerical value indicating the intrinsic desirability of that state. The aim is to find a policy which maximises the total reward in the long run, where rewards in states reached by a trajectory through the state space are taken into account. For simplicity we consider a deterministic *policies* $\pi(x)$, which assign each state a unique action, i.e., $\beta = \pi(x)$; it is a mapping from perceived states of the environment to actions to be taken when in those states.

There are many types of reinforcement learning problems: state dynamics known or not, discrete or continuous case, model-based or model-free, deterministic or stochastic [33]. What they all have in common is that they solve an optimal control problem, at least implicitly. The difference of reinforcement learning in comparison to optimal control problems is that the state dynamics and the reinforcement function are, a priori, at least partially unknown. Nevertheless, it is a problem of optimal control and the dynamic programming method is usually employed to estimate the best future cumulative reinforcement.

In this work we consider a deterministic *model-based* reinforcement learning approach in a continuous state space with unknown state dynamics, but known rewards. As in [9,23] and related methods our approach consists of two ingredients, a *model-learner* and a *planner*. By performing an action β in a state x the algorithm interacts with the environment and observes a sample $f(x, \beta)$ of the state dynamics. Based on such sample transitions $\{x_k, \beta_k, f(x_k, \beta_k)\}_{k=1,\ldots,K}$ the model-learner then estimates the state dynamics. On the other hand, given the current model the planner aims to find the best possible action β in a state x, i.e. those which is part of the trajectory starting at x with the highest total reward, and thereby determines an approximation π to the optimal policy π^*. With more and more samples of the state dynamics the model-learner is assumed to become more accurate, while the derived actions are supposed to get closer to the optimal ones from π^*.

For model-learning we use Gaussian process regression as in [23], while for the planner we employ adaptive sparse grid interpolation. The discretization technique of sparse grids allows to cope with the curse of dimensionality to some extent. It is based on a hierarchical multilevel basis [36] and a sparse tensor product construction. The underlying idea was first used for numerical integration and interpolation [34]. Subsequently, the sparse grid method has been developed for the solution of partial differential equations [37]. By now, it is also successfully used for, e.g., integral equations, stochastic differential equations, machine learning, or approximation, see the overview articles [10, 17] and the references cited therein.

For the representation of a function f defined over a d-dimensional domain, the conventional sparse grid approach employs $\mathcal{O}(h_n^{-1} \cdot \log(h_n^{-1})^{d-1})$ grid points in the discretization process, where $h_n := 2^{-n}$ denotes the mesh width. It can be shown that the order of approximation to describe a function f, provided that certain mixed smoothness conditions hold, is $\mathcal{O}(h_n^2 \cdot \log(h_n^{-1})^{d-1})$. This is in contrast to

conventional grid methods, which need $\mathcal{O}(h_n^{-d})$ for an accuracy of $\mathcal{O}(h_n^2)$, albeit for less stringent smoothness conditions. Thus, the curse of dimensionality of full grid methods arises for sparse grids to a much smaller extent. In case the smoothness conditions are not fulfilled, spatially adaptive sparse grids have been used with good success [6, 10, 15, 31]. There, as in any adaptive grid refinement procedure, the employed hierarchical basis functions are chosen during the actual computation depending on the function to be represented. In regard to adaptivity, closely related work in reinforcement learning was presented in [28, 30], in contrast to these approaches we investigate sparse grids in the planner and use a model-based setting.

The presented sparse grid approach for reinforcement learning is an extension of a semi-Lagrangian scheme for HJB-equations on an adaptive sparse grid, which was introduced in [6]. There it was empirically shown that for problems related to the front propagation model, the number of grid points needed in higher dimensions to approximately represent the involved functions with a given threshold error can be small. Thus, the approach is able to circumvent the curse of dimensionality of standard grid approaches for Hamilton-Jacobi Bellman equations to some extent. This work now shows numerical results for the case of reinforcement learning and gives evidence that adaptive sparse grids can be used there as well.

But note that the sparse grid scheme is not monotone as the interpolation with sparse grids is not monotone [29, 31]. Thus neither convergence towards the viscosity solutions of Hamilton-Jacobi Bellman equations nor stability of the scheme can presently be guaranteed, even for the linear advection equation. Consequently, these properties do not necessarily hold in the case of reinforcement learning either; numerically divergent behaviour of the adaptive sparse grid approach can be observed in certain situations. To this end, further analytical work on the scheme, both for the HJB and the RL case, is necessary.

9.2 Reinforcement Learning

Our reinforcement learning approach is based on the procedure presented in [23], a model-based online reinforcement learning approach for continuous domains with deterministic transitions. It separates function approximation in the model learner from the interpolation in the planner. For model-learning we use Gaussian process regression as in [23], but we replace the equidistant grid in the planner by an adaptive sparse grid procedure similar to the one used for HJB equations [6]. The overall approach assumes some properties of the reinforcement learning problems under consideration: We consider discrete actions, a smooth transition function, i.e. an action performed on states which are close in state space must lead to successor states that are close, deterministic transitions, and known reward functions. The latter two are mainly for simplicity, the ingredients of the approach can be extended to the non-deterministic case, and learning a reward function would just be one more function to be learned. Since our goal is to investigate the applicability of

adaptive sparse grids in the planning stage of a reinforcement learning setting, a simple setting is advantageous to concentrate on the effect of, and interplay with, unknown and only approximately learned state dynamics, which is the extension in comparison to [6].

We assume that the *state space* $\mathscr{X}$ is a hyperrectangle in $\mathbb{R}^d$, which is justified for many applications, and that we have a finite *action space* $\mathscr{B}$, this might involve discretizations of continuous controls. For simplicity we assign each action a unique number $1, \ldots, |\mathscr{B}|$. Note that we use here a setting where all actions involve the same time horizon τ, which therefore can be omitted from the exposition for simplification. In general, temporal aspects need to be taken into account, see e.g. [26, 30]. The function $f : \mathscr{X} \times \mathscr{B} \to \mathscr{X}$ describes the *state dynamics*. In our setting the state dynamics are (at least partially) unknown, only an approximate model $\hat{f} : \mathscr{X} \times \mathscr{B} \to \mathscr{X}$, which will be learneded from samples, is available with $f \approx \hat{f}$. Finally $r : \mathscr{X} \times \mathscr{B} \to \mathbb{R}$ is the *reward* function.

For a state $x \in \mathscr{X}$ one is interested in determining a sequence of actions $\beta_0, \beta_1, \ldots$ such that the accumulated reward is maximised, this is given by the optimal value function $v^*(x)$

$$v^*(x) := \max_{\beta_0, \beta_1, \ldots} \left\{ \sum_{t=0}^{\infty} \gamma^t \cdot r(x_t, \beta_t) \mid x_0 = x, x_{t+1} = f(x_t, \beta_t) \right\},$$

where $0 < \gamma < 1$ is the discount factor, which determines the importance of future rewards [5, 33].

The value iteration, the employed basic numerical scheme, is based on the dynamic programming principle and can be formulated as

$$v^{n+1}(x) = \max_{\beta \in \mathscr{B}} \left[\gamma \cdot v^n(\hat{f}(x, \beta)) + r(x, \beta) \right], \tag{9.1}$$

which computes the value function $v^*(x)$ in the limit $n \to \infty$, see e.g. [1, 25, 26]. Note that in a similar fashion the value function v^π for a fixed policy π can be computed. This formulation for the computation of the value function, the planning, is valid for both situations, a known model f and a to be learned model $\hat{f}$. In addition, a numerical discretization of the value function is necessary, in particular for continuous domains.

For any given value function v, e.g. a suitable numerical approximation $\hat{v}$ of the optimal value function v^* computed by value iteration using a discretization approach, the corresponding control policy $\pi(x)$ determined by v can easily be obtained, since at each state the optimal action can be chosen depending on the given value function v as follows

$$\pi(x) \in \arg\max_{\beta \in \mathscr{B}} \left[\gamma \cdot v(\hat{f}(x, \beta)) + r(x, \beta) \right].$$

The overall reinforcement learning approach now consists of three parts as outlined in the following algorithm.

Algorithm 1 Generic model-based reinforcement learning approach

while learning **do**
  interact with system and store observed transitions
  learn model $\hat{f}$ based on observed transitions
  for planning use model $\hat{f}$ to determine $\hat{v}^\pi$
end while

In the following we will describe the model learning using Gaussian process regression and the planning procedure in the next sections, where for the representation of the value function v we use a finite element approach, similar to [1,25,26], but based on a sparse grid.

9.2.1 Model Learning with Gaussian Processes

We now describe how we, following [23], learn the model from samples which are obtained by interactions with the environment. In its core, model learning is a regression problem. In this work we aim to concentrate on evaluating sparse grids for the planning stage, and therefore apply for the model learning a regression approach successfully used before in reinforcement learning, namely Gaussian processes (GPs) [13,23,32]. As the kernel the squared exponential is employed

$$k(x, x'; v_0, b, \theta) = v_0 \exp\left\{-0.5(x - x')^2\theta\right\},$$

where v_0, b, θ are the to be determined hyperparameters. In our view, a main advantage of Gaussian processes in this application is the possible automatic determination of the hyperparameters in a controlled fashion using a maximum likelihood approach. This is here in particular relevant since, as we will see, one needs to repeatedly compute, or update, regression models. Nevertheless, it would be interesting to investigate sparse grid regression [16, 31] in the model learning phase as well.

The input data for the regression algorithm are the taken actions and their resulting state, i.e. $\mathscr{S} = \{x_k, \beta_k, x_{k+1}\}_{k=1,2,\ldots}$, where $x_{k+1} = f(x_k, \beta_k)$. As noted before, the transition function $f : \mathscr{X} \times \mathscr{B} \to \mathscr{X}$ is d-dimensional: $f(x, \beta) = [f_1(x, \beta), \ldots, f_d(x, \beta)]^T$. One can either estimate f directly, i.e. the absolute transitions, or the relative change $x_{k+1} - x_k$, the latter we do as in [23]. We train multiple Gaussian processes, one for each action and output dimension, and use the combined predictions afterwards. In other words, a GP_{ij} is trained for the output in the i-th dimension of the j-th action, using the data when that action was taken, i.e. the input data for GP_{ij} is $\mathscr{S}_{ij} = \left\{x_k, x_{k+1}^{(i)} - x_k^{(i)}\right\}_{\beta_k = j, k=1,2,\ldots}$. The hyperparameters are computed for each individual GP_{ij} by optimizing the marginal likelihood. For any test point x the GP_{ij} gives a distribution over target values

$\mathcal{N}(\mu_{ij}, \sigma_{ij})$ with mean $\mu_{ij}(x)$ and variance $\sigma_{ij}^2(x)$. The change in the i-th coordinate of the state under the j-th action is predicted by the mean μ_{ij}, where the variance can be interpreted as the uncertainty of the prediction. The full learned model consists of $d \cdot |\mathcal{B}|$ Gaussian processes and the predicted successor state $\hat{f}(x, \beta)$ for any action β taken in state x is then

$$\hat{f}(x, \beta) := \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(d)} \end{bmatrix} + \begin{bmatrix} \mu_{1\beta}(x) \\ \vdots \\ \mu_{d\beta}(x) \end{bmatrix},$$

see [23] for more details on model learning with Gaussian processes.

9.3 Planning with Sparse Grids

We consider in the following the discretization needed for the planner in the case of a continuous state-space and concentrate on the function approximation aspect. These techniques are used in the context of Hamilton-Jacobi Bellman equations as a device for having a numerical representation of the value function. They discretize an HJB-equation (with a resolution ε) into a dynamic programming (DP) problem for some stochastic Markov Decision Process (MDP). Using DP techniques the MDP can then be solved. The convergence of the solution V^ε of the discrete MDP to the value function V of the continuous problem for $\varepsilon \to 0$ can be proven under assumptions on the discretization scheme, namely it being (using suitable definitions) a consistent, monotone and uniformly continuous numerical procedure to solve the underlying HJB-equation. Generalizing these proofs (in regard to the deterministic or stochastic setting, the regularity of the value function and the properties of the discretization scheme) is an active field of research [2–4,24,26,28]. Typical discretization schemes are of finite-difference type [7, 8, 25], (operator) splitting methods [35], or control schemes based on the dynamic programming principle (e.g. [11]).

The idea of a discretized HJB-equation was adopted to the field of reinforcement learning [26, 28, 30]. Since the state dynamics and the reinforcement function are not perfectly known, the original convergence proof [4] for DP was generalized to the case where only approximations of these are known. Convergence was shown when the number of iterations of the RL scheme (for the approximation of the state dynamics and the reinforcement function) goes to infinity and the discretization step tends to zero. The result applies in a general form to model-based or model-free RL algorithms, for off-line or on-line methods, for deterministic or stochastic dynamics and finite element or finite difference discretization [26].

Adaptive finite difference grids and a posteriori error estimates using the methodology from the numerical solution of partial differential equations were studied for the deterministic HJB-equation [19] and later generalized to the stochastic

case [20]. Adaptive schemes are particularly important since often the value function is non-smooth. In [28] different spatial refinement strategies were studied, in particular a heuristic is proposed which refines the grid mostly where there is a transition in the optimal control. These discretization approaches are limited in the number of dimensions due to the curse of dimensionality, i.e. the complexity grows exponentially with the number of dimensions.

An important aspect of these grid based approaches is the inherent locality in the schemes and its properties. Although, for example, formulated using a finite element representation, one does not solve a 'global' Galerkin-type problem, but the convergence is due to local properties of a function defined on a simplex (or box). This local view unfortunately does not hold for sparse grids, which renders the usual theoretical justifications of these grid approaches for HJB-equations invalid. Nevertheless, the empirical observations in [6] and this work give evidence that spatial adaptive sparse grids can be used in this setting, although further detailed investigations are necessary to provide criteria when this is the case.

In case of a suitably chosen, for now fixed, discretization grid Ω with corresponding function space V the planning step based on dynamic programming and using a model $\hat{f}$ can be written as

- Suitably initialize $v_0 \in V$.
- Iterate for $n = 0, 1, 2, \ldots$ until convergence, e.g. $|v^{n+1}(x) - v^n(x)| < tol \quad \forall x \in \Omega$,

$$v^{n+1}(x) = \max_{\beta \in \mathscr{B}} [\gamma \cdot v^n(\hat{f}(x, \beta)) + r(x, \beta)] \quad \forall x \in \Omega. \tag{9.2}$$

Here, $v^n \in V$ is the numerical solution computed by the scheme at DP step n, which in some cases can be interpreted as a time step, and the value $v^n\left(\hat{f}(x, \beta)\right)$ denotes the interpolation of v^n at point $\hat{f}(x, \beta)$. Note that for our later experiment the minimization over the set $\mathscr{B}$ can be done in a straightforward way by evaluating the function values for each possible action $b \in \mathscr{B}$. In case this number is too large, or if the set $\mathscr{B}$ is infinite, then a minimisation procedure which uses only evaluations of the objective function, and not its derivatives, could be performed without changing the main steps of the scheme.

Note that many reinforcement learning algorithms in continuous state-space, e.g. the ones employing approximate value iteration (AVI) [27, 28, 33] for the dynamic programming part, use supervised machine learning methods (i.e. regression algorithms [21]) to achieve the function approximation, also called *generalization* in this context [5, 33]. For example neural networks [5] or decision trees [12] can be used. But there is no general convergence of the algorithms, and the combination of DP methods with function approximation may produce unstable or divergent results even when applied to very simple problems [33]. Nevertheless, for schemes which are linear in the model parameters, a convergence proof is available (see [33]). Finally note that there are recent results indicating that the contribution of the error due to the approximation of the Bellman operator at each iteration is more prominent in later iterations of AVI and the effect of an error term in the earlier iterations decays

exponentially fast [14]. To put more emphasis on having a lower Bellman error at later iterations one could increase the number of samples during the scheme or use more powerful function approximators in the end.

Having set the background for function approximation for the value function in the planning stage we now describe our adaptive sparse grid procedure, which is based on [6].

9.3.1 Sparse Grids

For ease of presentation we will consider the domain $\Omega = [0,1]^d$ in this section. Let $\underline{l} = (l_1, \ldots, l_d) \in \mathbb{N}^d$ denote a multi-index. We define the anisotropic grid $\Omega_{\underline{l}}$ on Ω with mesh width $h_{\underline{l}} := (h_{l_1}, \ldots, h_{l_d}) := (2^{-l_1}, \ldots, 2^{-l_d})$. It has, in general, different but equidistant mesh widths h_{l_t} in each coordinate direction t, $t = 1, \ldots, d$. The grid $\Omega_{\underline{l}}$ thus consists of the points $x_{\underline{l},\underline{j}} := (x_{l_1,j_1}, \ldots, x_{l_d,j_d})$, with $x_{l_t,j_t} := j_t \cdot h_{l_t} = j_t \cdot 2^{-l_t}$ and $j_t = 0, \ldots, 2^{l_t}$. For any grid $\Omega_{\underline{l}}$ we define the associated space $V_{\underline{l}}$ of piecewise d-linear functions

$$V_{\underline{l}} := \operatorname{span}\{\phi_{\underline{l},\underline{j}} \mid j_t = 0, \ldots, 2^{l_t},\ t = 1, \ldots, d\}, \tag{9.3}$$

which is spanned by the conventional basis of d-dimensional piecewise d-linear hat functions

$$\phi_{\underline{l},\underline{j}}(\underline{x}) := \prod_{t=1}^{d} \phi_{l_t,j_t}(x_t). \tag{9.4}$$

The one-dimensional functions $\phi_{l,j}(x)$ are defined by

$$\phi_{l,j}(x) = \begin{cases} 1 - |x/h_l - j|, & x \in [(j-1)h_l, (j+1)h_l] \cap [0,1], \\ 0, & \text{otherwise.} \end{cases} \tag{9.5}$$

The multi-index $\underline{l} \in \mathbb{N}^d$ denotes the level, i.e. the discretization resolution, be it of a grid $\Omega_{\underline{l}}$, of a space $V_{\underline{l}}$, or of a function $f_{\underline{l}}$, whereas the multi-index $\underline{j} \in \mathbb{N}^d$ gives the position of a grid point $x_{\underline{l},\underline{j}}$ or its corresponding basis function $\phi_{\underline{l},\underline{j}}$.

We now define a hierarchical difference space $W_{\underline{l}}$ via

$$W_{\underline{l}} := V_{\underline{l}} \setminus \bigoplus_{t=1}^{d} V_{\underline{l}-\underline{e}_t}, \tag{9.6}$$

where $\underline{e}_t$ is the t-th unit vector. In other words, $W_{\underline{l}}$ is spanned by all $\phi_{\underline{k},\underline{j}} \in V_{\underline{l}}$ which are not included in any of the spaces $V_{\underline{k}}$ smaller[1] than $V_{\underline{l}}$. To complete the definition, we formally set $V_{\underline{l}} := \emptyset$, if $l_t = 0$ for at least one $t \in \{1, \ldots, d\}$. As can be easily seen from (9.3) and (9.6), the definition of the index set

$$\mathsf{B}_{\underline{l}} := \left\{ \underline{j} \in \mathbb{N}^d \;\middle|\; \begin{array}{ll} j_t = 1, \ldots, 2^{l_t} - 1, & j_t \text{ odd}, t = 1, \ldots, d, \text{ if } l_t > 1, \\ j_t = 0, 1, 2, & t = 1, \ldots, d, \text{ if } l_t = 1 \end{array} \right\} \tag{9.7}$$

leads to

$$W_{\underline{l}} = \operatorname{span}\{\phi_{\underline{l},\underline{j}} \,|\, \underline{j} \in \mathsf{B}_{\underline{l}}\}. \tag{9.8}$$

The family of functions

$$\left\{ \phi_{\underline{l},\underline{j}} \;\middle|\; \underline{j} \in \mathsf{B}_{\underline{l}} \right\}_{\underline{l}=(1,\ldots,1)}^{(n,\ldots,n)} \tag{9.9}$$

is just the hierarchical basis [36] of $V_n (:= V_{(n,\ldots,n)})$, which generalizes the one-dimensional hierarchical basis to the d-dimensional case with a tensor product ansatz. Observe that the supports of the basis functions $\phi_{\underline{l},\underline{j}}(\underline{x})$, which span $W_{\underline{l}}$, are disjoint for $\underline{l} > 1$.

Zenger [37] introduced so-called *sparse grids*, where hierarchical basis functions with a small support, and therefore a small contribution to the function representation, are not included in the discrete space of level n any more.

Formally, the sparse grid function space $V_n^s \subset V_n$ is defined as

$$V_n^s := \bigoplus_{|\underline{l}|_1 \le n+d-1} W_{\underline{l}}. \tag{9.10}$$

Every $f \in V_n^s$ now can be represented as

$$f_n^s(\underline{x}) = \sum_{|\underline{l}|_1 \le n+d-1} \sum_{\underline{j} \in \mathsf{B}_{\underline{l}}} \alpha_{\underline{l},\underline{j}} \phi_{\underline{l},\underline{j}}(\underline{x}). \tag{9.11}$$

The resulting grid which corresponds to the approximation space V_n^s is called sparse grid and is denoted by Ω_n^s.

The sparse grid space V_n^s has a size of order $\dim V_n^s = \mathcal{O}(2^n \cdot n^{d-1})$, see [10]. It thus depends on the dimension d to a much smaller degree than a standard full grid space whose number of degrees of freedom is $\mathcal{O}(2^{nd})$. Note that for the approximation of a function f by a sparse grid function $f_n^s \in V_n^s$ the error relation

[1] We call a discrete space $V_{\underline{k}}$ smaller than a space $V_{\underline{l}}$ if $\forall_t k_t \le l_t$ and $\exists t : k_t < l_t$. In the same way a grid $\Omega_{\underline{k}}$ is smaller than a grid $\Omega_{\underline{l}}$.

$$\|f - f_n^s\|_2 = \mathcal{O}\left(2^{-2n} \cdot n^{d-1}\right)$$

holds, provided that f fulfils the smoothness requirement $|f|_{H^2_{mix}} < \infty$ [10]. Therefore, sparse grids need much fewer points in comparison to a full grid to obtain an error of the same size.

9.3.2 Spatially Adaptive Sparse Grids

The sparse grid structure (9.10) defines an a priori selection of grid points that is optimal if certain smoothness conditions are met, i.e. if the function has bounded second mixed derivatives, and no further knowledge of the function is known or used. If the aim is to approximate functions which either do not fulfil this smoothness condition at all or show strongly varying behaviour due to finite but nevertheless locally large derivatives, adaptive refinement may be used. There, depending on the characteristics of the problem and the function at hand, adaptive refinement strategies decide which points and corresponding basis functions should be incrementally added to the sparse grid representation to increase the accuracy.

In the sparse grid setting, usually an error indicator stemming directly from the hierarchical basis is employed [15, 18, 31]: depending on the size of the hierarchical surplus $\alpha_{\underline{l},\underline{j}}$ it is decided whether a basis function is marked for further improvement or not. This is based on two observations: First, the hierarchical surplus indicates the absolute change in the discrete representation at point $x_{\underline{l},\underline{j}}$ due to the addition of the corresponding basis function $\phi_{\underline{l},\underline{j}}$, i.e. it measures its contribution to a given sparse grid representation (9.11) in the maximum-norm. And second, a hierarchical surplus represents discrete second mixed derivatives and hence can be interpreted as a measure of the smoothness of the considered function at point $x_{\underline{l},\underline{j}}$.

In the adaptive procedure we use an index set $\mathcal{I}$ to track the indices of the employed basis functions and denote the corresponding sparse grid by $\Omega_{\mathcal{I}}$ and the associated sparse grid space by $V_{\mathcal{I}}$, respectively. We start with a coarse initial sparse grid function $f_n^s \in V_n^s$ for some given small n as in (9.11). The index set is thus initialized as $\mathcal{I} := \{(\underline{l}, \underline{j}) \mid |\underline{l}|_1 \leq n + d - 1\}$. We proceed as follows: If, for any given index $(\underline{l}, \underline{j}) \in \mathcal{I}$, we have

$$|\alpha_{\underline{l},\underline{j}}| \cdot \|\phi_{\underline{l},\underline{j}}\| > \varepsilon \tag{9.12}$$

for some given constant $\varepsilon > 0$, then the index will be *marked*. Here, $\|\cdot\|$ is typically either the L^∞- or L^2-norm, but other norms or weighted mixtures of norms are used in practice as well. If an index is marked, all its $2d$ so-called *children* will be added to the index set $\mathcal{I}$ to refine the discretization, i.e. all $(\tilde{\underline{l}}, \tilde{\underline{j}})$ with $\tilde{\underline{l}} = \underline{l} + \underline{e}_t$ and $\tilde{\underline{j}} = \underline{j} + j_t \, \underline{e}_t \pm 1$ will be added to $\mathcal{I}$ for $t = 1, \ldots, d$. For the indices added that way it is possible that not all *parents* in all dimensions are already contained in the grid; note that in such cases, for algorithmic and consistency reasons, these missing parents have to be added to $\mathcal{I}$ as well. Thus for any $(\underline{l}, \underline{j}) \in \mathcal{I}$ all parents $(\tilde{\underline{l}}, \tilde{\underline{j}})$

with $\tilde{\underline{l}} \leq \underline{l}$ and $supp(\phi_{\tilde{\underline{l}},\tilde{\underline{j}}}) \cap supp(\phi_{\underline{l},\underline{j}}) \neq \emptyset$ are also in the index set $\mathscr{I}$. In other words, "holes" in the hierarchical structure are not allowed. In Algorithm 2 we give the adaptive refinement procedure. If the function values at the newly added grid points are easily available, the refinement step can be repeated until no indices are added any more [6]. Note that if a global error criterion is available one can perform an additional outer loop with successively decreasing ε until the measured global error falls below a given threshold ε_{glob}.

In a similar way one can use the value $|\alpha_{\underline{l},\underline{j}}| \cdot \|\phi_{\underline{l},\underline{j}}\|$ to *coarsen* the grid in case of over-refinement. If this value is smaller than some coarsening constant η, and no children of $(\underline{l},\underline{j})$ are in $\mathscr{I}$, the index will be removed from this set. In Algorithm 3 we give the coarsening step, where the procedure is repeated until no indices are being removed. The coarsening will in particular be relevant once we consider problems where the region in need of a higher resolution changes during the computation. This is relevant for time-dependent problems, but also for the planning considered in this work, which in some sense can be viewed as a time-dependent problem. More importantly, the value function to be represented can change during the computation due to changes in the learned model.

Algorithm 2 Spatially adaptive refinement step

Data: initial index set $\mathscr{I}$, to be refined function $v_{\mathscr{I}}$, refinement threshold ε
Result: refined index set $\mathscr{I}$, refined function $v_{\mathscr{I}}$
for $(\underline{l},\underline{j}) \in \mathscr{I}$ **do**
  if $|\alpha_{\underline{l},\underline{j}}| \cdot \|\phi_{\underline{l},\underline{j}}\| > \varepsilon$ **then**
    for $t = 1,\ldots,d$ **do**
      if $(\tilde{\underline{l}},\tilde{\underline{j}}) \notin \mathscr{I}$ for $\tilde{\underline{l}} = \underline{l} + \underline{e}_t$ and $\tilde{\underline{j}} \in \{\underline{j} + j_t\,\underline{e}_t \pm 1\}$ **then**
        $\mathscr{I} = \mathscr{I} \cup (\tilde{\underline{l}},\tilde{\underline{j}})$ ▷ add children which are not in $\mathscr{I}$
      end if
    end for
  end if
end for
check $\forall(\underline{l},\underline{j}) \in \mathscr{I}$ holds: $(\tilde{\underline{l}},\tilde{\underline{j}}) \in \mathscr{I}$ for $\tilde{\underline{l}} \leq \underline{l}$ and $supp(\phi_{\tilde{\underline{l}},\tilde{\underline{j}}}) \cap supp(\phi_{\underline{l},\underline{j}}) \neq \emptyset$
for all added indices $(\underline{l},\underline{j}) \in \mathscr{I}$ **do**
  initialize $\alpha_{\underline{l},\underline{j}} = 0$
end for

Algorithm 3 Spatially adaptive coarsening

Data: index set $\mathscr{I}$, coarsening threshold η, and $\alpha_{\underline{l},\underline{j}}\ \forall(\underline{l},\underline{j}) \in \mathscr{I}$
Result: coarsened index set $\mathscr{I}$
while indices are removed from $\mathscr{I}$ **do**
  for $(\underline{l},\underline{j}) \in \mathscr{I}$ **do**
    if $|\alpha_{\underline{l},\underline{j}}| \cdot \|\phi_{\underline{l},\underline{j}}\| < \eta$ **then**
      if $\forall t = 1,\ldots,d$: $(\tilde{\underline{l}},\tilde{\underline{j}}) \notin \mathscr{I}$ for $\tilde{\underline{l}} = \underline{l} + \underline{e}_t$ and $\tilde{\underline{j}} \in \{\underline{j} + j_t\,\underline{e}_t \pm 1\}$ **then**
        $\mathscr{I} = \mathscr{I} \backslash (\underline{l},\underline{j})$ ▷ remove if no children in $\mathscr{I}$
      end if
    end if
  end for
end while

9.4 Sparse Grid Based Scheme for Reinforcement Learning

With these ingredients we now can formulate our approach, which in the end is quite similar to the semi-Lagrangian scheme for Hamilton-Jacobi Bellman equations using spatial adaptive sparse grids introduced in [6], but with only an approximate model for the state dynamics.

The planning scheme is given in Algorithm 4. We perform p steps of the DP equation (9.2) and then do one refinement step. Note that the initial steps are randomly chosen since no infomation is exploitable, i.e. no samples from the state dynamics exist. This is repeated a certain number of times, or stops early in case the change in a DP step is small. Observe that it is not useful to aim for a convergence of the DP scheme in the initial stages of the overall RL procedure [14]. Since the learned model will take some time to be a reasonable approximation of the state dynamics, and can be quite coarse in the beginning, aiming for convergence of the planning step can even lead to wrong behaviour. Therefore we limit the number of DP steps per planning stage. The same reasoning is behind the idea of calling the refinement algorithm only after every p DP steps, the resolution of the discretization of the value function shall only grow slowly. Once the value iteration is finished we coarsen the obtained sparse grid to reduce the further computational effort.

Algorithm 4 Adaptive SG-planning in reinforcement learning

Data: initial index set $\mathcal{I}(0)$, initial function v^0, refinement constant ε, coarsening constant η, model $\hat{f}$, $k = 0$
Result: adaptive sparse grid solution $v^{k_e} \in V_{\mathcal{I}(k_e)}$
repeat
  $k = k + 1$
  $v^k(x) = \max_{\beta \in \mathcal{B}}[\gamma \cdot v^{k-1}(\hat{f}(x, \beta)) + r(x, \beta)] \quad \forall x \in \Omega_{\mathcal{I}}$ ▷ DP step (9.2)
  if $k \bmod p = 0$ **then**
    call Alg. 2 with $\mathcal{I}(k-1)$, v^k, ε ▷ refine $v^k \in V_{\mathcal{I}(k)}$ every p steps
  end if
until $|v^k(x) - v^{k-1}(x)| < tol \quad \forall x \in \Omega_{\mathcal{I}(k)}$ and $k < k_{\max}$
$k_e = k$
call Alg. 3 with $\mathcal{I}(k_e)$, η and v^{k_e} ▷ coarsen v^{k_e}

The overall reinforcement learning procedure is presented in Algorithm 5. The algorithm performs a number of interactions with the environment and thereby observes new state transitions $(x, \beta, f(x, \beta))$. Every T actions first the model is updated to integrate the newly observed data, and then the value function is updated in the planning step. Updating after every new transition information would be a quite costly procedure, which we hereby avoid. Like in [23] the experiment we are investigating is performed in episodes, i.e. the trajectory through the state space is re-started once the target state, if defined, is reached or after a certain number of steps, i.e. $T \cdot \#ep_{max}$. The latter prevents the algorithm from investigating uninteresting regions of the state space when the transition path does not follow a 'good' trajectory after 'bad' choices of actions due to model uncertainty or early stages of convergence to the value function.

Algorithm 5 SG-GP-scheme in reinforcement learning

Data: suitable initial index set $\mathscr{I}(0)$, refinement constant ε, coarsening constant η, initial state x_{init}, and episode length $\#ep_{max}$
Result: adaptive sparse grid solution $v^{n_e} \in V_{\mathscr{I}(n_e)}$
$\#ep = 0, n = 0, x_0 = x_{init}$
repeat
  for $t = 1, \ldots, T$ **do**
    action β_t is chosen according to $\beta_t = \arg\max_{\beta \in \mathscr{B}} [\gamma \cdot v^n(\hat{f}(x_{t-1}, \beta)) + r(x_{t-1}, \beta)]$
    interact with system by executing action β_t
    observe next state x_t and store transition $\mathscr{S} = \mathscr{S} \cup \{x_{t-1}, \beta_t, x_t\}$
  end for
  based on $\mathscr{S}$ learn model $\hat{f} = \{\mathsf{GP}_{ij}\}_{i=1,\ldots,d, j=1,\ldots,|\mathscr{B}|}$
  call Alg. 4 with $\mathscr{I}(n)$, v^n, ε, η, $\hat{f}$ ▷ for planning compute value function v^{n+1} on $\Omega_{\mathscr{I}(n+1)}$
  $n = n + 1, \#ep = \#ep + 1$ ▷ count number of chunks
  if $\#ep \bmod \#ep_{max} = 0$ or x_T is target state **then**
    $x_0 = x_{init}$
  else
    $x_0 = x_T$
  end if
until $|v^n(x) - v^{n-1}(x)| < tol \quad \forall x \in \Omega_{\mathscr{I}(n)}$
$n_e = n$

9.5 Experiments

We evaluate our approach on a well-known reinforcement learning benchmark, the mountain car problem [33]. The goal is to drive a car from the bottom of a valley to the top of a mountain. Since the car is underpowered it cannot climb the mountain directly, but instead has to build up momentum by first going in the opposite direction. The state space is two-dimensional, the position of the car is described by $x_1 \in [-1.2, 0.5]$ and $x_2 \in [-0.07, 0.07]$ is its velocity. The actions are accelerating by a fixed value to the left, right, or no acceleration, encoded by actions $\beta \in \{-1, +1, 0\}$. Every step gives a reward of -1, until the top of the mountain on the right satisfying $x_1 \geq 0.5$ is reached, the goal is therefore to have as few steps as possible to reach the top. This experimental setup is the same as in [23], which in modification to [33] sets a maximal episode length of 500, while model and value function are relearned every 50 steps, uses a discount factor $\gamma = 0.99$ and every episode starts at $x_0 = (\pi/6, 0)$. The parameters of the sparse grid approach were $\varepsilon = 0.01$ and $\eta = 0.002$, while an initial grid of $n = 3$ was used. In Algorithm 4 we used $p = 20$ and $k_{max} = 200$.

In Fig. 9.1 we give the final value function v^{n_e} and the employed adaptive sparse grid $\Omega_{\mathscr{I}(n_e)}$, which consists of 4,054 grid points. We observe that, as expected, the majority of the grid points are where the value function has a large gradient. In context of the problem there is a high gradient, where on one side the mountain can be reached directly, on the other side only with first gaining momentum by going in

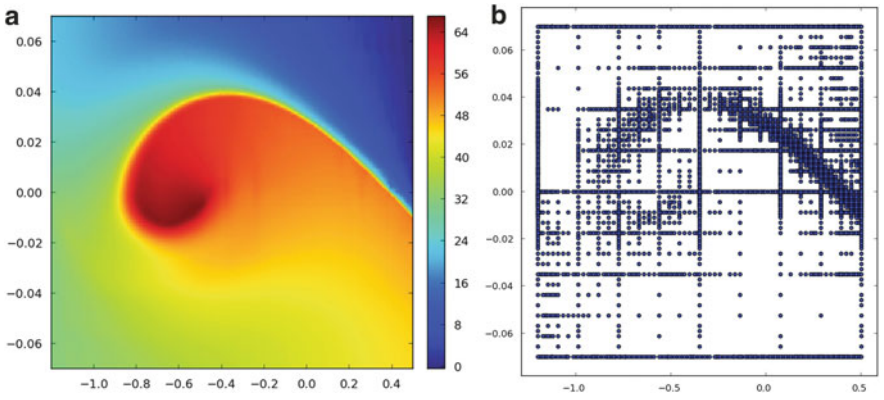


Fig. 9.1 Results for the mountain car problem. (**a**) Value function v^{n_e}. (**b**) Adaptive sparse grid $\Omega_{\mathscr{I}(n_e)}$

the opposite direction. The algorithm converged after 5 episodes, with 479 different states visited in total. The goal was reached in each episode as follows:

Episode	1	2	3	4	5
Steps to goal	159	105	103	104	104

Note that the optimal number of steps for this problem is 103 [23] and the number steps we observe per episode is essentially (only a graph is shown) as in that paper, which uses a full grid. The standard online model-free RL algorithm Sarsa(λ) with tile coding [33] needs many more episodes to get below 150 steps and still is a couple of steps away from 103 even after 1,000 episodes [23]. During the course of the Sarsa(λ) algorithm many more states are visited than in our procedure.

9.6 Conclusion

The combination of the Gaussian processes approach for model learning, which achieves a low sample complexity, with the adaptive sparse grid procedure, which breaks the curse of dimensionality to some extent and allows function representation in higher dimensional state spaces, is applicable for reinforcement learning.

However, due to the lack of monotonicity of the sparse grid approach, the stability of the scheme presently cannot be guaranteed. In practise, we do observe divergent behaviour of the proposed scheme with unfavourable settings of the parameters of the refinement algorithm. Additionally, the initial randomly chosen steps can lead the algorithm off-track, although this can also happen for other reinforcement learning approaches. In such a case the combination of adaptivity and a too wrong model of the state dynamics can lead to divergence, even with otherwise, i.e. for other initial samples, suitable refinement parameters. In any case, theoretical investigations are needed to provide criteria, in particular useable for the actual

computation, under which conditions on the problem and which settings of the algorithm the scheme can successfully be used in reinforcement learning.

Furthermore, a bottleneck in the time complexity of the algorithm is the evaluation of the adaptive sparse grid function, which at this stage prevents us from detailed numerical experiments in higher dimensions, although an adaptive sparse grid would cope with the higher dimensional setting. Recently it was shown that using a specific reordering of the steps of the point evaluations together with a GPU-based parallelisation can achieve speed-ups of almost 50 in comparison to the standard implementation of adaptive sparse grid [22], employing this procedure in our scheme would improve the runtime significantly and allow higher dimensional examples to be finished in reasonable time.

References

1. Bardi, M., Capuzzo-Dolcetta, I.: Optimal control and viscosity solutions of Hamilton-Jacobi-Bellman equations. In: Systems and Control: Foundations and Applications. Birkhäuser, Boston (1997)
2. Barles, G., Jakobsen, E.R.: On the convergence rate of approximation schemes for Hamilton-Jacobi-Bellman equations. M2AN Math. Model. Numer. Anal. **36**(1), 33–54 (2002)
3. Barles, G., Jakobsen, E.R.: Error bounds for monotone approximation schemes for parabolic Hamilton-Jacobi-Bellman equations. Math. Comput. **76**(240), 1861–1893 (2007)
4. Barles, G., Souganidis, P.: Convergence of approximation schemes for fully nonlinear second order equations. Asymptot. Anal. **4**(3), 271–283 (1991)
5. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming. Athena Scientific, Belmont (1996)
6. Bokanowski, O., Garcke, J., M-Griebel, Klompmaker, I.: An adaptive sparse grid semi-Lagrangian scheme for first order Hamilton-Jacobi Bellman equations. J. Sci. Comput. **55**(3), 575–605 (2013)
7. Bonnans, J.F., Ottenwaelter, E., Zidani, H.: A fast algorithm for the two dimensional HJB equation of stochastic control. M2AN, Math. Model. Numer. Anal. **38**(4), 723–735 (2004)
8. Bonnans, J.F., Zidani, H.: Consistency of generalized finite difference schemes for the stochastic HJB equation. SIAM J. Numer. Anal. **41**(3), 1008–1021 (2003)
9. Brafman, R., Tennenholtz, M.: R-max-a general polynomial time algorithm for near-optimal reinforcement learning. J. Mach. Learn. Res. **3**, 213–231 (2002)
10. Bungartz, H.J., Griebel, M.: Sparse grids. Acta Numer. **13**, 1–123 (2004)
11. Camilli, F., Falcone, M.: An approximation scheme for the optimal control of diffusion processes. RAIRO, Modélisation Math. Anal. Numér. **29**(1), 97–122 (1995)
12. Chapman, D., Kaelbling, L.P.: Input generalization in delayed reinforcement learning: an algorithm and performance comparisons. In: Proceedings of the 12th International Joint Conference on Artificial Intelligence, San Mateo, pp. 726–731 (1991)
13. Deisenroth, M.P., Rasmussen, C., Peters, J.: Gaussian process dynamic programming. Neurocomputing **72**(7–9), 1508–1524 (2009)
14. Farahmand, A.M., Munos, R., Szepesvári, C.: Error propagation for approximate policy and value iteration. In: NIPS. Lafferty, J., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A. (eds.) Advances in Neural Information Processing Systems, vol. 23, pp. 568–576. (2010)
15. Feuersänger, C.: Sparse grid methods for higher dimensional approximation. Dissertation, Institut für Numerische Simulation, Universität Bonn (2010)
16. Garcke, J.: Regression with the optimised combination technique. In: Cohen, W., Moore, A. (eds.) Proceedings of the 23rd ICML'06, Pittsburgh, pp. 321–328. ACM, New York (2006)

17. Garcke, J.: Sparse grids in a nutshell. In: Sparse Grids and Applications. Lecture Notes in Computational Science and Engineering, vol. 88, pp. 57–80. Springer, Berlin/New York (2013)
18. Griebel, M.: Adaptive sparse grid multilevel methods for elliptic PDEs based on finite differences. Computing **61**(2), 151–179 (1998)
19. Grüne, L.: An adaptive grid scheme for the discrete Hamilton-Jacobi-Bellman equation. Numer. Math. **75**(3), 319–337 (1997)
20. Grüne, L.: Error estimation and adaptive discretization for the discrete stochastic Hamilton-Jacobi-Bellman equation. Numer. Math. **99**(1), 85–112 (2004)
21. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, New York (2001)
22. Heinecke, A., PflügerS, D.: Multi- and many-core data mining with adaptive sparse grids. In: Proceedings of the 8th ACM International Conference on Computing Frontiers, CF'11, Ischia, pp. 29:1–29:10. ACM (2011)
23. Jung, T., Stone, P.: Gaussian processes for sample efficient reinforcement learning with RMAX-Like exploration. In: Balcázar, J.L., Bonchi, F., Gionis, A. Sebag, M. (eds.) ECML/PKDD 2010 (1). Lecture Notes in Computer Science, vol. 6321, pp. 601–616. Springer, Berlin/New York (2010)
24. Krylov, N.V.: The rate of convergence of finite-difference approximations for Bellman equations with Lipschitz coefficients. Appl. Math. Optim. **52**(3), 365–399 (2005)
25. Kushner, H., Dupuis, P.: Numerical Methods for Stochastic Control Problems in Continuous Time. No. 24 in Applications of Mathematics, 2nd edn. Springer, New York (2001)
26. Munos, R.: A study of reinforcement learning in the continuous case by the means of viscosity solutions. Mach. Learn. **40**(3), 265–299 (2000)
27. Munos, R.: Performance bounds in L_p-norm for approximate value iteration. SIAM J. Control Optim. **46**(2), 541–561 (2007)
28. Munos, R., Moore, A.: Variable resolution discretization in optimal control. Mach. Learn. **49**(2–3), 291–323 (2002)
29. Noordmans, J., Hemker, P.: Application of an adaptive sparse grid technique to a model singular perturbation problem. Computing **65**, 357–378 (2000)
30. Pareigis, S.: Adaptive choice of grid and time in reinforcement learning. In: NIPS. MIT, Cambridge (1997).
31. Pflüger, D.: Spatially Adaptive Sparse Grids for High-Dimensional Problems. Verlag Dr. Hut, München (2010)
32. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT, Cambridge (2006)
33. Sutton, R.S., Barto, A.: Reinforcement Learning: An Introduction. MIT, Cambridge (1998)
34. Smolyak, S.A.: Quadrature and interpolation formulas for tensor products of certain classes of functions. Dokl. Akad. Nauk SSSR **148**, 1042–1043 (1963)
35. Tourin, A.: Splitting methods for Hamilton-Jacobi equations. Numer. Methods Partial Differ. Equ. **22**(2), 381–396 (2006)
36. Yserentant, H.: On the multi-level splitting of finite element spaces. Numerische Mathematik **49**, 379–412 (1986)
37. Zenger, C.: Sparse grids. In: Hackbusch, W. (ed.) Parallel Algorithms for Partial Differential Equations, Proceedings of the Sixth GAMM-Seminar, Kiel, 1990. Notes on Numerical Fluid Mechanics, vol. 31, pp. 241–251. Vieweg, Braunschweig (1991)

Chapter 10
A Review on Adaptive Low-Rank Approximation Techniques in the Hierarchical Tensor Format

Jonas Ballani, Lars Grasedyck, and Melanie Kluge

Abstract The hierarchical tensor format allows for the low-parametric representation of tensors even in high dimensions d. On the one hand, this format provides a robust framework for approximate arithmetic operations with tensors based on rank truncations, which can be exploited in iterative algorithms. On the other hand, it can be used for the direct approximation of high-dimensional data stemming, e.g., from the discretisation of multivariate functions. In this review, we discuss several strategies for an adaptive approximation of tensors in the hierarchical format by black box type techniques, including problems of tensor reconstruction and tensor completion.

10.1 Introduction

High-dimensional problems are encountered in many areas of practical interest, as e.g. in stochastics, quantum chemistry, or optimisation. In this review, we consider high-dimensional data that can be represented or approximated by a *tensor*

$$A \in \mathbb{R}^{n_1 \times \ldots \times n_d}$$

of *order* (or *dimension*) $d \in \mathbb{N}$ with $n_1, \ldots, n_d \in \mathbb{N}$. As soon as the order d is large enough, the explicit representation of A in terms of all its entries $A_{(i_1,\ldots,i_d)}$, $i_\mu = 1, \ldots, n_\mu, \mu = 1, \ldots, d$, becomes prohibitively expensive. This has motivated the development of data-sparse tensor representations that can be applied even in high dimensions d. For a detailed introduction to tensor representations we refer the reader to [21, 22, 27].

A crucial question in applications is how to find a data-sparse representation of A from the evaluation of a (small) subset of its entries $A_{(i_1,\ldots,i_d)}$. For example,

J. Ballani (✉) • L. Grasedyck • M. Kluge
RWTH Aachen, Templergraben 55, 52056 Aachen, Germany
e-mail: ballani@igpm.rwth-aachen.de; lgr@igpm.rwth-aachen.de; kluge@igpm.rwth-aachen.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_10

this setting arises when the tensor A is given by the values of some function $\phi : [0,1]^d \to \mathbb{R}$ on a tensor grid, i.e.

$$A_{(i_1,\dots,i_d)} := \phi(\xi_{1,i_1},\dots,\xi_{d,i_d}), \qquad \xi_{\mu,i_\mu} \in [0,1],\ \mu = 1,\dots,d.$$

The function ϕ might, e.g., be induced by a functional Φ of the solution $u : S \times [0,1]^d \to \mathbb{R}$ of a parameter-dependent PDE posed on a physical domain $S \subset \mathbb{R}^3$. In this case, each evaluation of $\phi(y) := \Phi(u(\cdot, y))$ for $y \in [0,1]^d$ requires the (possibly costly) solution of a PDE in S.

The availability of certain entries of A strongly depends on the nature of the application. In this review, we distinguish three typical scenarios for the data-sparse approximation of A from a subset $\Omega \subset \mathcal{I}$ of its entries, where $\mathcal{I} := \times_{\mu=1}^{d} \{1 \dots, n_\mu\}$:

1. Adaptive Tensor Sampling: The user can permanently interact with the application. Upon request, the application returns an entry A_i for any $i \in \mathcal{I}$. The set $\Omega \subset \mathcal{I}$ can therefore be determined adaptively by the user during the approximation process.
2. Non-adaptive Tensor Sampling: The user can interact with the application only once. The set $\Omega \subset \mathcal{I}$ is fixed by the user in advance and given to the application. The application returns the values A_i for all $i \in \Omega$.
3. Tensor Completion: The user cannot interact with the application. The application gives a set $\Omega \subset \mathcal{I}$ and the corresponding values A_i for all $i \in \Omega$ to the user.

Given either of the three policies, the choice of a suitable data-sparse representation of A is typically up to the user. Here, we are interested in approximations of A that possess a certain low-rank structure. Low-rank tensor techniques have, e.g., been used in iterative algorithms for the solution of parametric linear systems and eigenvalue problems [4, 29, 40], for the approximation of multivariate functions and parameter-dependent integrals [2, 3, 8], and for the solution of parametric PDEs [5, 25, 31].

A quite general framework for the low-rank representation of tensors has been introduced in [23] which we further analysed in [18]. In the so-called *hierarchical tensor* (or *hierarchical Tucker*) *format*, a tensor is represented by a number of parameters that scales only linearly in the dimension d. As a key ingredient, this format relies on an appropriate hierarchy of subspaces which can be related to specific matrix representations of a given tensor. Based on this strong connection to matrices, powerful tools have been developed that allow for (approximate) arithmetic operations with tensors even in high dimensions d.

As an application for a hierarchical low-rank representation, we may consider a problem from the field of uncertainty quantification. Given the function ϕ from above and a probability density function $f : [0,1]^d \to \mathbb{R}$, the aim is to estimate the expected value

$$\mathbb{E}[\phi] = \int_{[0,1]^d} \phi(y) f(y) \, \mathrm{d}y.$$

This value can, e.g., be approximated by Monte Carlo or quasi-Monte Carlo methods [11] by generating a finite number of random or quasi-random sampling points $y_i \in [0,1]^d$ which yields an approximation of $\mathbb{E}[\phi]$ by a simple average. This sampling strategy falls into the second category from above, but it does not exploit any structural information from ϕ. A low-rank approximation of ϕ in the hierarchical tensor format reduces the computation of $\mathbb{E}[\phi]$ to the evaluation of a simple scalar product [5]. Moreover, the low-rank structure can be exploited by adaptive sampling strategies from the first [7, 34, 36, 37] or second [26] category from above in order to decrease the number of required samples dramatically to reach a given target accuracy.

The rest of this review is organised as follows. In Sect. 10.2, we recall the most important tensor representations from the literature. In Sects. 10.3 and 10.4, we study the approximation of tensors in the hierarchical format by strategies that belong to the first and second category from above. The problem of tensor completion is considered in Sect. 10.5.

10.2 Low-Rank Tensor Representations

Given $d \in \mathbb{N}$ and $n_1, \dots, n_d \in \mathbb{N}$, let $\mathscr{I}_\mu := \{1, \dots, n_\mu\}$ for $\mu = 1, \dots, d$ and define

$$\mathscr{I} := \mathscr{I}_1 \times \dots \times \mathscr{I}_d.$$

A full representation of a tensor $A \in \mathbb{R}^{\mathscr{I}}$ in terms of all entries typically leads to a storage complexity of $\mathscr{O}(n^d)$, $n := \max_{\mu=1,\dots,d} n_\mu$, which quickly becomes intractable if d gets large. Therefore, different data-sparse representations of tensors have been developed which we shortly introduce in the following.

10.2.1 Canonical (CP) Format

Any tensor $A \in \mathbb{R}^{\mathscr{I}}$ can be represented as the finite sum of *elementary tensors* $u = u_1 \otimes \dots \otimes u_d \in \mathbb{R}^{\mathscr{I}}$ with $u_\mu \in \mathbb{R}^{\mathscr{I}_\mu}$ for all $\mu = 1, \dots, d$. This motivates the following definition.

Definition 10.1 (canonical format, tensor rank). Let $r \in \mathbb{N}_0$. The subset $\mathscr{C}_r \subset \mathbb{R}^{\mathscr{I}}$ is defined by

$$\mathscr{C}_r := \left\{ \sum_{j=1}^{r} \bigotimes_{\mu=1}^{d} u_{\mu,j} : u_{\mu,j} \in \mathbb{R}^{\mathscr{I}_\mu},\ j = 1, \ldots, r,\ \mu = 1, \ldots, d \right\}.$$

A tensor $A \in \mathscr{C}_r$ is said to be represented in *canonical format* with (canonical) *representation rank* r. Given $A \in \mathbb{R}^{\mathscr{I}}$, the integer

$$\operatorname{rank}(A) := \min\{r \in \mathbb{N}_0 : A \in \mathscr{C}_r\}$$

is called the *tensor rank* of A.

The storage complexity of a tensor $A \in \mathscr{C}_r$ with $n := n_1 = \ldots = n_d$ is given by

$$N_{\text{storage}}(\mathscr{C}_r) = \mathscr{O}(drn),$$

which remains moderate even for large d provided that r is small.

In general, the determination of the tensor rank of a given tensor $A \in \mathbb{R}^{\mathscr{I}}$ is an NP-hard problem (cf. [24]). Moreover, the set $\mathscr{C}_r$ is not closed for $d \geq 3$ and $r \geq 2$ (cf. [14]). This may lead to severe difficulties for the usage of the canonical format in approximation problems. For a discussion on regularisation techniques and local optimisation algorithms in the canonical format we refer the reader to [15, 27] and the references therein.

10.2.2 Tucker Format

In linear algebra, matrices are well-understood objects which can be characterised by a number of important properties like, e.g., the matrix rank. It turns out that the interpretation of tensors as matrices leads to powerful tools that can be applied even in high dimensions.

Definition 10.2 (matricisation). Let $D := \{1, \ldots, d\}$. Given a subset $t \subset D$ with complement $[t] := D \setminus t$, the *matricisation*

$$\mathscr{M}_t : \mathbb{R}^{\mathscr{I}} \to \mathbb{R}^{\mathscr{I}_t} \otimes \mathbb{R}^{\mathscr{I}_{[t]}}, \qquad \mathscr{I}_t := \times_{\mu \in t} \mathscr{I}_\mu,\ \mathscr{I}_{[t]} := \times_{\mu \in [t]} \mathscr{I}_\mu,$$

of a tensor $A \in \mathbb{R}^{\mathscr{I}}$ is defined by its entries

$$\mathscr{M}_t(A)_{(i_\mu)_{\mu \in t}, (i_\mu)_{\mu \in [t]}} := A_{(i_1, \ldots, i_d)}, \qquad i_\mu \in \mathscr{I}_\mu,\ \mu \in D.$$

If $t = \{\mu\}$ for $\mu \in D$, we shortly write $\mathscr{M}_\mu$ instead of $\mathscr{M}_{\{\mu\}}$.

Definition 10.3 (Tucker format, Tucker rank). Let $k = (k_1, \ldots, k_d) \in \mathbb{N}_0^d$. The subset $\mathscr{T}_k \subset \mathbb{R}^{\mathscr{I}}$ is defined by

$$\mathscr{T}_k := \left\{A \in \mathbb{R}^{\mathscr{I}} : \operatorname{rank}\left(\mathscr{M}_\mu(A)\right) \le k_\mu,\ \mu = 1, \ldots, d\right\}.$$

A tensor $A \in \mathscr{T}_k$ is said to be represented in *Tucker format* with *Tucker representation rank* k. Given $A \in \mathbb{R}^{\mathscr{I}}$, *the* Tucker rank $k := (k_1, \ldots, k_d) \in \mathbb{N}_0^d$ of A is defined by

$$k_\mu = \operatorname{rank}\left(\mathscr{M}_\mu(A)\right), \quad \mu = 1, \ldots, d.$$

The Tucker format has been introduced for $d = 3$ in [42] in the context of psychometrics. An explicit representation of a tensor $A \in \mathscr{T}_k$ is often given in the form

$$A = \sum_{j_1=1}^{k_1} \cdots \sum_{j_d=1}^{k_d} C_{(j_1,\ldots,j_d)} \bigotimes_{\mu=1}^{d} u_{\mu,j_\mu}$$

with a *core tensor* $C \in \mathbb{R}^{k_1 \times \ldots \times k_d}$ and $u_{\mu,j_\mu} \in \mathbb{R}^{\mathscr{I}_\mu}$, $j_\mu = 1, \ldots, k_\mu$, $\mu = 1, \ldots, d$. This leads to a storage complexity of

$$N_{\text{storage}}(\mathscr{T}_k) = \mathscr{O}\left(k^d + dnk\right)$$

with $n_\mu = n$ and $k_\mu \le k$ for all $\mu = 1, \ldots, d$.

For a given tensor $A \in \mathbb{R}^{\mathscr{I}}$ one can compute its Tucker rank by standard linear algebraic tools. Moreover, a truncation procedure is available [13] that computes a quasi-best approximation $A_{\text{HOSVD}} \in \mathscr{T}_k$ in the sense

$$\|A - A_{\text{HOSVD}}\|_2 \le \sqrt{d} \min_{A_{\text{best}} \in \mathscr{T}_k} \|A - A_{\text{best}}\|_2.$$

10.2.3 Hierarchical Format

In the Tucker format, we have considered matricisations of tensors with respect to subsets $t = \{\mu\} \in D$. In order to allow for a structured and data-sparse representation of tensors, more general subsets $t \subset D$ which can be organised as a binary tree are of special interest.

Definition 10.4 (dimension tree). Let $D := \{1, \ldots, d\}$. A tree T_D is called a *dimension tree* if the following three conditions hold:

(a) The index set D is the root of the tree T_D,
(b) All vertices $t \in T_D$ are non-empty subsets $t \subset D$,
(c) Every vertex $t \in T_D$ with $\#t \ge 2$ has two sons $t_1, t_2 \in T_D$ with the property

$$t = t_1 \cup t_2, \qquad t_1 \cap t_2 = \emptyset.$$

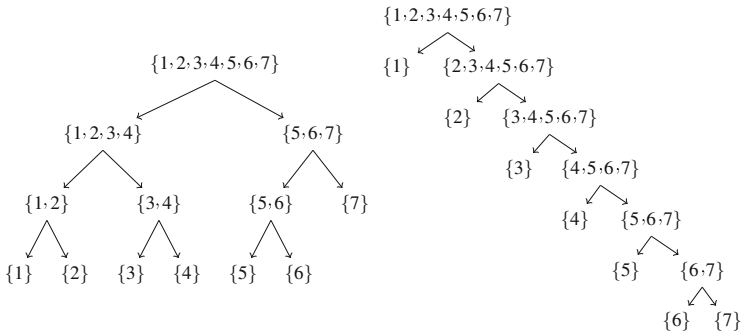


Fig. 10.1 *Left*: Balanced binary tree. *Right*: Linear TT tree

The set of leaves of T_D is defined by $\mathscr{L}(T_D) := \{t \in T_D : \#t = 1\}$. For all $t \in T_D \setminus \mathscr{L}(T_D)$, we denote the set of sons of t by $\mathrm{sons}(t)$.

Example 10.1. (a) In a balanced binary tree T_D, each node $t \in T_D \setminus \mathscr{L}(T_D)$ with $t = \{\mu_1, \ldots, \mu_q\} \subset D$, $q > 1$, has two sons $t_1, t_2 \in T_D$ of the form

$$t_1 = \{\mu_1, \ldots, \mu_r\}, \qquad t_2 = \{\mu_{r+1}, \ldots, \mu_q\}, \qquad r := \lceil q/2 \rceil.$$

An example for $d = 7$ is depicted in Fig. 10.1. The balanced tree is of minimal depth $\lceil \log_2 d \rceil$.

(b) In the *TT format* [33] (or *MPS* representation [43, 44]), the dimension tree is a simple linear tree, where all nodes $t \in T_D$ are of the form

$$t = \{q\} \quad \text{or} \quad t = \{q, \ldots, d\}, \qquad q = 1, \ldots, d.$$

An example for $d = 7$ is depicted in Fig. 10.1. The TT tree is of maximal depth $d - 1$.

Based on the concept of the matricisation of tensors and the definition of a dimension tree, we can now introduce the hierarchical tensor format.

Definition 10.5 (hierarchical format, hierarchical rank). Let T_D be a dimension tree and let $k := (k_t)_{t \in T_D} \in \mathbb{N}_0^{T_D}$. The subset $\mathscr{H}_k \subset \mathbb{R}^{\mathscr{I}}$ is defined by

$$\mathscr{H}_k := \{A \in \mathbb{R}^{\mathscr{I}} : \mathrm{rank}(\mathscr{M}_t(A)) \leq k_t,\ t \in T_D\}.$$

A tensor $A \in \mathscr{H}_k$ is said to be represented in *hierarchical format* with *hierarchical representation rank* k. Given $A \in \mathbb{R}^{\mathscr{I}}$, *the* hierarchical rank $k := (k_t)_{t \in T_D} \in \mathbb{N}_0^{T_D}$ of A is defined by

$$k_t = \mathrm{rank}\left(\mathscr{M}_t(A)\right), \quad t \in T_D.$$

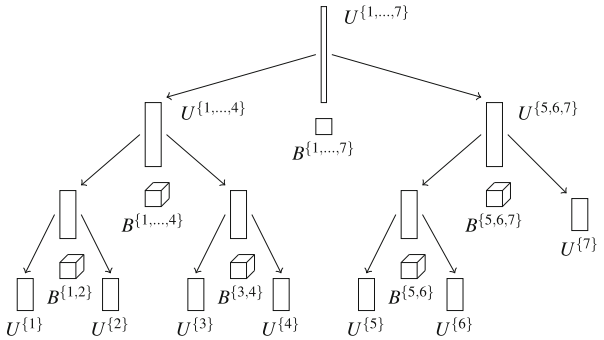


Fig. 10.2 Hierarchical tensor format with a balanced dimension tree for $d = 7$ with the nestedness property (10.2)

Given a tensor $A \in \mathscr{H}_k$, the subspaces $\mathscr{U}_t := \operatorname{image}(\mathscr{M}_t(A)) \subset \mathbb{R}^{\mathscr{I}_t}$, $t \in T_D$, fulfil the so-called *nestedness property*

$$\mathscr{U}_t \subset \mathscr{U}_{t_1} \otimes \mathscr{U}_{t_2}, \qquad t \in T_D \setminus \mathscr{L}(T_D),\ \operatorname{sons}(t) = \{t_1, t_2\}. \tag{10.1}$$

This allows for a recursive representation of A by the relation

$$U^t_{:,j} = \sum_{j_1=1}^{k_{t_1}} \sum_{j_2=1}^{k_{t_2}} B^t_{j,j_1,j_2} U^{t_1}_{:,j_1} \otimes U^{t_2}_{:,j_2}, \quad j = 1, \dots, k_t, \tag{10.2}$$

for all $t \in T_D \setminus \mathscr{L}(T_D)$ with $\operatorname{sons}(t) = \{t_1, t_2\}$ where $B^t \in \mathbb{R}^{k_t \times k_{t_1} \times k_{t_2}}$ and $U^t \in \mathbb{R}^{\mathscr{I}_t \times k_t}$ such that $\mathscr{M}_D(A) = U^D_{:,1}$. The nestedness property is illustrated for a balanced dimension tree in Fig. 10.2.

Due to (10.2), one only needs to store the matrices $U^t \in \mathbb{R}^{\mathscr{I}_\mu \times k_t}$ in the leaves $t = \{\mu\} \in \mathscr{L}(T_D)$ and the *transfer tensors* $B^t \in \mathbb{R}^{k_t \times k_{t_1} \times k_{t_2}}$ for all inner nodes $t \in T_D \setminus \mathscr{L}(T_D)$ in order to represent a tensor in $\mathscr{H}_k$. The storage complexity for this representation then sums up to

$$N_{\text{storage}}(\mathscr{H}_k) = \mathscr{O}\left(dk^3 + dnk\right).$$

with $n_\mu = n$ for all $\mu = 1, \dots, d$ and $k_t \le k$ for all $t \in T_D$.

Remark 10.1. In the TT format, for each node $t \in T_D \setminus \mathscr{L}(T_D)$ with $\operatorname{sons}(t) = \{t_1, t_2\}$ one either has $t_1 \in \mathscr{L}(T_D)$ or $t_2 \in \mathscr{L}(T_D)$. Hence, either $k_{t_1} \le n$ or $k_{t_2} \le n$ which leads to the bound $\mathscr{O}\left(dnk^2\right)$. An explicit representation of a tensor $A \in \mathbb{R}^{\mathscr{I}}$ in the TT format is often given in the form

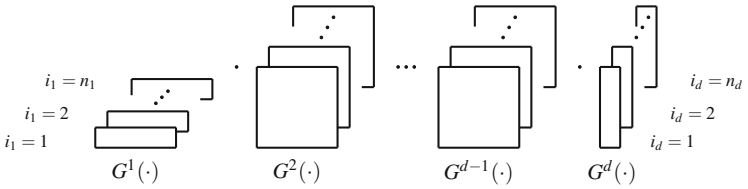


Fig. 10.3 TT format with matrices $G^\mu(i_\mu)$

$$A_{(i_1,\dots,i_d)} = \prod_{\mu=1}^{d} G^\mu(i_\mu)$$

with matrices $G^\mu(i_\mu)$ of matching sizes defined by

$$G^\mu(i_\mu)_{j,\ell} := \sum_{\nu=1}^{k_\mu} B^{\{\mu,\dots,d\}}_{j,\nu,\ell} U^{\{\mu\}}_{i_\mu,\nu}, \quad \mu = 1,\dots,d-1, \quad G^d(i_d)_{j,1} := U^{\{d\}}_{i_d,j}.$$

A visualisation of the TT format can be found in Fig. 10.3.

Similar to the Tucker format, the hierarchical rank of a tensor $A \in \mathbb{R}^{\mathcal{I}}$ can be computed by standard linear algebraic tools. Moreover, we have developed a truncation procedure [18] that computes an approximation of the best approximation of a tensor $A \in \mathbb{R}^{\mathcal{I}}$ in $\mathscr{H}_k$. This hierarchical singular value decomposition ($\mathscr{H}$-SVD) yields a tensor $A_{\mathscr{H}\text{-SVD}} \in \mathscr{H}_k$ with the property that

$$\|A - A_{\mathscr{H}\text{-SVD}}\|_2 \le \sqrt{2d-3} \min_{A_{\text{best}} \in \mathscr{H}_k} \|A - A_{\text{best}}\|_2.$$

If the input tensor A is already given in hierarchical format, i.e. $A \in \mathscr{H}_k$, the $\mathscr{H}$-SVD can be computed in

$$\mathscr{O}(dk^4 + dnk^2).$$

Similar results have been obtained for the TT format in [33].

10.3 Adaptive Tensor Sampling

In many applications, a tensor $A \in \mathbb{R}^{\mathcal{I}}$ is not already given in a data-sparse representation and one only knows how to evaluate certain entries A_i for $i \in \mathcal{I}$. In this section, we assume that the set $\Omega \subset \mathcal{I}$ of available entries can freely be chosen by the user. In a first scenario, the user is able to permanently communicate with

the application which means that the set Ω can be constructed adaptively during the approximation process. In the second scenario (Sect. 10.4), the user communicates with the application only once and gives a predetermined set Ω to the application which returns the corresponding entries A_i, $i \in \Omega$, to the user.

In both cases, we assume that there exists an approximation of A in $\mathscr{H}_k$ which is unknown to us and which we would like to reconstruct. Since the set $\mathscr{H}_k$ is characterised by certain rank bounds on the matricisations of A, it is a natural idea to try to approximate the matricisations by low-rank matrices. In [2, 7], we have developed an adaptive approximation scheme which relies on a thorough combination of cross approximation techniques with adaptive sampling strategies.

10.3.1 Cross Approximation

Let T_D be a dimension tree and let $t \in T_D$ with complement $[t] := D \setminus t$. The approximation of a matrix $M \in \mathbb{R}^{\mathscr{I}_t \times \mathscr{I}_{[t]}}$ by the outer product of particular rows and columns of M has been analysed in [17].

Theorem 10.1 ([17]). *Let $M \in \mathbb{R}^{\mathscr{I}_t \times \mathscr{I}_{[t]}}$. If there exists a matrix $R \in \mathbb{R}^{\mathscr{I}_t \times \mathscr{I}_{[t]}}$ with the properties $\|M - R\|_2 \leq \varepsilon$ and $\operatorname{rank}(R) \leq k$ then there exist a subset $\mathscr{P}_t \subset \mathscr{I}_t$ of row indices and a subset $\mathscr{Q}_t \subset \mathscr{I}_{[t]}$ of column indices with $\#\mathscr{P}_t = \#\mathscr{Q}_t = k$ and a matrix $S_t \in \mathbb{R}^{\mathscr{P}_t \times \mathscr{Q}_t}$ such that*

$$\tilde{M} := M|_{\mathscr{I}_t \times \mathscr{Q}_t} \cdot S_t^{-1} \cdot M|_{\mathscr{P}_t \times \mathscr{I}_{[t]}}$$

approximates M with an error of

$$\|M - \tilde{M}\|_2 \leq \varepsilon \left(1 + 2\sqrt{k}\left(\sqrt{\#\mathscr{I}_t} + \sqrt{\#\mathscr{I}_{[t]}}\right)\right).$$

A practical construction which is based on successive rank-1 approximations has been introduced in [9]. The idea is to construct rank-1 approximations of the remainder. For initial pivot elements $\boldsymbol{p}_1 \in \mathscr{I}_t$, $\boldsymbol{q}_1 \in \mathscr{I}_{[t]}$, one defines

$$X_t^1 := M|_{\mathscr{I}_t \times \{\boldsymbol{q}_1\}} M^{-1}_{(\boldsymbol{p}_1, \boldsymbol{q}_1)} M|_{\{\boldsymbol{p}_1\} \times \mathscr{I}_{[t]}}.$$

For a given approximation X_t^{j-1} and pivots $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_j \in \mathscr{I}_t$, $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_j \in \mathscr{I}_{[t]}$, the next approximation X_t^j is defined by

$$X_t^j := X_t^{j-1} + R|_{\mathscr{I}_t \times \{\boldsymbol{q}_j\}} R^{-1}_{(\boldsymbol{p}_j, \boldsymbol{q}_j)} R|_{\{\boldsymbol{p}_j\} \times \mathscr{I}_{[t]}}, \qquad R := M - X_t^{j-1}, \tag{10.3}$$

for all $j = 2, \ldots, k$. The final approximation is then given by $\tilde{M} := X_t^k$. Using the notation of Theorem 10.1, we have that

$$S_t = M|_{\mathscr{P}_t \times \mathscr{Q}_t}, \quad \mathscr{P}_t := \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k\}, \quad \mathscr{Q}_t := \{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_k\}.$$

The construction (10.3) is adaptive in the sense that the absolute value of the pivot element $R_{(\boldsymbol{p}_j, \boldsymbol{q}_j)}$ gives an estimate for the norm $\|M - X_t^{j-1}\|_\infty$. The number of pivots k_t can therefore be chosen adaptively in order to reach some (heuristic) target accuracy ε.

10.3.2 Nested Approximation

Due to the nestedness property (10.1), the approximations of the matricisations $\mathscr{M}_t(A)$ for all $t \in T_D$ are not completely independent. This property can be ensured by an a priori restriction on the choice of the pivot elements traversing the tree T_D from the root to the leaves. Let $f \in T_D \setminus \mathscr{L}(T_D)$ with $\mathrm{sons}(f) = \{s, t\}$ and let $\mathscr{Q}_f \subset \mathscr{I}_{[f]}$ be known column pivots for the father f of t. Then we seek pivots $(\boldsymbol{p}_j, \boldsymbol{q}_j)$ for t such that

$$\boldsymbol{p}_j \in \mathscr{P}_t \subset \mathscr{I}_t, \qquad \boldsymbol{q}_j \in \mathscr{Q}_t \subset \mathscr{I}_s \times \mathscr{Q}_f \subset \mathscr{I}_{[t]}. \tag{10.4}$$

In [7] we prove that by this restriction the matrices S_t and the index sets $\mathscr{P}_t$, $\mathscr{Q}_t$ can be constructed for all $t \in T_D$ in a recursive way such that the final approximation $\tilde{A} \approx A$ fulfils $\tilde{A} \in \mathscr{H}_k$. The setup of $\tilde{A} \in \mathscr{H}_k$ requires only

$$N_{\mathrm{setup}} = \mathscr{O}\left(dk^4 + \mathrm{depth}(T_D)k^2 \sum_{\mu=1}^{d} n_\mu\right), \qquad k := \max_{t \in T_D} k_t,$$

operations. Moreover, we show that if

$$\mathrm{rank}(S_t) = \mathrm{rank}(\mathscr{M}_t(A)), \quad t \in T_D, \tag{10.5}$$

then $\tilde{A}$ is a reconstruction of the original tensor A. The number of entries required from A crucially depends on *how* the index sets $\mathscr{P}_t, \mathscr{Q}_t$ are chosen.

10.3.3 Adaptive Sampling Strategies

The pivot sets $\mathscr{P}_t, \mathscr{Q}_t$ are best chosen such that S_t has maximal volume. Since this is practically impossible, we suggest to seek for large entries in modulus in the remainder $R := M - X_t^{j-1}$ in the incremental construction (10.3). As the index sets $\mathscr{I}_t$ and $\mathscr{I}_{[t]}$ are typically very large, the remainder cannot be formed explicitly. However, one can compute entries of R directly from the low-rank representation (10.3).

A quite successful strategy proposed in [16] for the canonical format aims at finding large entries on *fibres* $R_{(i_1,\dots,i_{\mu-1},\cdot,i_{\mu+1},\dots,i_d)} \in \mathbb{R}^{\mathscr{I}_\mu}$. In a greedy approach, one can alternate through the directions $\mu \in D$ such that in each step the quality of the pivot element increases. Using this construction, the overall number of entries required from A is given by

$$N_{\text{entries}}^{\text{fibres}} = \mathscr{O}\left(dk^3 + \operatorname{depth}(T_D)k^2 \sum_{\mu=1}^{d} n_\mu\right), \qquad k := \max_{t\in T_D} k_t.$$

Alternatively, we suggested in [6] to (randomly) select small index sets $\mathscr{J}_t \subset \mathscr{I}_t$, $\mathscr{J}_{[t]} \subset \mathscr{I}_{[t]}$ such that we may consider explicit cross approximations of the submatrix $\hat{M} := M|_{\mathscr{J}_t \times \mathscr{J}_{[t]}}$. Assuming $\#\mathscr{J}_t, \#\mathscr{J}_{[t]} \le K$, the construction of the pivot index sets $\mathscr{P}_t \subset \mathscr{J}_t$, $\mathscr{Q}_t \subset \mathscr{J}_{[t]}$ requires the evaluation of $\mathscr{O}(k_t K)$ entries of A. The overall number of entries required from A is then given by

$$N_{\text{entries}}^{\text{submatrix}} = \mathscr{O}\left(dk^3 + dkK + k\sum_{\mu=1}^{d} n_\mu\right), \qquad k := \max_{t\in T_D} k_t.$$

Provided that $K = \mathscr{O}(k^2)$, this is slightly less than in the fibre-based strategy.

Example 10.2 ([5]). Let $S = (0,1)^2$ with circular inclusions $S_\mu \subset S$ as depicted in Fig. 10.4 (left). Consider the parametric PDE problem

$$-\operatorname{div}(a\nabla u) = 1 \text{ in } S, \quad u|_{\partial S} = 0,$$

with diffusion coefficient $a|_{S_\mu} = p_\mu$, $\mu = 1,\dots,d$, and $a = 1$ elsewhere for parameters $p = (p_1,\dots,p_d) \in P = [\frac{1}{2}, 2]^d$. Let $\phi(p) := \int_S u(x,p)\,dx$ be an

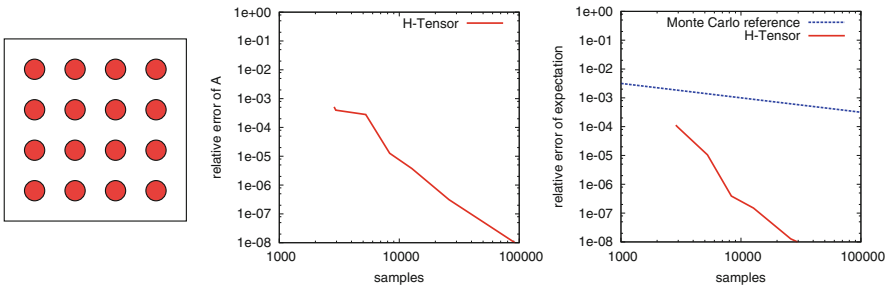


Fig. 10.4 Parametric diffusion problem from [5] for $d = 16$. *Left*: domain $S = (0,1)^2$ with circular inclusions $S_1,\dots,S_d$. *Middle*: relative error of an approximation of $A \in \mathbb{R}^{\mathscr{I}}$, $\mathscr{I} := \{1,\dots,10\}^d$, with $A_i := \phi(p_i)$ in $\mathscr{H}_k$ constructed by an adaptive sampling strategy based on random submatrices. *Right*: relative error of expectation $\mathbb{E}[\phi]$ for a uniform distribution of p in P

associated quantity of interest. A discretisation of ϕ on a tensor grid $\{p_i \in P\}_{i \in \mathcal{I}}$ defines a tensor $A \in \mathbb{R}^{\mathcal{I}}$ by $A_i := \phi(p_i)$. Figure 10.4 (middle) shows the attainable accuracy of an approximation of A in $\mathscr{H}_k$ by an adaptive sampling strategy. In an a posteriori step, one can then compute approximations of the expected value $\mathbb{E}[\phi]$ for a given distribution of p in P, cf. Fig. 10.4 (right).

10.3.4 Tree Adaptivity

Up to now, we have considered the approximation of tensors $A \in \mathbb{R}^{\mathcal{I}}$ in the hierarchical format given a fixed dimension tree T_D. In [19] we have shown that the hierarchical rank of a tensor A may be very sensitive to the choice of the tree. Therefore, in general a suitable tree needs to be constructed in dependence on the given input data. In principle, one needs to solve the following minimisation problem.

Problem 10.1. Let $A \in \mathbb{R}^{\mathcal{I}}$ and let $k_t := \operatorname{rank}(\mathscr{M}_t(A))$ for all $t \subset D$. Among all possible dimension trees T_D find a minimiser of

$$\sum_{\substack{t \in T_D \setminus \mathscr{L}(T_D) \\ \operatorname{sons}(t) = \{t_1, t_2\}}} k_t k_{t_1} k_{t_2}.$$

Unfortunately, the number of possible trees scales exponentially in the dimension d which makes a global optimisation of the tree structure too expensive. In [6] we address this problem by an agglomerative strategy that constructs a suitable dimension tree from the leaves up to the root. In each agglomerative step we aim at joining strongly coupled subsets of the directions $1, \ldots, d$.

The separability of a subset $t \subset D$ from $[t] := D \setminus t$ can be measured in terms of the rank k_t. However, it is not directly obvious whether or not it makes sense to join two disjoint subsets $t_1, t_2 \subset D$ to a new subset $t := t_1 \cup t_2$. In our agglomerative strategy, we suggest to use the ratio $k_t/(k_{t_1} k_{t_2})$ as a cluster criterion. Since $k_t = \dim \mathscr{U}_t$, $\mathscr{U}_t := \operatorname{image} \mathscr{M}_t(A)$, this ratio gives an indication how well the nestedness property (10.1) at a node $t \in T_D$ is fulfilled. It can therefore be used in an iterative construction of a dimension tree with a (small) polynomial dependence on d.

10.4 Non-adaptive Tensor Sampling

If the user can interact with the application only once, all pivot elements from A have to be fixed a priori. In this case, one can still use the second approach from above by a minor modification [26]. Descending from the root of T_D down to the leaves, one

may first fix (randomly chosen) pivot sets $\mathscr{P}_t \subset \mathscr{I}_t$, $\mathscr{Q}_t \subset \mathscr{I}_{[t]}$ of size $\#\mathscr{P}_t, \#\mathscr{Q}_t \leq K$ that additionally fulfil the nestedness condition (10.4) for all $t \in T_D$. Using these pivot sets, one can directly construct an approximation of A in $\mathscr{H}_k$ with $k_t \leq K$ for all $t \in T_D$. The overall number of entries required from A is then given by

$$N_{\text{entries}}^{\text{fixed}} = \mathscr{O}\left(dK^3 + K\sum_{\mu=1}^{d} n_\mu\right).$$

If the selected pivot sets fulfil Eq. (10.5) the reconstruction of the original tensor is successful.

10.5 Tensor Completion

The challenge of tensor completion is to find a low-rank tensor $A^{\mathscr{F}}$ in a certain tensor format $\mathscr{F}$ to a set of a priori given entries $\{A_i \in \mathbb{R} : i \in \Omega\}$ such that

$$A^{\mathscr{F}} = \operatorname*{argmin}_{\tilde{A}^{\mathscr{F}} \in \mathscr{F}} \left\| A - \tilde{A}^{\mathscr{F}} \right\|_\Omega \text{ with } \|A - Y\|_\Omega^2 := \sum_{i \in \Omega} (A_i - Y_i)^2 .$$

This task is entirely different from the two scenarios considered in the previous sections since the tensor entries cannot be freely chosen by the user. Due to this fact, completion problems are often harder to handle. In order to allow a low-rank approximation in the hierarchical format we require a certain slice density of the samples measured by the oversampling factor

$$c_{\text{ov}}(j_\mu) := \sharp\{i \in \Omega : i_\mu = j_\mu\} \geq C, \qquad j_\mu \in \mathscr{I}_\mu.$$

In practice, the required oversampling factor depends on the original tensor A and the desired target accuracy of the approximation. If $c_{\text{ov}}(j_\mu) = 0$ for any $j_\mu \in \mathscr{I}_\mu$, then the tensor slice $A|_{i_\mu = j_\mu} \in \mathbb{R}^{\mathscr{I}_1 \times \ldots \times \mathscr{I}_{\mu-1} \times \mathscr{I}_{\mu+1} \times \ldots \times \mathscr{I}_d}$ is not determined and could by completed arbitrarily.

The problem of tensor completion has been studied for the canonical tensor format in [1, 10, 12, 30, 41] and for the Tucker format in [28, 32, 35, 38]. In the hierarchical format a promising manifold optimisation strategy is pursued in [39]. In [20] we consider the tensor completion problem in the hierarchical format with respect to a given TT tree (cf. Example 10.1).

With the representation system $G = (G^1, \ldots, G^d)$ from Remark 10.1, the task is to find

$$G = \operatorname*{argmin}_{\tilde{G}} \left\| A - A^{\tilde{G}} \right\|_\Omega .$$

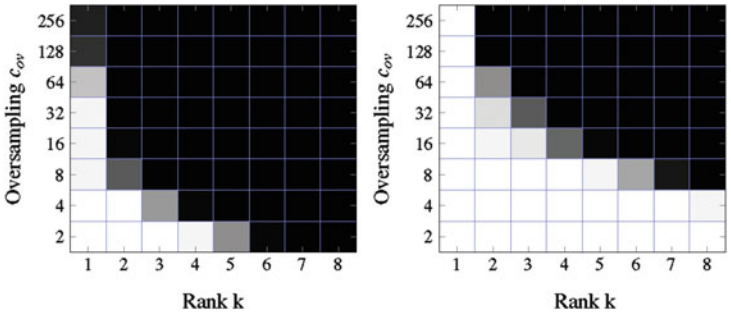


Fig. 10.5 Random TT-tensor of rank k with $n = 20$, where all entries of the representation system are uniformly distributed. The completion is successful if the relative error in the test set is $<10^{-3}$. Every combination of rank and oversampling factor was tested 20 times. The success is depicted by the coloured scale from *white* (0 of 20) to *black* (20 of 20). *Left*: $d = 4$. *Right*: $d = 5$. [20]

If the norm were the standard Euclidean norm of the tensor, then this minimisation problem could be treated by an alternating minimisation of the cores G^μ, each being just a linear least squares problem. In order to reduce the problem to this type, an augmentation by an auxiliary tensor Z is necessary:

$$(G, Z) = \operatorname*{argmin}_{\tilde{G},\tilde{Z}} \left\| \tilde{Z} - A^{\tilde{G}} \right\|_2 \text{ with } \tilde{Z}_i = A_i \ \forall i \in \Omega.$$

For this the first order optimality conditions lead to a system of equations that can be solved by a (block) successive over relaxation (SOR) type iterative method, the so-called alternating direction fitting (ADF) algorithm [20]. Whereas the convergence of the adaptive sampling strategies with respect to the number of samples was quasi-optimal, this is completely different for the tensor completion in this section.

Example 10.3 ([20]). We consider a random generated TT-tensor A of exact rank k and $n = 20$, where the entries of the representation system $G^\mu(i_\mu)_{j,\ell}$ are uniformly distributed in $[-0.5, 0.5]$ for all $\mu = 1, \ldots, d$. The completion is successful if the relative error for a random generated test set Ω_T is smaller than 10^{-3} with $\sharp\Omega_T = \sharp\Omega$. In Fig. 10.5 a statistic of the completion success is shown where the algorithm runs 20 times for every combination of rank k and oversampling factor c_{ov} with a coloured scale from white (0 of 20) to black (20 of 20).

It turns out that the necessary oversampling factor depends on both the rank as well as the order of the tensor. Moreover, the iterative minimisation is considerably slower than the direct reconstruction via cross approximation, so that one should always use a sampling rule when this is possible.

References

1. Acar, E., Dunlavy, D.M., Kolda, T.G., Mørup, M.: Scalable tensor factorizations for incomplete data. Chemom. Intell. Lab. Syst. **106**(1), 41–56 (2011). doi:10.1016/j.chemolab.2010.08.004
2. Ballani, J.: Fast evaluation of near-field boundary integrals using tensor approximations. Ph.D. thesis, Universität Leipzig (2012)
3. Ballani, J.: Fast evaluation of singular BEM integrals based on tensor approximations. Numer. Math. **121**(3), 433–460 (2012)
4. Ballani, J., Grasedyck, L.: A projection method to solve linear systems in tensor format. Numer. Linear Algebra Appl. **20**(1), 27–43 (2013)
5. Ballani, J., Grasedyck, L.: Hierarchical tensor approximation of output quantities of parameter-dependent PDEs. Preprint 385, IGPM, RWTH-Aachen, www.igpm.rwth-aachen.de (2013)
6. Ballani, J., Grasedyck, L.: Tree adaptive approximation in the hierarchical tensor format. SIAM J. Sci. Comput. **36**(4), A1415–A1431 (2014)
7. Ballani, J., Grasedyck, L., Kluge, M.: Black box approximation of tensors in hierarchical Tucker format. Linear Algebra Appl. **438**(2), 639–657 (2013)
8. Ballani, J., Meszmer, P.: Tensor structured evaluation of singular volume integrals. Comput. Vis. Sci. **15**(2), 75–86 (2012)
9. Bebendorf, M.: Approximation of boundary element matrices. Numer. Math. **86**(4), 565–589 (2000)
10. Beylkin, G., Garcke, J., Mohlenkamp, M.: Multivariate regression and machine learning with sums of separable functions. SIAM J. Sci. Comput. **31**, 1840–1857 (2009)
11. Caflisch, R.E.: Monte Carlo and quasi-Monte Carlo methods. Acta Numer. **7**, 1–49 (1998)
12. Dauwels, J., Garg, L., Earnest, A., Pang, L.: Tensor factorizations for missing data imputation in medical questionnaires. In: ICASSP 2012, Kyoto (2012)
13. De Lathauwer, L., Moor, B.D., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**(4), 1253–1278 (2000)
14. De Silva, V., Lim, L.H.: Tensor rank and the ill-posedness of the best low-rank approximation problem. SIAM J. Matrix Anal. Appl. **30**, 1084–1127 (2008)
15. Espig, M.: Effiziente bestapproximation mittels summen von elementartensoren in hohen dimensionen. Ph.D. thesis, Universität Leipzig (2008)
16. Espig, M., Grasedyck, L., Hackbusch, W.: Black box low tensor-rank approximation using fiber-crosses. Constr. Approx. **30**(3), 557–597 (2009). doi:10.1007/s00365-009-9076-9. http://www.mis.mpg.de/de/publications/preprints/2008/prepr2008-60.html
17. Goreinov, S.A., Tyrtyshnikov, E.E., Zamarashkin, N.L.: A theory of pseudoskeleton approximations. Linear Algebra Appl. **261**, 1–22 (1997)
18. Grasedyck, L.: Hierarchical singular value decomposition of tensors. SIAM J. Matrix Anal. Appl. **31**, 2029–2054 (2010)
19. Grasedyck, L., Hackbusch, W.: An introduction to hierarchical ($\mathscr{H}$-) rank and TT-rank of tensors with examples. Comput. Methods Appl. Math. **11**(3), 291–304 (2011). http://www.cmam.info/issues/CMAMv11p291-304.pdf
20. Grasedyck, L., Kluge, M., Krämer, S.: Alternating directions fitting (ADF) of hierarchical low rank tensors. Preprint 149, DFG SPP-1324 (2013)
21. Grasedyck, L., Kressner, D., Tobler, C.: A literature survey of low-rank tensor approximation techniques. GAMM-Mitteilungen **36**(1), 53–78 (2013)
22. Hackbusch, W.: Tensor Spaces and Numerical Tensor Calculus. Springer, Berlin (2012)
23. Hackbusch, W., Kühn, S.: A new scheme for the tensor representation. J. Fourier Anal. Appl. **15**(5), 706–722 (2009). doi:10.1007/s00041-009-9094-9. http://www.mis.mpg.de/de/publications/preprints/2009/prepr2009-2.html
24. Håstad, J.: Tensor rank is NP-complete. J. Algorithms **11**, 644–654 (1990)
25. Khoromskij, B.N., Oseledets, I.: Quantics-TT collocation approximation of parameter-dependent and stochastic elliptic PDEs. Comput. Methods Appl. Math. **10(4)**, 376–394 (2010)

26. Kluge, M.: Sampling rules for tensor reconstruction in hierarchical Tucker format. Preprint 162, DFG SPP-1324 (2014)
27. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**(3), 455–500 (2009). doi:10.1137/07070111X
28. Kressner, D., Steinlechner, M., Vandereycken, B.: Low-rank tensor completion by Riemannian optimization. Technical report, EPF Lausanne (2013). Technical report 19.2013
29. Kressner, D., Tobler, C.: Low-rank tensor Krylov subspace methods for parameterized linear systems. SIAM J. Matrix Anal. Appl. **32**(4), 1288–1316 (2011)
30. Krishnamurthy, A., Singh, A.: Low-rank matrix and tensor completion via adaptive sampling (2013). ArXiv:1304.4672
31. Litvinenko, A., Matthies, H.G., El-Moselhy, T.A.: Low-rank tensor approximation of the response surface. accepted by MCQMC (2013)
32. Liu, Y., Shang, F.: An efficient matrix factorization method for tensor completions. IEEE Signal Process. Lett. **20**(4), 307–310 (2013)
33. Oseledets, I.: Tensor-train decomposition. SIAM J. Sci. Comput. **33**(5), 2295–2317 (2011)
34. Oseledets, I.V., Tyrtyshnikov, E.E.: TT-cross approximation for multidimensional arrays. Linear Algebra Appl. **432**(1), 70–88 (2010)
35. Rauhut, H., Schneider, R., Stojanac, Z.: Low rank tensor recovery via iterative hard thresholding. In: SampTA 2013, 10th International Conference on Sampling Theory and Application, Jacobs University Bremen, Bremen (2013)
36. Savostyanov, D.: Quasioptimality of maximum-volume cross interpolation of tensors. Linear Algebra Appl. **458**, 217–244 (2014)
37. Savostyanov, D.V., Oseledets, I.V.: Fast adaptive interpolation of multi-dimensional arrays in tensor train format. In: 7th International Workshop on Multidimensional Systems (nDS). IEEE, University of Poitiers, Poitiers (2011)
38. Signoretto, M., Tran Dinh, Q., De Lathauwer, L., Suykens, J.: Learning with tensors: a framework based on convex optimization and spectral regularization. Mach. Learn. **94**(3), 303–351 (2014)
39. Silva, C.D., Herrmann, F.J.: Hierarchical Tucker tensor optimization – applications to tensor completion (2013). In: SampTA 2013, 10th International Conference on Sampling Theory and Application, Jacobs University Bremen, Bremen
40. Tobler, C.: Low-rank tensor methods for linear systems and eigenvalue problems. Ph.D. thesis, ETH Zürich (2012)
41. Tomasi, G., Bro, R.: PARAFAC and missing values. Chemom. Intell. Lab. **75**(2), 163–180 (2005)
42. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**, 279–311 (1966)
43. Vidal, G.: Efficient classical simulation of slightly entangled quantum computations. Phys. Rev. Lett. **91**(14) (2003)
44. White, S.R.: Density matrix formulation for quantum renormalization groups. Phys. Rev. Lett. **69**(19), 2863–2866 (1992)

Chapter 11
A Bond Order Dissection ANOVA Approach for Efficient Electronic Structure Calculations

Michael Griebel, Jan Hamaekers, and Frederik Heber

Abstract In this article, we present a new decomposition approach for the efficient approximate calculation of the electronic structure problem for molecules. It is based on a dimension-wise decomposition of the space the underlying Schrödinger equation lives in, i.e. $\mathbb{R}^{3(M+N)}$, where M is the number of nuclei and N is the number of electrons. This decomposition is similar to the ANOVA-approach (analysis of variance) which is well-known in statistics. It represents the energy as a finite sum of contributions which depend on the positions of single nuclei, of pairs of nuclei, of triples of nuclei, and so on. Under the assumption of locality of electronic wave functions, the higher order terms in this expansion decay rapidly and may therefore be omitted. Furthermore, additional terms are eliminated according to the bonding structure of the molecule. This way, only the calculation of the electronic structure of local parts, i.e. small subsystems of the overall system, is necessary to approximate the total ground state energy. To determine the required subsystems, we employ molecular graph theory combined with molecular bonding knowledge. In principle, the local electronic subproblems may be approximately evaluated with whatever technique is appropriate, e.g. HF, CC, CI, or DFT. From these local energies, the total energy of the overall system is then approximately put together in a telescoping sum like fashion. Thus, if the size of the local subproblems is independent of the size of the overall molecular system, linear scaling is directly obtained. We discuss the details of our new approach and apply it to both, various small test systems and interferon alpha as an example of a large biomolecule.

M. Griebel (✉) • F. Heber
Institute for Numerical Simulation, University of Bonn, Wegelerstr. 6, 53115 Bonn, Germany
e-mail: griebel@ins.uni-bonn.de; heber@ins.uni-bonn.de

J. Hamaekers
Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, 53754 Sankt Augustin, Germany
e-mail: jan.hamaekers@scai.fraunhofer.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_11

11.1 Introduction

The coupling of the micro- and the mesoscale of chemical reactions is currently a field of intensive research. Where the microscale is the realm of quantum mechanical effects, the mesoscale is described by statistical mechanics and macroscopic thermodynamics. Nevertheless, there are additional strong influences onto the mesoscale by effects from the microscale. Numerically, the microscale is usually treated with Hartree-Fock (HF), Configuration Interaction (CI), Coupled Cluster (CC), or Density Functional Theory (DFT) methods which yield approximate solutions to the underlying quantum-mechanical (QM) Schrödinger equation (SE), whereas the mesoscale is covered by classical molecular mechanics (MM) methods which use Newton's mechanics with empirically fitted potential functions.

The ultimate goal would be a seamless coupling of quantum mechanical computations where needed and classical molecular mechanics simulations where sufficient. Such approaches are generally referred to as multi-scale methods, an extensive overview is given in [28]. Any starting point must be the general Schrödinger equation for the electrons and nuclei of the system under consideration. The Schrödinger equation however lives in $3(M + N)$ dimensions, where M denotes the number of nuclei and N denotes the number of electrons. This renders a direct numerical treatment impossible due to the curse of dimension and one has to resort to model approximations. As a first step, in the Born-Oppenheimer molecular dynamics (MD) approach, the wave functions of the nuclei and electrons are separated, the subsystem of the nuclei is treated classically with Newton's mechanics and the remaining $3N$-dimensional electronic Schrödinger equation is further approximated by one of the aforementioned methods. The potential needed for Newton's mechanics is obtained from the electronic solution by the Hellmann-Feynman theorem. This way, QM and MM are globally coupled. However, a global electronic QM solution is, at least for larger molecules, still too expensive as conventional methods scale at best with $O(M^3)$ due to the underlying problem of matrix diagonalization.

Thus, general *linear scaling* electronic structure methods are employed to overcome the dimensionality problem. As a first step, for the long-range Coulomb interaction, the use of the fast multipole method [16] has resulted in $O(M \log M)$ complexity. Furthermore, a cutoff radius such as for the MP2 theory [5] was used in a Divide&Conquer approach to take advantage of the exponential decay properties of electronic wave functions. Altogether, this resulted in linear scaling [14]. Another common method is the Density Matrix Minimization technique [8, 27]. There, the density matrix is unconstrainedly minimized via a conjugate gradient scheme, using idempotency and normalization. The Fock matrix is the minimized output, after off-diagonal elements have also been truncated at a cutoff radius. The electronic localization in non-metallic systems can also be exploited for plane wave basis sets [34]. Again, a cutoff then allows for linear scaling. Note however that there is a crossover point up to which the standard cubic scaling approaches still perform faster due to smaller prefactors in their computational complexity counts [15].

In order to reduce the constants and thus to shift this crossover point, one tries to somehow further decompose the full global electronic Hamiltonian into local parts

and employs local QM there. Let us briefly summarize the most common *decomposition approaches* in the following. One of the first, the Force-Matching Method by Ercolessi [13], tries to automatically generate empirical potentials by a least-square fitting of the forces of ab-initio calculations to general many-body potential forms such as that of the Embedded Atom approach [11] or that of Abell-Tersoff [1, 36]. Then, there is a range of methods which employ a decomposition[1] directly in $\mathbb{R}^3$, like e.g. the SIBFA (Sum of Interactions Between Fragments computed Ab initio) procedure [17] and its generalization, the so-called Fragmentation Reconstruction Method (FRM) [2]. Further schemes are the so-called IMOMM ansatz proposed by Morokuma [30], the ONIOM approach [39] and the well-known Fragment Molecular Orbital (FMO) method [25]. A scheme for modeling the electrostatic impact of a passive MM environment on the active QM system is described in [26]. Moreover, in [3] and [33] an interface regime between QM and MM with "link" atoms is proposed to account for the cutting of bonds. Similar techniques are used in [37, 38]. The common basic idea of most approaches is to use a telescoping sum over two regions to describe the total energy, like e.g. Ω_1 and $\Omega_2 \subseteq \Omega_1$, where the energy is split as $E^{QM/MM} = E^{MM}_{\Omega_1} + E^{QM}_{\Omega_2} - E^{MM}_{\Omega_2}$, where to our knowledge all procedures involve stringent chemical knowledge to choose the regions (or cuts) as best as possible but to still keep the ground-state electronic density intact. Another approach divides *time* instead of space in order to generate a coupling between QM and MM. One of these methods is learn-on-the-fly [10], which is similar to the Force-Matching method, but is run during the computation: At intermittent time steps certain clusters of the simulation domain are locally computed by QM, and the obtained local forces are used to correct the MM calculation.

While all of the above methods have promising features, we feel that they generally either involve too many additional parameters, unchemically cut bonds in separating active from passive regions, or even worse, add unphysical pseudo-atoms in order to compensate for the different energy and time scales and to avoid spill-out effects of electronic density or energy. Moreover, they are plainly too simple or do not grasp the problem in its full complexity, since only a matching or interpolation with respect to energy or forces between the QM and MM parts of the overall approach is employed.

There is one more group of methods that build upon *additivity models*, well-known in chemistry, see [19] and references therein. The central idea is to construct molecular properties of a system by adding up the corresponding known properties of its fragments. The principal hope is that a high-dimensional system such as a complex molecule depends strongly only on few input variables. Rabitz et al. [19] describe a High-Dimensional Model Representation (HDMR) that can also be understood as an *ANalysis Of VAriance* (ANOVA), which is well-known from statistics. They address the problem of the estimation of the enthalpy of formation of a broad range of organic molecules based on experimental data, but they do not

[1] Ultimately, the aim would be a decomposition of $\mathbb{R}^{3(M+N)}$, the space where the full Schrödinger equation lives in.

assess the possibilities of the ansatz in the field of electronic structure calculations. Deev and Collins [9, 12] use this additivity model ansatz by calculating the total electronic energy of fragments of a system under consideration to obtain a good approximation of the energy of the total molecule. They do give an algorithmic description, however which we feel is not fully consistent with the mathematical basics, governed by the ANOVA or HDMR scheme.

In this article, we propose a more sophisticated algorithm. The additivity models place their hope on the same grounds as do many-body potential such as Tersoff's [36], where the energy and the forces of an atom are assumed to depend on its local coordination. Here, for a proof-of-concept, we concentrate on covalent bonding, hence on charge-neutral molecular systems and subsystems.[2] We will use this knowledge of coordination and bonds between nuclei to decompose the space $R^{3(M+N)}$ of the underlying Schrödinger equation in a dimension-wise fashion. This decomposition is similar to the ANOVA-approach. It represents the energy as a finite sum of contributions which depend on the positions of single nuclei, of pairs of nuclei, of triples of nuclei, and so on. Under the assumption of locality of electronic wave functions, the higher order terms in this expansion decay rapidly and may therefore be omitted. Furthermore, additional terms are eliminated according to the bonding structure of the molecule. This way, only the calculation of the electronic structure of local parts, i.e. small overlapping subsystems of the overall molecule system, is necessary to approximate the total ground state energy. To determine the required subsystems, we employ molecular graph theory combined with molecular bonding knowledge. Here, modern graph algorithms are used to create proper local subproblems as overlapping fragments of the overall molecular system. Furthermore, hydrogenization is used to close shells and saturate bonds that have been cut. We thus also exploit locality, however not by an explicit cutoff radius as most conventional methods do, but by implicitly using it in the inherent bond structure of the molecular system. In principle, the local electronic subproblems may be approximately evaluated with whatever QM technique is appropriate, e.g. HF, CI, CC, or DFT. From these local energies, the total energy of the overall system is then approximately put together in a telescoping sum like fashion. Thus, if the size of the local subproblems is independent of the size of the overall molecular system, linear scaling is directly obtained. The $3(M+N)$-dimensional full global Hamiltonian is broken down within the Born-Oppenheimer Approximation to $O(M)$ components, the i-th of them with $M_i^{(k)}$ degrees of freedom, with an upper bound $\max_i\{M_i^{(k)}\}$ controlled by a single parameter k which we name the *bond order* of the approximation. This ansatz specifically combines the smaller prefactor of the cubic scaling methods with a general linear scaling behavior. As the size of each subproblem depends on the bond coordination of the involved atoms, we coined the method BOSSANOVA (Bond Order diSSection ANOVA).

The remainder of this article is organized as follows: In Sect. 11.2 we briefly summarize the basics of the underlying Schrödinger equation. In Sect. 11.3 we describe the ANOVA-like decomposition of the energy of the Schrödinger equation

[2] Note however that our approach should work equally well also in the non-charge neutral case.

in the context of molecular graph theory. In Sect. 11.4 we give numerical results for a broad range of organic molecules. We end with some concluding remarks in the final section.

11.2 Schrödinger Equation in the Born-Oppenheimer Approximation

Let us consider a molecular system consisting of M nuclei and N electrons. Its time-dependent state function can be written in general as

$$\Psi = \Psi(R_1, \ldots, R_M, r_1, \ldots, r_N, t),$$

where R_i and r_j denote positions in three-dimensional space $\mathbb{R}^3$ associated to the ith nucleus and the jth electron, respectively. The variable t denotes the time-dependency of the state function. The vector space (space of configurations), in which the coordinates of the particles are given, is therefore of dimension $3(M+N)$. In the following we will abbreviate $(R_1, \ldots, R_M)$ and $(r_1, \ldots, r_N)$ with the shorter notation $\mathbf{R}$ and $\mathbf{r}$, respectively. Also, we assume that Ψ is normalized to $\int \Psi^*(\mathbf{R}, \mathbf{r}, t)\Psi(\mathbf{R}, \mathbf{r}, t) d\mathbf{R} d\mathbf{r} = 1$.

Nuclei and electrons are charged particles. The electrostatic potential (Coulomb potential) of a point charge is $1/r$ in atomic units, where r is the distance from the position of the charged particle. An electron moving in this potential possesses the potential energy $V(r) = -1/r$. Neglecting spin and relativistic interactions and assuming that no external forces act on the system, the Hamilton operator in position representation associated to the system of nuclei and electrons is given as the sum over the operators for the kinetic energy and the Coulomb potentials,

$$\begin{aligned} &H(N, M, Z_1, m_1, \ldots, Z_M, m_M; \mathbf{R}, \mathbf{r}) := \\ &\underbrace{-\frac{1}{2}\sum_{k=1}^{N} \Delta_{r_k} + \sum_{k<j}^{N} \frac{1}{\| r_k - r_j \|} - \sum_{k=1}^{N}\sum_{j=1}^{M} \frac{Z_j}{\| r_k - R_j \|} + \sum_{k<j}^{M} \frac{Z_k Z_j}{\| R_k - R_j \|}}_{H_e(N, M, Z_1, m_1, \ldots, Z_M, m_M; \mathbf{R}, \mathbf{r})} \\ &- \frac{1}{2}\sum_{k=1}^{M} \frac{1}{m_k} \Delta_{R_k}, \end{aligned} \tag{11.1}$$

where we use a semicolon to distinguish between parameters (i.e. the number M of atoms, the number N of electrons, the nuclei mass in atomic units m_j and the atomic number Z_j) and the degrees of freedom (i.e. the positions $\mathbf{R}$ and $\mathbf{r}$). Here, $\|r_k - r_j\|$ are the distances between electrons, $\|r_k - R_j\|$ are distances between electrons and nuclei and $\|R_k - R_j\|$ are distances between nuclei. We will omit parameters from this list if they are clear from the context. This will later especially be $N, M, Z_1, m_1, \ldots, Z_M, m_M$.

Now, a system of equations for the electronic and for the nuclei degrees of freedom is usually derived with the *Born-Oppenheimer approximation*. To this end, the large difference in masses between electrons and atomic nuclei is exploited to decouple the motion of the electrons from that of the nuclei.[3] Then, one assumes that the electrons adapt instantaneously to a change in the nuclear configuration and are thus always in the quantum mechanical ground state denoted by $\phi_0(\mathbf{R}(t);\mathbf{r})$, which is associated to the actual position of the nuclei $\mathbf{R}(t)$. Note that this allows us to write $H_e(\mathbf{R}(t);\mathbf{r})$ instead of $H_e(\mathbf{R}(t),\mathbf{r})$ since the movement of the nuclei during the adaptation of the electron positions is negligibly small in the sense of classical dynamics. This justifies to set $\Psi(\mathbf{R},\mathbf{r},t) \approx \Psi^{BO}(\mathbf{R},\mathbf{r},t) := \sum_{j=0}^{\infty} \chi_j(\mathbf{R},t)\phi_j(\mathbf{R};\mathbf{r})$, which allows to separate the fast from the slow variables. We then obtain the following set of equations:

$$M_k \ddot{R}_k(t) = -\nabla_{R_k} \underbrace{\min_{|\phi_0(\mathbf{R}(t);\cdot)|=1} \left\{ \int \phi_0^*(\mathbf{R}(t);\mathbf{r}) H_e(\mathbf{R}(t);\mathbf{r}) \phi_0(\mathbf{R}(t);\mathbf{r}) d\mathbf{r} \right\}}_{=: V_e^{BO}(\mathbf{R}(t))} \tag{11.2}$$

$$H_e(\mathbf{R}(t);\mathbf{r})\phi_0(\mathbf{R}(t);\mathbf{r}) = E_0(\mathbf{R}(t))\phi_0(\mathbf{R}(t);\mathbf{r}). \tag{11.3}$$

In the end, after time discretization, we have to perform in each time step the following tasks: First, we have to compute an approximate solution of the electronic Schrödinger equation in (11.3) for fixed positions $\mathbf{R}$ of the nuclei, then we have to compute from its solution the forces on the nuclei and finally we have to compute the positions of the nuclei at the next time step by e.g. a Verlet time step for Newton's equations of motion of the nuclei in (11.2). To this end, we use the *Hellmann-Feynman Theorem* to obtain the electronic forces

$$F_k(\mathbf{R}(t)) = -\nabla_{R_k} \int \phi_0^*(\mathbf{R}(t)) H_e(\mathbf{R}(t)) \phi_0(\mathbf{R}(t)) d\mathbf{r}$$

acting on the nuclei. Variants of this approach are the Ehrenfest molecular dynamics and the Car-Parrinello method. For details of the derivation, see [18] and the references cited therein.

11.3 ANOVA Decomposition Scheme

So far, the Born-Oppenheimer molecular dynamics was employed to split the full Schrödinger problem into two parts, i.e. a classical Newton's equation of motion for the nuclei, and, in each discretized time step, the electronic eigenproblem (11.3) which may approximately be solved by e.g. the Hartree Fock, Configuration

[3] The ratio of the velocity v_K of a nucleus to the velocity of an electron v_e is in general smaller than 10^{-2}.

Interaction, Coupled Cluster, or Density Functional method, see [31, 35]. However, such an overall approach is only feasible for small molecules due to the high complexity of any approximate solution method for the electronic problem. To overcome this difficulty, the aforementioned coupling techniques and linear scaling methods had been developed. They basically all exploit locality of the electronic wave function in one way or another to reduce the complexity of the electronic problem.[4,5]

In the following, we also resort to a certain locality of the electronic wave function. It is expressed in the bond structure of the molecular system. We decompose the overall electronic problem into small subproblems which then may be handled efficiently. To this end, we introduce an ANOVA decomposition scheme for the energy of a molecular system into local parts by means of the bond order of the nuclei in the system.

11.3.1 ANOVA Expansion

We will now define the energy function for a molecular system and its ANOVA series expansion. To this end, we consider a molecular system which consists of N electrons and M nuclei, each with coordinate vector $R_i \in \mathbb{R}^3$ and atomic number $Z_i \in \mathbb{N}$, $i \in \{1, \dots, M\}$. We restrict ourselves to charge-neutral systems, i.e. the number of electrons N is equal to $\sum_i^M Z_i$ for reasons of simplicity. Finally, we consider the systems only in their electronic ground state in the framework of the Born-Oppenheimer molecular dynamics. To this end, we separate the time-independent electronic Schrödinger equation as in (11.3) and define a total ground state energy function $E^M : (\mathbb{N} \times \mathbb{R}^3)^M \to \mathbb{R}$. It depends on the parameters that completely identify the system under consideration, namely the coordinates R_i and the atomic number Z_i of each nuclei with fixed and unique label $i \in \{1, \dots, M\}$, i.e.

$$E^M(\underbrace{(Z_1, R_1)}_{=:X_1}, \dots, \underbrace{(Z_M, R_M)}_{=:X_M}) :=$$

$$\min_{|\phi_0(\mathbf{R}(t);\cdot)|=1} \int \phi_0^*(\mathbf{R}(t);\mathbf{r}) H_e\big(N = \sum_{i=1}^{M} Z_i, X_1, \dots, X_M)\phi_0(\mathbf{R}(t);\mathbf{r}\big) d\mathbf{r}, \tag{11.4}$$

where we further simplify the notation by defining $X_i := (Z_i, R_i)$. I.e. X_i combines the atomic number and the coordinates of the nuclei i. Note that, due to the

[4]This excludes in general metallic systems, whose electrons may be delocalized due to a vanishing band gap.

[5]Furthermore, the notion of the locality of the wave function is important as it leads to the general chemical understanding of molecules from the general bond structure up to nucleophilic sites.

charge-neutrality condition $N = \sum_i^M Z_i$, the parameter N may now be eliminated from the parameter list of the Hamiltonian H.

Now we will decompose the function E^M in a multivariate telescoping sum, i.e. in a finite series expansion in the nucleic parameters, in a similar way as the ANOVA decomposition[6] [21]. This decomposition involves a splitting of the M-dimensional function into contributions which depend on the positions of single nuclei and associated charges, of pairs of nuclei and associated charges, of triples of nuclei and charges, and so on. To this end, we consider the subset of the nuclei parameters $\{X_i\}_{i\in I}$ described by a set of labels I with cardinality $|I| = k$ and call it the *molecular fragment* associated to I with size k. Note that we here do not need to consider the electronic degrees of freedom $\mathbf{r}$, as the system is assumed to be in ground state and, hence, the electronic state functions are all fixed by the minimum condition in (11.4).

First, we define the total electronic ground state energy of lower-dimensional subsystems of the molecular system under consideration, described by the set of indices $I = \{i_1, \ldots, i_k\}$,

$$E_{\{i_1,\ldots,i_k\}}(X_1, \ldots, X_k) := \min_{|\phi_0|=1} \int \phi_0^*(\mathbf{r}) H_e\Big(\sum_{j=1}^{k} Z_{i_j}, X_{i_1}, \ldots, X_{i_k}\Big)\phi_0(\mathbf{r})\Big) d\mathbf{r}. \tag{11.5}$$

Note that this is in form very similar to (11.4). In the notation of the electronic ground state wave functions ϕ_0, the dependency on $\mathbf{R}(t)$ was dropped as it is clear from the context.

Then, the energy function E^M is decomposed analogously to the ANOVA approach as

$$\begin{aligned} E^M(X_1, \ldots, X_M) = {} & F_\emptyset \\ & + \sum_{i_1}^{M} F_{\{i_1\}}(X_{\{i_1\}}) \\ & + \sum_{i_1<i_2}^{M} F_{\{i_1,i_2\}}(X_{\{i_1,i_2\}}) \end{aligned}$$

[6]The ANOVA decomposition of a M-dimensional function $f : [0,1]^M \to \mathbb{R}$ reads $f = \sum_{u\subseteq\{1,\ldots,M\}} f_u$ with f_u depending only on the variables indicated in u. The functions f_u satisfy the recurrence relation $f_\emptyset = L_{\{1,\ldots,M\}}(f)$, $f_u = L_{\{1,\ldots,M\}/u}(f) - \sum_{v\subset u} f_v$ with $L_w(f) = \int_{[0,1]^{|w|}} f(x_1, \ldots, x_M)\, dx_w$. Thus, f is decomposed into a constant, a sum of one-dimensional functions, a sum of two-dimensional functions, and so on. The involved functions are generated by proper partial integration and telescopic corrections according to the recurrence relation.

$$
\begin{aligned}
&+ \sum_{i_1<i_2<i_3}^{M} F_{\{i_1,i_2,i_3\}}(X_{\{i_1,i_2,i_3\}}) \\
&+ \dots \\
&+ F_{\{i_1,\dots,i_M\}}(X_{\{i_1,\dots,i_M\}}) \\
&=: \sum_{U\subseteq\{1,\dots,M\}} F_U(X_U),
\end{aligned}
$$

where X_U denotes the set of variables $\{X_i\}_{i\in U}$ and $U \subseteq \{1,\dots,M\}$.

Here, each term $F_{\{i_1,\dots,i_k\}}$ is defined as follows:

$$
\begin{aligned}
F_\emptyset &= 0 \\
F_{\{i_1\}}(X_{\{i_1\}}) &= \gamma_{\{i_1\}}\big(E_{\{i_1\}}(X_{\{i_1\}}) - F_\emptyset\big) \\
F_{\{i_1,i_2\}}(X_{\{i_1,i_2\}}) &= \gamma_{\{i_1,i_2\}}\big(E_{\{i_1,i_2\}}(X_{\{i_1,i_2\}}) - F_{\{i_1\}}(X_{\{i_1\}}) - F_{\{i_2\}}(X_{\{i_2\}}) - F_\emptyset\big) \\
&\dots \quad \dots \\
F_{\{i_1,\dots,i_k\}}(X_{\{i_1,\dots,i_k\}}) &= \gamma_{\{i_1,\dots,i_k\}}\big(E_{\{i_1,\dots,i_k\}}(X_{\{i_1,\dots,i_k\}}) \\
&\quad - \sum_{U\subseteq I,|U|=k-1} F_U(X_U) \\
&\quad - \sum_{U\subseteq I,|U|=k-2} F_U(X_U) \\
&\quad \dots \\
&\quad - \sum_{U\subseteq I,|U|=1} F_U(X_U) - F_\emptyset\big) \\
&\dots \quad \dots,
\end{aligned}
$$

where the constant function $F_\emptyset$ is set equal to zero since it corresponds to the energy of an empty molecular system and a set $\{\gamma_I\}_{I\subseteq\{1,\dots,M\}}$ of weights $\gamma_I \in \{0,1\}$ is involved to switch on and off the considered interaction terms. I.e. we have

$$
E^M(X_1,\dots,X_M) = \sum_{U\subseteq\{1,\dots,M\}} F_U(X_U), \tag{11.6}
$$

where

$$
F_U(X_U) = \gamma_U\big(E_U(X_U) - \sum_{k=0}^{|U|-1} \sum_{V\subseteq U,|V|=k} F_V(X_V)\big) \tag{11.7}
$$

and $E_\emptyset = 0$. Let us for the moment assume that all γ_I are set to one. Then the decomposition is exact and contains 2^M different terms due to the power set construction. In general it might be that all terms are equally important up to the last, M-dimensional one, or, in the extreme case that the last term might be the only important one and thus nothing is gained from this decomposition. However, if the size of the terms decay fast with e.g. the order of the terms, then a proper truncation of the ANOVA series expansion results in a substantial reduction in computational complexity. We then only have to deal with a sequence of lower-dimensional subproblems which are associated to the remaining lower-dimensional energy terms of the decomposition.

Let us remark that the energy functions $F_{\{i_1,\ldots,i_k\}}$ in (11.6) may be recognized as an expansion of many-body interaction contributions, as in [29]. This leads us to the following assumption which is central to our further approach: There is a certain decay in the contribution of each order k of the ANOVA expansion and this results in a monotone convergence of the approximation error with rising order. Consequently, from a certain order onward, we may neglect the higher higher-order terms in the ANOVA decomposition. This results in a good approximation to the true result[7] with an accuracy which is related to the order parameter at which the truncation was executed. This assumption is also strongly supported by the success of conventional two- and many-body potential functions used in classical molecular dynamics, such as short range pair-potentials like harmonic springs, the Morse potential and the Lennard-Jones potential, three- and four-body potential like angle and dihedral potential functions and more advanced many-body potential functions which involve a local coordination number (that is the local density of atoms) like Tersoff's potential [36], the embedded atom method [11] or Brenner's reactive bond order potential for hydrocarbons [7]. Here, in any case, only a small number of neighboring atoms are involved in the potential forms, for further details see [18].

Our ansatz is as follows: We decompose the total energy function (11.4) in an ANOVA series expansion as in (11.6) where we include only terms up to a certain order k, which we call the *bond order* of the approximation. Now, let $G = (P, K)$ be the associated graph that represents the bond structure of the molecular system under consideration. For reasons of simplicity we assume that this graph is connected. Then, we neglect in a second step even further interaction terms in the ANOVA expansion. These terms contain as parameters the degrees of freedom which belong to nuclei in I that are not connected by a path in the graph G_I, i.e. we additionally eliminate those terms whose induced subgraph G_I is not connected by setting γ_I to zero. Note here that each set $I = \{i_1, \ldots, i_{|I|}\}$ of nuclei parameters indices for each term $E_{\{i_1,\ldots,i_{|I|}\}}(X_{\{i_1,\ldots,i_{|I|}\}})$ in (11.6) is directly associated to an induced subgraph $G_I = (P_I, K_I)$ of the total graph G with $P_I = \{v_i\}_{i\in I}$ and $K_I = \{\{v_1, v_2\} \in K : v_1 \in I, v_2 \in I\}$. This second elimination step is motivated by the locality of the electronic wave functions: Atoms that share a bond to a nearby atom will be

[7] Note that, in practice, the global electronic problem is only solved approximately anyway, by e.g. DFT, CC, CI.

strongly influenced by changes in the chemical vicinity of nearest or next-nearest bonding partners whereas atoms that share no bond to a nearby atom will not.

Altogether, this can be described by an approximation to the ground state energy according to (11.6) and (11.7). To this end, let $G = (P, K)$ be the interaction graph of the molecular system under consideration. We then define a set of graph-related weights $\{\gamma_U^G\}_{U \subseteq \{1,\dots,M\}}$ by

$$\gamma_U^G = \begin{cases} 1, & \text{if the subgraph } G_U \text{ of } G \text{ (induced by } U\text{) is connected,} \\ 0, & \text{else.} \end{cases} \tag{11.8}$$

This definition is motivated from the following observation. Let us assume that $E_{A \cup B}(X_A, X_B) = E_A(X_A) + E_B(X_B)$ for all pairs of disconnected subgraphs G_A and G_B which are induced by disjoint subsets $A, B \subseteq P$, $A \bigcap B = \emptyset$ and for simplicity let us further assume that all weights are set to one. Then we can derive the following statement:

Lemma 11.1. *Let $G = (P, K)$ be an interaction graph. Let $A, B \subseteq P$, $A \bigcap B = \emptyset$ and let the subgraphs G_A and G_B induced by A and B, respectively, be disconnected. Then*

$$F_{A \cup B}(X_{A \cup B}) = 0.$$

Proof. We use induction: The base case can be easily seen for graphs $G = (P, K)$ with sets $|P| \leq 2$. Let us assume that the statement holds for graphs $G = (P', K')$ with $|P'| \leq n$. Now let $G = (P, K)$ with $|P| = n + 1$. Note that from the recursive definition of F_U it immediately follows that

$$E_U(X_U) = \sum_{u \subseteq U} F_U(X_U)$$

holds for all $U \subseteq P$. With the assumption $E_{A \cup B}(X_{A \cup B}) = E_A(X_A) + E_B(X_B)$ and $F_\emptyset = 0$, we then obtain

$$\begin{aligned} F_{A \cup B}(X_{A \cup B}) &= E_{A \cup B}(X_{A \cup B}) - \sum_{a \subseteq A, a \neq \emptyset} F_a(X_a) - \sum_{b \subseteq B, b \neq \emptyset} F_b(X_b) \\ &\quad - \sum_{\substack{a \subset A, b \subset B \\ a \neq \emptyset, b \neq \emptyset, |a \cup b| < |A \cup B|}} F_{a \cup b}(X_{a \cup b}) - F_\emptyset \\ &= E_A(X_A) + E_B(X_B) - \sum_{a \subseteq A} F_a(X_a) \\ &\quad - \sum_{b \subseteq B} F_b(X_b) - \sum_{\substack{a \subset A, b \subset B \\ a \neq \emptyset, b \neq \emptyset, |a \cup b| < |A \cup B|}} F_{a \cup b}(X_{a \cup b}). \end{aligned}$$

Now, we apply the induction hypothesis to each $F_{A\bigcup B}$: $|a\bigcup b| < |A\bigcup B| \leq |P| = n+1$ and finally obtain

$$\begin{aligned} F_{A\bigcup B}(X_{A\bigcup B}) &= E_A(X_A) - \sum_{a\subseteq A} F_a(X_a) + E_B(X_B) - \sum_{b\subseteq B} F_b(X_b) \\ &\quad - \sum_{\substack{a\subset A, b\subset B \\ a\neq\emptyset, b\neq\emptyset, |a\bigcup b|<|A\bigcup B|}} F_{a\bigcup b}(X_{a\bigcup b}) \\ &= - \sum_{\substack{a\subset A, b\subset B \\ a\neq\emptyset, b\neq\emptyset, |a\bigcup b|<|A\bigcup B|}} F_{a\bigcup b}(X_{a\bigcup b}) = 0. \end{aligned}$$ □

11.3.2 Saturation with Hydrogen

After the motivation of the basic principles of our decomposition scheme in the last section, we now have to face a technical difficulty: A cut-out fragment may have a total spin unequal zero while the molecular system itself has a total spin of zero. As closed-shell calculations are algorithmically both simpler and more stable, this situation would complicate the proposed linear-scaling ansatz.

A step to remedy this situation is a saturation of the dangling bonds of the fragments by adding hydrogen at the places where bonds were cut, causing the total spin of the fragment system to be zero. Due to our telescopic sum approach the effect of the hydrogen atoms actually goes unnoticed.

This correction is schematically depicted in Fig. 11.1 where we just show two atoms and its vertex but omitted for simplicity any further vertices and edges these atoms might be connected to. Here, let us assume that, after cutting the edge k_i, Atom1 should belong to an induced subgraph G', while Atom2 should not. Then, edge $k_i = \{\text{Atom1, Atom2}\}$ is not present in this subgraph. Now, we insert two new terminal vertices H1 and H2 and two new edges $k_1^{(H)} = \{\text{Atom1, H1}\}$ and $k_2^{(H)} = \{\text{Atom2, H2}\}$ so that all dangling bonds are closed. Hence, the new vertex H1 and the edge $k_1^{(H)}$ would be added to G' next to Atom1. By this saturation procedure, we only calculate closed-shell atoms. In particular, the electronic density of the cut edges is thus conserved to a higher degree. Note that this approach is still

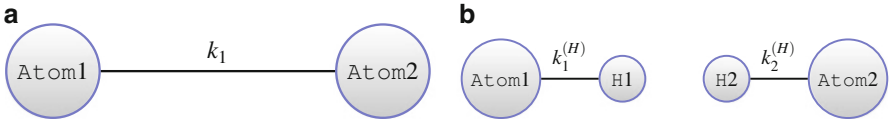


Fig. 11.1 Cut of an edge k_i between two vertices and replacement with two edges $k_1^{(H)}$ and $k_2^{(H)}$ to two newly introduced terminal vertices (hydrogen atoms H1 and H2). (**a**) Edge k_i before cutting. (**b**) Edge k_i disconnected

tunable by the bond length used between new hydrogen vertices and cut-vertices. In our subsequent implementation we use here the equilibrium hydrogen bond lengths of certain small molecules taken from [24].

This saturation procedure can be understood as a re-definition of the electronic Hamiltonian H_e in (11.5): From the known graph G of the molecular system l additional hydrogen vertices, bonds and their graph-dependent coordination $R_i^H(G)$, $1 \leq i \leq l$, are derived and the ground state energy evaluated for this system is defined as:

$$\hat{E}_{i_1,\ldots,i_k}(X_1,\ldots,X_k) := \min_{|\phi_0|=1} \int \phi_0^*(\mathbf{r}) H_e\big(l + \sum_{j=1}^{k} Z_{i_j}, X_{i_1},\ldots,X_{i_k}, R_1^H(G),\ldots,R_l^H(G)\big)\phi_0(\mathbf{r})d\mathbf{r}. \tag{11.9}$$

Note that this saturated energy function is denoted by $\hat{E}$.

The saturation procedure by means of hydrogen renders the role of hydrogen special in our approach. Thus, it is is useless to cut out a fragment at an edge involving only one hydrogen nucleus, as this will only create an additional hydrogen molecule while leaving the edge as it was before. Here, the best procedure is to remove the hydrogen nuclei degrees of freedom from the ANOVA decomposition algorithm, i.e. to drop them completely from the graph G, or to combine them with their bonding partners since they are always terminal vertices anyway, see Fig. 11.2 for an illustration. Hence, in the following, we will not take further heed of the hydrogen atoms which are present in the molecular system. This is also advantageous since e.g. about half of the atoms in organic molecules are hydrogens. Thus, we strongly reduce the necessary number of fragments to be evaluated.

Altogether, to a given bond graph G we define the BOSSANOVA approximate energy up to order k by

$$E^{\mathrm{ANOVA}}(k) = \sum_{U\subseteq\{1,\ldots,M\},|U|\leq k} F_U(X_U), \tag{11.10}$$

with F_U according to the recursive definition (11.7) using energies determined by (11.9) and weights $\{\gamma_U^G\}_{U\subseteq\{1,\ldots,M\}}$ chosen like in (11.8).

11.3.3 Scaling Behavior

We give some theoretical limits on the scaling behavior of the proposed approach along with a constructive proof which our actual implementation follows closely. Here, just the dependence of the number of fragments is to be considered.

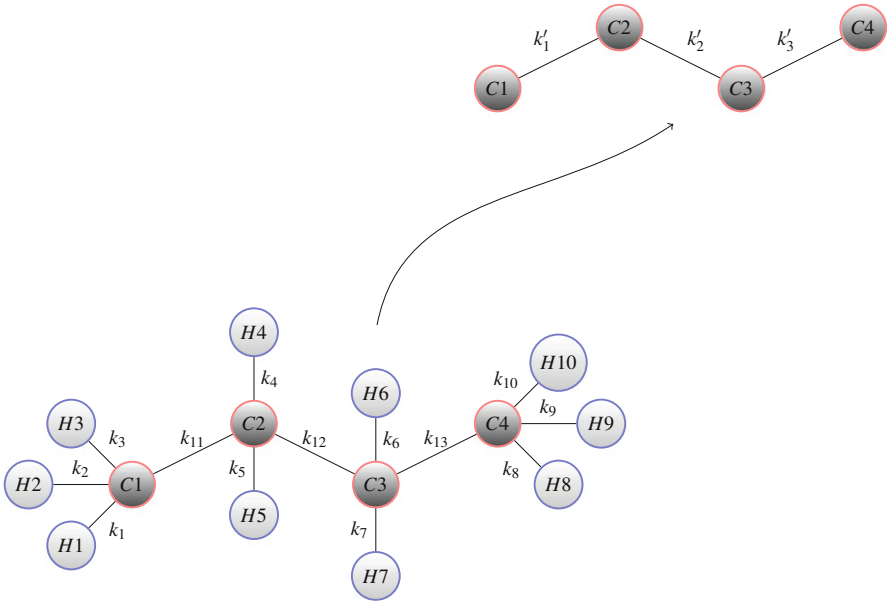


Fig. 11.2 Hydrogen vertices in *light gray* are combined with their bonding partners in *dark gray* to new single vertices. The remaining edges and new vertices have been relabeled, denoted by single digits

The maximum number of fragments possible for a molecular system consisting of M nuclei is given by the power set 2^M. Generally, we obtain for the power set, truncated to contain at most k nuclei, the following relation $\sum_{l=0}^{k} \frac{M!}{l!(M-l)!} \approx M^k$ for small k. However, in our ansatz many fragments are discarded when they do not constitute a connected subgraph of the molecular system. Hence, the true number of fragments is actually a lot smaller as is shown with the following lemma.

Lemma 11.2 (Upper bound on number of connected subgraphs). *Let a connected graph $G = (P, K)$ be given. Let the number of edges per vertex be bounded from above by $c > 0$.*

Then, the number of induced subgraphs $G' = (P', K')$ containing a specific vertex $s \in P$ that are connected, and whose vertex count $|P'| \leq k$ is bounded by the order k, is bounded from above by

$$\sum_{j=1}^{k-1} 2^{c(k-j)} = \sum_{j=1}^{k-1} \Big(\underbrace{2^c}_{=:C} \Big)^{k-j} = \sum_{j=1}^{k-1} C^{k-j} < \sum_{j=0}^{k-1} C^j \overset{C \geqq 2}{=} \frac{C^k - 1}{C - 1} \leq C^k. \tag{11.11}$$

Proof. We will give a constructive proof by starting from a specific vertex and by adding further vertices to the current subgraph, moving along connected edges only.

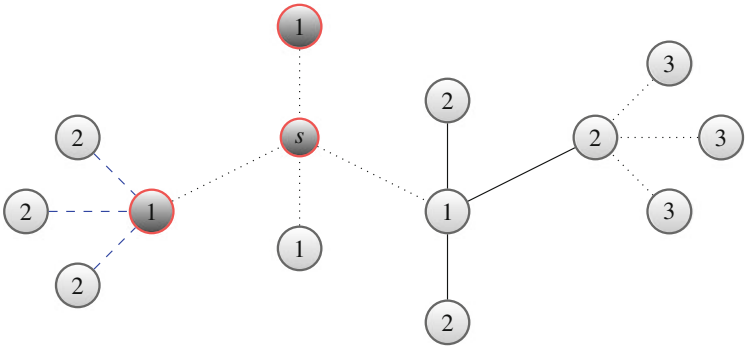


Fig. 11.3 Depiction of the reduced edge set $K'_s(2)$ with *dashed lines* for a given subgraph $G' \subset G$ consisting of vertices in *dark gray* with root vertex s. The vertices of the graph G are designated by the distance to s, all edges outside of the full edge $K_s(2)$ are *dotted*

Let a vertex $s \in P$ be given. We split the edges in equidistant levels with respect to s. To this end, let $K_s(j)$ be the set of terminal edges connecting any $v \in P$ to s via a shortest path of distance $d(s, v) = j$.

Consider now a possible subgraph G' with $s \in P'$. Let $K'_s(j)$ be the reduced set of edges of $K_s(j)$ for which only one of either associated vertices is in P', see Fig. 11.3 for a depiction of these sets. This set is the exploration boundary of G' at distance j with respect to s.

The cardinality of the power set of the reduced set of edges $K'_s(j)$ is $2^{|K'_s(j)|}$ for a level j. Therefore, we obtain $\sum_{j=1}^{k-1} 2^{|K'_s(j)|}$ possible sets by summing over all $k-1$ levels and ignoring that the number per level is actually not independent. With the upper bound on the vertex degree it follows that $|K'_s(j)|$ is bounded from above by $c|I_{j-1}|$ where I_{j-1} denotes the set of vertices added on level $j-1$. In Fig. 11.3 these are nodes designated with "1" and colored in dark gray. Furthermore, $|I_j|$ is bounded from above by $k-j$ because at least one vertex has to be added per level and there is already one root vertex. Putting it all together and using the partial sum of the geometric series results finally in (11.11). □

Hence, the sum of all possible subgraphs with at most k vertices only depends on the bond order k and the highest degree c over all vertices v in G. As we go over all vertices $s \in P$ as root vertices, the number of fragments scales as $O(M \cdot C^k)$.

11.4 Numerical Results

Now we present the results of our numerical experiments. This section is divided into three parts. In the first part, we look at the scaling behavior in terms of runtime to assure linear complexity. In the second part, we give the approximate total energy for smaller molecules to indicate the good approximation quality of the approach. In

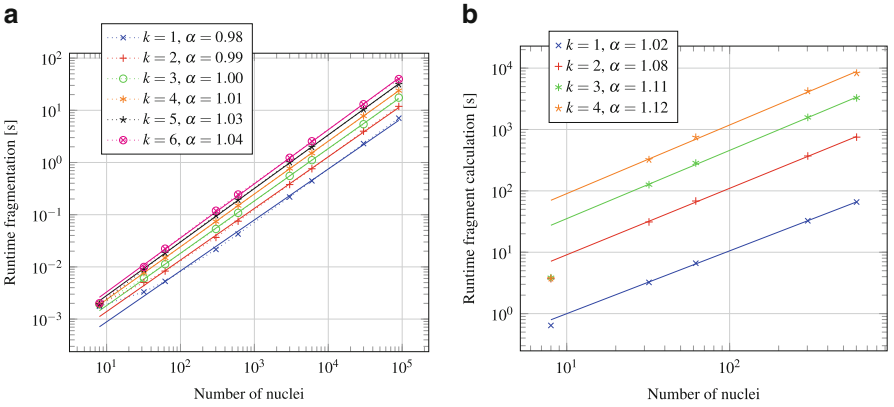


Fig. 11.4 Runtimes for the fragmentation and subsequent calculation of the individual fragments for alkanes of increasing length and varying truncation order $k = 1, \ldots, 6$. (**a**) Fragmentation. (**b**) Calculation

the third and final part, we look at a large biomolecule and assess the applicability of the approach for large-scale calculations.

As approximate computational method for the electronic subproblems associated with the different fragments we have chosen the closed-shell Hartree Fock method with the Gaussian basis "6-311*G" set as implemented in MPQC [23]. We use evaluations of the total molecule as reference results (full HF) to compare the approximation error against.

11.4.1 Scaling Study

In the first part of this subsection we investigate the computational scaling behavior of our BOSSANOVA implementation with respect to the number of nuclei M and with respect to the truncation order k. From the theoretical considerations of the previous section, we here expect a linear scaling complexity with M.

To this end, we studied alkanes of varying length. In Fig. 11.4a the total runtime for the fragmentation procedure is given and in Fig. 11.4b the cumulative runtime for the calculation of all fragment problems is depicted. Both show linear scaling behavior with the number of nuclei M as expected. Additionally, we see that the time required for the calculation indeed increases polynomially with the truncation order.

Finally, we measure the crossover point for our ansatz—that is when the fragment calculations require less time than the reference calculation of the full system.[8] To

[8] As can be seen from Fig. 11.4, the fragmentation times are negligible.

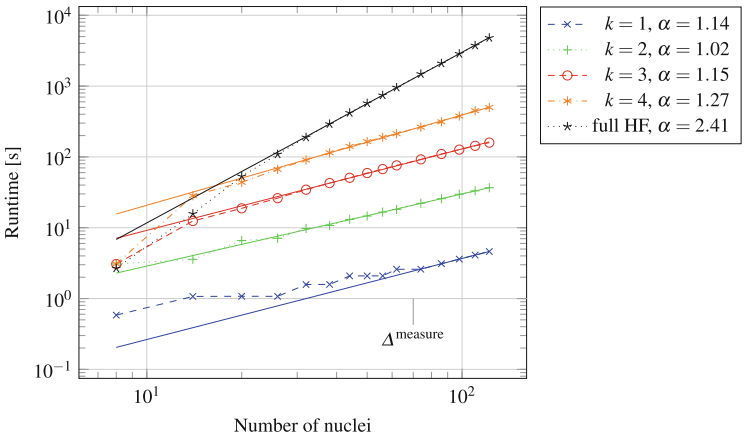


Fig. 11.5 Measured runtime of the calculation of alkanes of increasing length, via standard closed-shell HF and via BOSSANOVA for orders $k = 1, \ldots, 4$. *Solid lines* give linear regression fits to overall behavior

this end, we use four solvers for the fragment problems in parallel and compare against the runtime of MPQC running on four processes for the reference calculation in Fig. 11.5. The respective crossover point is where the black curve intersects the other curves associated with the varying truncation order k.

We notice that at order $k = 4$ we obtain a crossover in runtimes at $M \approx 20$ which is roughly an order of magnitude in number of atoms, or three orders in total run time, lower than that achieved by other linear-scaling schemes, e.g. ONETEP [34], see also [15].

11.4.2 Qualitative Study

In this section we investigate the approximation quality of the proposed approach. To this end, let us first give some remarks on what a threshold for a good approximation would be. HF calculations do not give the so-called correlation energy. However, due to the finite basis set they also never reach the true HF ground state energy but only an upper bound. For the employed "6-311G*" basis set we have estimated this finite basis set error (by employing even larger basis sets) to be 1.81×10^{-4} with respect to the true HF energy of alkanes. Hence, if we find the relative error $\Delta E(k)$ of the approximated energy E^{ANOVA} according to (11.10),

$$\Delta E(k) = \frac{E^{\text{SCF}} - E^{\text{ANOVA}}(k)}{E^{\text{SCF}}}, \tag{11.12}$$

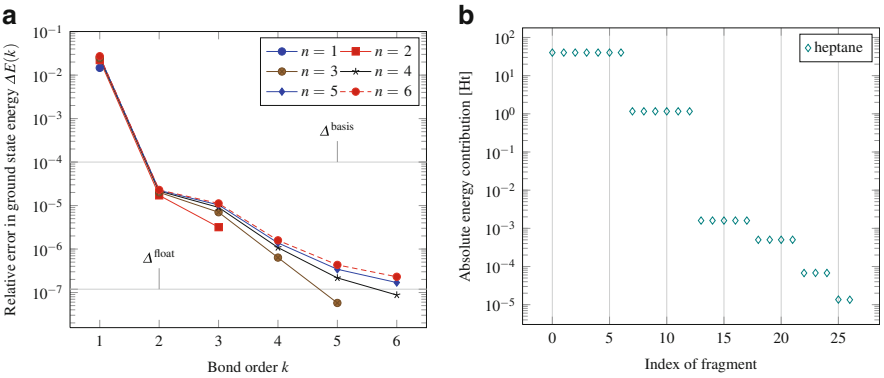


Fig. 11.6 Approximation of the total ground state energy for various alkanes over the truncation order k and absolute value of the energy contribution of each fragment of heptane sorted by increasing number of nuclei. (**a**) Alkanes. (**b**) Heptane

Table 11.1 Relative error $\Delta E(k,n)$ for increasing truncation order k and varying chain length n of the alkane molecule

k	$\Delta E(n=3)$	$\Delta E(n=4)$	$\Delta E(n=5)$	$\Delta E(n=6)$
1	$2.47 \cdot 10^{-2}$	$2.60 \cdot 10^{-2}$	$2.67 \cdot 10^{-2}$	$2.72 \cdot 10^{-2}$
2	$2.02 \cdot 10^{-5}$	$2.16 \cdot 10^{-5}$	$2.24 \cdot 10^{-5}$	$2.29 \cdot 10^{-5}$
3	$7.01 \cdot 10^{-6}$	$9.06 \cdot 10^{-6}$	$1.03 \cdot 10^{-5}$	$1.12 \cdot 10^{-5}$
4	$5.95 \cdot 10^{-7}$	$1.08 \cdot 10^{-6}$	$1.35 \cdot 10^{-6}$	$1.55 \cdot 10^{-6}$
5	$8.50 \cdot 10^{-8}$	$1.91 \cdot 10^{-7}$	$3.06 \cdot 10^{-7}$	$4.26 \cdot 10^{-7}$
6	0.0	$6.38 \cdot 10^{-8}$	$1.53 \cdot 10^{-7}$	$2.13 \cdot 10^{-7}$

to be closer than $\Delta^{\text{basis}} = 10^{-4}$ to the reference calculation E^{SCF}, we define the approximation to be good. As a second threshold value we use $\Delta^{\text{float}} = 1.19 \times 10^{-7}$ as the output precision of values, i.e. below that value numerical rounding artifacts may appear.

In the following we give the numerical results for various chain molecules, namely alkanes, alkenes, alkines, and homologous chains consisting of boron and nitrogen. Let us remark that, for certain small lengths, we already reach the exact result for small truncation order k due to the nature of the telescopic sum.

In Fig. 11.6a and Table 11.1 we give the relative error of the energy calculated for alkanes of length n with formula $C_{2n}H_{2n+4}$. We notice that we are below the estimated threshold Δ^{basis} already for $k = 2$. Also, the error grows only very slowly for longer chains. Hence, the approximation works very well for these linear chain molecules, whose graph forms a tree and each edge represents only a single bond.

Furthermore, we depict the absolute value of the contribution to the total energy per fragment for heptane in Fig. 11.6b. Due to the symmetry of the molecular system we clearly see levels of equal values in the graph. The difference between these levels closely follows the error obtained, e.g. 10^{-2} between level $k = 1$ and $k = 2$ and 10^{-3} between level $k = 2$ and $k = 3$. Hence, we feel that this can be taken as a rough error estimate when a full calculation is unavailable.

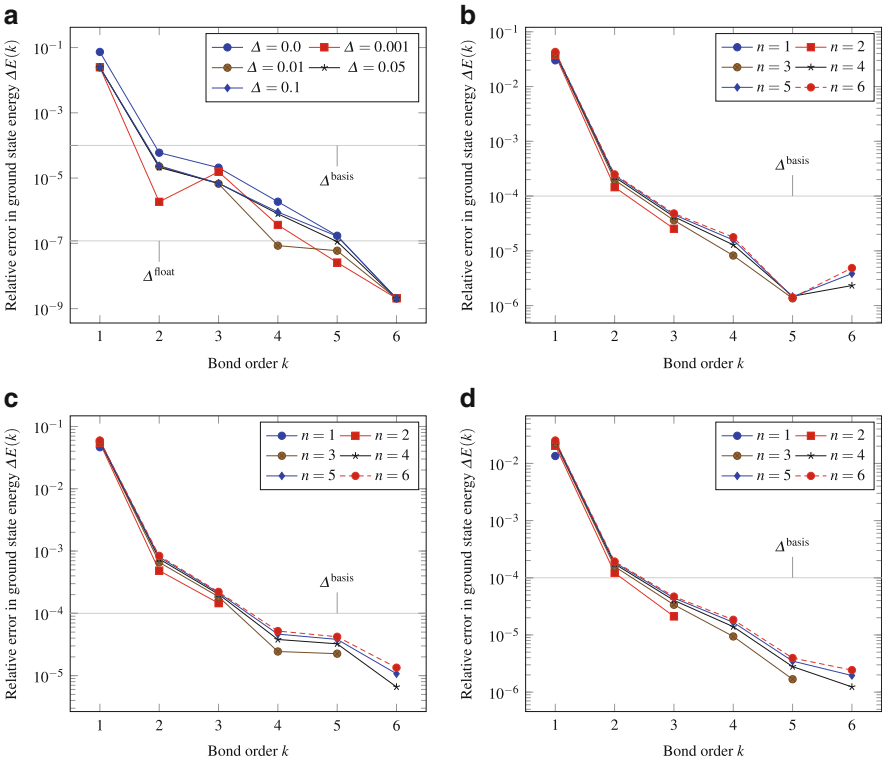


Fig. 11.7 Approximation of the total energy for hexane with nuclei coordinates under random perturbation of magnitude Δ, alkenes and alkines with double and triple bonds, and boron-nitride chain molecules of varying length n. (**a**) Hexane with distored nuclei coordinates. (**b**) Alkenes. (**c**) Alkines. (**d**) Boron-nitrogen chain

The approximation for hexane, alkenes and alkines, and boron-nitrogen chains of varying lengths with distored coordinates, higher bond degrees or different nuclei elements are depicted in Fig. 11.7.

We see that perturbation affects the approximation quality only negligibly. A stronger effect is seen with double and triple bonds as in alkenes and alkines or for different nuclei elements as with the boron nitrogen chain. However, we still reach the threshold Δ^{basis} at $k = 3$ and notice that the decrease with chain length n is very small.

Moving on from these simple chain molecules to more complex bond graphs we come to molecular systems with aromatic rings. These are particularly difficult as the gain in energy due to the delocalized π-electrons is captured only when the complete ring is taken into account as a fragment. As an example we take naphthalene which consists of two interconnected aromatic rings and coronene which consists of six interconnected aromatic rings. In Fig. 11.8 we have calculated the approximative energy of these molecules two times: Once, we calculated the

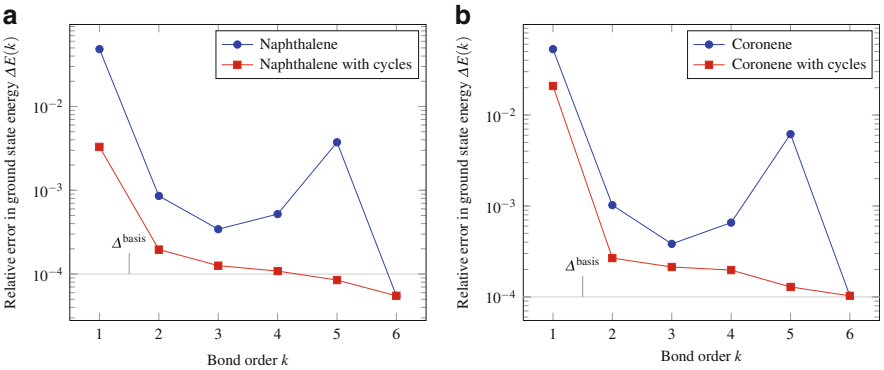


Fig. 11.8 Relative error of the total energy for molecules with delocalized electrons over the truncation order k. In the second calculation cycles in the interaction graph are taken into account irrespectively of the truncation order. (**a**) Naphthalene. (**b**) Coronene

energy in the proposed fashion with increasing truncation order. The second time, however, we took the full cycles in the graph as extra fragments into account.

We immediately notice the effect: While with the first calculation the approximation error decreases up to $k = 3$, it increases afterwards as higher-order fragments are strained due to the ring-like geometry of the full system. In the second calculation this decrease is absent although we never calculate the full system consisting of multiple interconnected rings. Moreover, we reach the threshold Δ^{basis} at around $k = 5$.

11.4.3 Quantitative Study

As an example of a truly large molecule we have chosen the interferon alpha (1ITF), taken from the Protein Data Bank [6] and amended it by hydrogens from topological knowledge via [22] that go undetected in the x-ray spectroscopy of the structure. The structure consists in total of 2,698 nuclei.

Due to the larger number of nuclei a reference calculation is infeasible. Instead, we give in Fig. 11.9 the contributions to the telescopic sum from each individual fragment sorted by the number of nuclei. Each absolute energy value is given as a tiny dash in the figure that as a whole emphasizes certain levels of similar values, compare with Fig. 11.6b. Also, we give exemplary fragments to each of the more prominent levels in Fig. 11.9b.

We notice a similar decay in the absolute magnitude as for alkane. This indicates a good approximation of the total energy of the structure. Judging from our previous remarks when investigating the approximation quality with alkane we see that the obtained ground state energy value of $-68{,}467.41$ Ht is accurate to relative precision

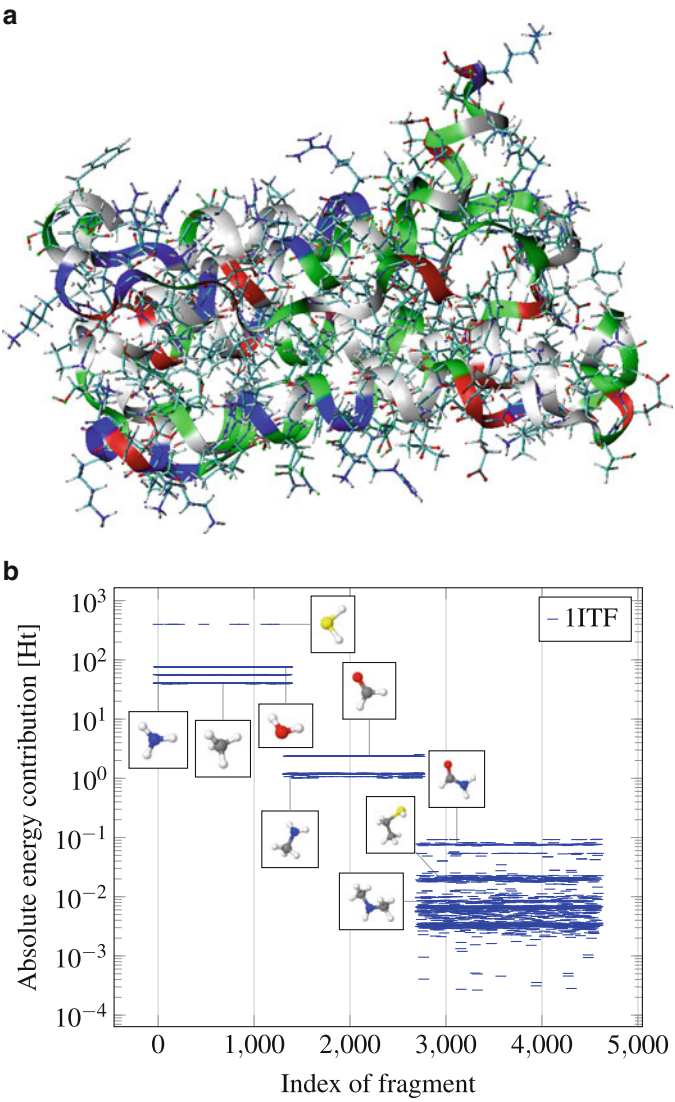


Fig. 11.9 Ball and stick model of interferon alpha (PDB key: 1ITF) combined with a ribbon view of the main chain. The configuration is split up into fragments of up to $k = 3$ for which we give each's contribution to the total energy and examples of typical fragment subsystems. (**a**) Ball-stick-model. (**b**) Absolute contribution per fragment

of 10^{-3} to 10^{-4}. This especially underlines the usefulness of the empirical potential approaches for these large biomolecules, see [32].

We remark that the cumulated solver runtime is 6.28 h for this system. Hence, we see that our proposed scheme is especially well-suited to large molecules. Note

furthermore that long-range Coulomb interactions can additionally be computed via one of the well-known schemes [4] in a first-order perturbation calculation [20] under the assumption that the wavefunctions do not change significantly anymore.

Concluding Remarks

In this article we presented the BOSSANOVA decomposition approach for the approximate solution to the electronic Schrödinger equation for a given molecular system. It involves an ANOVA series expansion of an electronic energy function in the framework of the Born-Oppenheimer molecular dynamics. A truncation of this series at a certain *bond order* and the elimination of certain further terms by a locality constraint of the electronic wavefunction plus some additional hydrogen saturation results in a set of fragments of the overall molecule. Now, each of the associated electronic subproblems may be solved with e.g. HF, CI, CC, or DFT methods. A proper combination of these solutions of the subproblems then leads to an approximate total ground state energy. This is an extension of the so-called additivity models which are well-known in chemistry.

We gave a description how this truncated BOSSANOVA expansion can be derived for any given graph. Furthermore we showed theoretically as well as practically that our new method indeed scales linearly with the number M of atoms in the overall problem. We gave numerical results for chain molecules where the obtained relative accuracy was well below 10^{-4} for $k = 3$, which is the relative precision of the reference calculation with respect to an infinite basis set. We also investigated aromatic systems with delocalized electrons where an inclusion of full cycles aids the approximation significantly to also achieve 10^{-4} relative precision. This is roughly the precision available to HF calculations with moderately sized basis sets.

Note that the impact of the neglected long-range Coulomb energy on the accuracy of the method and ways to recover this contribution is given elsewhere, see [20]. Note furthermore that our BOSSANOVA approach is not rid of empirical parameters due to the necessity to saturate dangling bonds with hydrogen in the fragmentation process. Since the typical bond lengths and angles of hydrogenated systems are well assessed by measurements, we hope that a careful collection of robust values into a database may enable a broad range of application for the BOSSANOVA method.

Let us also point out that our approach is trivial to parallelize since the evaluation of each fragment by an appropriate solver can be done independently, see [20]. Furthermore, since each fragment only contains a number of atoms which is roughly equal to the bond order k (neglecting hydrogen), the evaluation of the subproblems is possible already on very small machines with minimal memory requirements. Of course, also the memory cost scales only linearly. Thus, if the energy of a single fragment is calculated in seconds by

(continued)

e.g. a solver which is specifically tailored to the fast but precise evaluation of small and isolated systems, even a number of 10^5 or 10^6 fragments is within reach and the approximate total ground state energy evaluation of huge homogeneous molecular systems becomes computationally feasible. This has been shown by the calculation of the ground state energy of interferon alpha.

Finally, let us remark on how the BOSSANOVA method may be incorporated into a general coupling scheme of QM and MM. The BOSSANOVA fragmentation would be executed only in a given local domain, i.e. the active region where QM is locally needed. The resulting fragments are then forwarded to a suitable QM solver whereas the surrounding passive environment would not be fragmented but is directly passed on to a MM solver. Our BOSSANOVA scheme is closely related to conventional many-body potentials (however in an ab-initio fashion) with variable many-body order. Furthermore, due to the fragmentation process, the interface region is not sharply defined. Therefore, we believe that this approach also remedies the problems of energy and electron density leaking of other local coupling methods to a certain extent.

Acknowledgements This research was funded by the Deutsche Forschungsgemeinschaft (DFG) within the framework of the priority program SPP1324.

References

1. Abell, G.C.: Empirical chemical pseudopotential theory of molecular and metallic bonding. Phys. Rev. B **31**(10), 6184–6196 (1985)
2. Amovilli, A., Cacelli, I., Campanile, S., Prampolini, G.: Calculation of the intermolecular energy of large molecules by a fragmentation scheme: application to the 4-n-pentyl-4-cyanobiphenyl (5CB) dimer. J. Chem. Phys. **117**, 3003–3012 (2002)
3. Antes, I., Thiel, W.: Adjusted connection atoms for combined quantum mechanical and molecular mechanical methods. J. Phys. Chem. A **103**(46), 9290–9295 (1999)
4. Arnold, A., Bolten, M., Dachsel, H., Fahrenberger, F., Gähler, F., Halver, R., Heber, F., Hofmann, M., Holm, C., Iseringhausen, J., Kabadshow, I., Lenz, O., Pippig, M., Potts, D., Sutmann, G.: Comparison of scalable fast methods for long-range interactions. Phys. Rev. E **88**(6), 063,308 (2013)
5. Ayala, P.Y., Scuseria, G.E.: Linear scaling second-order Moeller-Plesset theory in the atomic orbital basis for large molecular systems. J. Chem. Phys. **110**(8), 3660–3671 (1999)
6. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The protein data bank (2000). http://www.pdb.org/
7. Brenner, D.W.: A second-generation reactive bond order (REBO) potential energy expression for hydrocarbons. J. Phys.: Condens. Matter **14**, 783–802 (2002)
8. Challacombe, M.: A simplified density matrix minimization for linear scaling self-consistent field theory. J. Chem. Phys. **110**, 2332–2342 (1999)
9. Collins, M.A., Deev, V.A.: Accuracy and efficiency of electronic energies from systematic molecular fragmentation. J. Chem. Phys. **125**, 104,104 (2006)

10. Csyani, G., Albaret, T., Payne, M.C., De Vita, A.: Learn on the fly: a hybrid classical and quantum-mechanical molecular dynamics simulation. Phys. Rev. Lett. **93**(17), 175,503 (2004)
11. Daw, M.S., Baskes, M.I.: Embedded-atom method: derivation and application to impurities, surfaces and other defects in metals. Phys. Rev. B **29**(12), 6443–6453 (1984)
12. Deev, V., Collins, M.A.: Approximate ab initio energies by systematic molecular fragmentation. J. Chem. Phys. **122**(15), 154,102 (2005)
13. Ercolessi, F., Adams, J.B.: Interatomic potentials from 1st-principles calculations – the force-matching method. Europhys. Lett. **26**(8) 583–588 (1994)
14. Fonseca Guerra, C., Snijders, J.G., te Velde, G., Baerends, E.J.: Towards an order-N DFT method. Theor. Chem. Acc. **99**(6), 391–403 (1998)
15. Goedecker, S.: Linear scaling electronic structure methods. Rev. Mod. Phys. **71**(4), 1085–1123 (1999)
16. Greengard, L., Rokhlin, V.: The fast multipole method for gridless particle simulation. Comput. Phys. Commun. **48**, 117–125 (1988)
17. Gresh, N., Claverie, P., Pullman, A.: Theoretical studies of molecular conformation. Derivation of an additive procedure for the computation of intramolecular interaction energies. Comparison with ab-initio SCF computations. Theor. Chim. Acta **66**, 1–20 (1984)
18. Griebel, M., Knapek, S., Zumbusch, G.: Numerical Simulation in Molecular Dynamics – Numerics, Algorithms, Parallelization, Applications. Springer, Heidelberg (2007)
19. Hayes, M.Y., Li, B., Rabitz, H.: Estimation of molecular properties by high-dimensional model representation. J. Phys. Chem. **110**, 264–272 (2006)
20. Heber, F.: Ein systematischer, linear skalierender Fragmentansatz für das Elektronenstukturproblem. Ph.D. thesis, Rheinische Friedrich-Wilhelms-Universität Bonn (2014)
21. Hoeffding, W.: A class of statistics with asymptotically normal distribution. Ann. Math. Stat. **19**(3), 293–325 (1948)
22. Humphrey, W., Dalke, A., Schulten, K.: VMD – visual molecular dynamics. J. Mol. Graph. **14**, 33–38 (1996)
23. Janssen, C.L., Nielsen, I.B., Leininger, M.L., Valeev, E.F., Kenny, J.P., Seidl, E.T.: The Massively Parallel Quantum Chemistry Program (MPQC), Version 2.3.0. Sandia National Laboratories, Livermore (2008). http://www.mpqc.org/
24. Johnson, R.D., III: NIST computational chemistry comparison and benchmark database, NIST Standard Reference Database Number 101 (2006). http://srdata.nist.gov/cccbdb
25. Kitaura, K., Ikeo, E., Asada, T., Nakano, T., Uebayasi, M.: Fragment molecular orbital method: an approximate computational method for large molecules. Chem. Phys. Lett. **313**, 701–706 (1999)
26. Laio, A., Van de Vondele, J., Rothlisberger, U.: A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations. J. Comput. Chem. **116**(16), 6941–6947 (2002)
27. Li, X.P., Nunes, R.W., Vanderbilt, D.: Density-matrix electronic-structure method with linear system-size scaling. Phys. Rev. B **47**, 10,891–10,894 (1993)
28. Liu, W.K., Karpov, E.G., Zhang, S., Park, H.S.: An introduction to computational nanomechanics and materials. Comput. Methods Appl. Mech. Eng. **193**, 1529–1578 (2004)
29. Marx, D., Hutter, J.: Ab initio molecular dynamics: theory and implementation. In: Modern Methods and Algorithms of Quantum Chemistry. NIC Series, vol. 1, pp. 301–440. Forschungszentrum Juelich, Deutschland (2000)
30. Maseras, F., Morokuma, K.: IMOMM – a new integrated ab-initio plus molecular mechanics geometry optimization scheme of equilibrium structures and transition-states. J. Comput. Chem. **16**(9), 1170–1179 (1995)
31. Parr, R.G., Yang, W.: Density-Functional Theory of Atoms and Molecules. Oxford Science Publications, New York (1989)
32. Ponder, J.W., Case, D.A.: Force fields for protein simulation. Adv. Protein Chem. **66**, 27–85 (2003)
33. Sauer, J., Sierka, M.: Combining quantum mechanics and interatomic potential functions in ab initio studies of extended systems. J. Comput. Chem. **21**(16), 1470–1493 (2000)

34. Skylaris, C.K., Haynes, P.D., Mostofi, A.A., Payne, M.C.: Introducing ONETEP: linear-scaling density functional simulations on parallel computers. J. Chem. Phys. **122**(8), 84,119 (2005)
35. Szabo, A., Ostlund, N.S.: Modern Quantum Theory – Introduction to Advanced Electronic Structure Theory. Dover, New York (1996)
36. Tersoff, J.: Modeling solid-state chemistry: interatomic potentials for multicomponent systems. Phys. Rev. B **39**, 5566–5568 (1989)
37. Van der Vaart, A., Gogonea, V., Dixon, S.L., Merz, K.M., Jr.: Linear scaling molecular orbital calculations of biological systems using the semiempirical divide and conquer method. J. Comput. Chem. **21**(16), 1494–1504 (2000)
38. Velde, G.T., Bickelhaupt, F.M., Baerends, E.J., Guerra, C.F., Van Gisbergen, S.J.A., Snijders, J.G., Ziegler, T.: Chemistry with ADF. J. Comput. Chem. **22**(9), 931–967 (2001)
39. Vreven, T., Morokuma, K.: On the application of the IMOMO (integrated molecular orbital + molecular orbital) method. J. Comput. Chem. **21**(16), 1419–1432 (2000)

Chapter 12
Tensor Spaces and Hierarchical Tensor Representations

Wolfgang Hackbusch and Reinhold Schneider

Abstract In the present report we provide a brief introduction into recently developed hierarchical tensor representations. The new formats extend the well-known Tucker format into a hierarchical framework, by combining its favourable characteristics with low-order scaling properties. We demonstrate the basic concept of subspace approximation and higher order SVD (HOSVD), and how to extend this in a hierarchical way. We highlight that the present tensor representations are constituting smooth manifolds, and give a perspective how these properties can be used to develop numerical solvers for tensor equations and tensor optimisation problems.

12.1 Introduction

Supported by the DFG Priority program *SPP 1324*, we started a joint German/ Russian project in 2008. Independently in 2009 the groups in Leipzig and Moscow introduced new tensor representations. Since that time strong activities have been started to develop these ideas further and the relationship to established approaches in quantum physics has been recognised [49]. Due to the amount of material, we do not intend to provide a complete overview over all developments since that time, most can be found in the monograph [20]. We only try to describe the basic concepts up to the state of the art in a nutshell. Moreover we have left several topics to the other contributions (see Chaps. 19 and 10 in this volume), e.g. we do not consider the problem of tensor completion here. To provide a short and comprehensive introduction we keep the presentation as self-contained as possible. Nevertheless more detailed information will be required for getting deeper insight into the material [20]. In our bibliography, we emphasise the citations which has

W. Hackbusch (✉)
MPI Mathematik in den Naturwiss., Inselstr. 22, 04103 Leipzig, Germany
e-mail: wh@mis.mpg.de

R. Schneider
Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: schneidr@math.tu-berlin.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_12

been relevant for the present SPP 1324 project, or have been supported by this project, far of being complete. For an exhaustive collection of articles (until 2012) we refer to recent survey articles, e.g. [16, 17, 20, 22].

We confine ourselves to the finite-dimensional setting and introduce tensors as multi-indexed arrays. While real vectors have entries $v_k \in \mathbb{R}$ depending on one index k, the entries of matrices $a_{i,j}$ depend on two indices, tensors $\mathbf{v}_{j_1,\dots,j_d}$ of *order* d carry d indices. For $j \in \{1,\dots,d\}$, we fix index sets $I_j = \{1,\dots,n_j\}$ and the Cartesian product of these index sets,

$$\mathbf{I} = I_1 \times \cdots \times I_d \quad , \quad \mathbf{i} := (i_1,\dots,i_d) \in \mathbf{I}\,. \tag{12.1}$$

A tensor $\mathbf{v}$ is defined by its entries $\mathbf{v}_\mathbf{i} = \mathbf{v}[\mathbf{i}] = \mathbf{v}[i_1,\dots,i_d]$, i.e. the tensor $\mathbf{v}$ can be considered as a mapping

$$\mathbf{v} : \mathbf{I} \to \mathbb{R}\,, \qquad (i_1,\dots,i_d) \mapsto \mathbf{v}[i_1,\dots,i_d]\,.$$

We may write $\mathbf{v} = (\mathbf{v}[\mathbf{i}])_{\mathbf{i}\in\mathbf{I}}$. Hence, we may express the set of the tensors considered above by[1]

$$\mathbb{R}^\mathbf{I} = \{\mathbf{v} : (\mathbf{v}[\mathbf{i}])_{\mathbf{i}\in\mathbf{I}}\}\,.$$

The dimension of the linear space $\mathbb{R}^\mathbf{I}$ is $\Pi_{j=1}^d n_j$, where $n_j = \#I_j$.

The tensor structure is introduced by the following tensor product. Let $u^j \in \mathbb{R}^{I_j}$ be vectors for $j \in \{1,\dots,d\}$. Then the *elementary* tensor product $\mathbf{v} = u^1 \otimes u^2 \otimes \dots \otimes u^d$ is defined entry-wise by

$$\mathbf{v}[i_1,\dots,i_d] = u^1_{i_1} \cdot u^2_{i_2} \cdot \ldots \cdot u^d_{i_d}.$$

It is easy so see that $\mathbb{R}^\mathbf{I}$ is the span of all elementary products. The fact that the tensor space $\mathbb{R}^\mathbf{I}$ is produced by the linear spaces $\mathbb{R}^{I_1}, \mathbb{R}^{I_2}, \dots, \mathbb{R}^{I_d}$, is expressed by the notation

$$\mathbb{R}^\mathbf{I} = \mathbb{R}^{I_1} \otimes \mathbb{R}^{I_2} \otimes \dots \otimes \mathbb{R}^{I_d} = \bigotimes_{j=1}^d \mathbb{R}^{I_j}.$$

For $d = 2$, the tensor product $\mathbf{u} \otimes \mathbf{v}$ can be considered as the matrix $\mathbf{u}\mathbf{v}^\mathsf{T}$.

We recall $\dim \mathbb{R}^\mathbf{I} = \prod_{j=1}^d n_j$, where $n_j = \dim \mathbb{R}^{I_j}$. Set $n := \max_{1\le j\le d} n_j$. Then, the dimension of $\mathbb{R}^\mathbf{I}$ is in general $\mathcal{O}(n^d)$, scaling exponentially in d. This is often referred as the *curse of dimensionality*, since even for moderate d the dimension becomes prohibitively large [4, 34, 36]. Therefore it is impossible to store

[1] Given a finite index set J, $\mathbb{R}^J$ can be considered as the set of tuples $\mathbf{w} = (\mathbf{w}_i)_{i\in J}$ with $\mathbf{w}_i \in \mathbb{R}$ or, equivalently, as set of mappings $\mathbf{w} : J \to \mathbb{R}$. $\mathbb{R}^J$ is a vector space of dimension $\sharp J$, where $\sharp J$ denotes the cardinality of the finite index set J.

a tensor $\mathbf{v}$. Even for the smallest nontrivial case $n_j = 2$ and $d = 500$, 2^{500} is a number larger than the estimated number of atoms in the universe. In practice, we cannot deal with these spaces without further approximation or restrictions. What we can hope for are smaller *subclasses* which can be parametrised by much less parameters, e.g. lower dimensional manifolds in which we treat our problems satisfactorily accurate. Additionally we want to identify ways to increase these classes in order to achieve any desired accuracy ε. As an ultimate goal we intend to present algorithms which provide an accuracy ε by a computational effort $\leq C_d(\varepsilon^{-s})$ (storage as well as computational work) with some positive and hopefully rather small s and moderate C_d which grows at most polynomially in d.

Above we started from $I_j = \{1, \ldots, n_j\}$. Instead, we can consider $I_j = J_j \times K_j$, where $J_j = \{1, \ldots, n_j\}$ and $K_j = \{1, \ldots, m_j\}$. Now $\mathbb{R}^{I_j} = \mathbb{R}^{J_j \times K_j}$ is the vector space of $n_j \times m_j$ matrices. Note that $\mathbf{I} = I_1 \times \cdots \times I_d$ has the same cardinality as $\mathbf{J} \times \mathbf{K}$ with $\mathbf{J} = J_1 \times \cdots \times J_d$ and $\mathbf{K} = K_1 \times \cdots \times K_d$. Therefore the tensor space $\mathbb{R}^{\mathbf{I}}$ can be identified with the matrix space $\mathbb{R}^{\mathbf{J} \times \mathbf{K}}$ corresponding to linear mappings from $\mathbb{R}^{\mathbf{J}}$ to $\mathbb{R}^{\mathbf{K}}$. In the latter case, the tensor product is also called the *Kronecker product*.

As mentioned above, we consider the real field $\mathbb{R}$. There are some delicate differences between real and complex tensors, but here they are mostly irrelevant.

The next generalisation replaces the particular vector spaces $\mathbb{R}^{I_j}$ by general vector spaces V_j of any (also infinite) dimension. The abstract definition of a tensor space (see [18]) uses the right diagram. Then, the (algebraic) tensor space $\mathbf{V} = {}_a\bigotimes_{j=1}^d V_j$ and the tensor product $\otimes$ are defined uniquely—up to isomorphism—by the requirement that for any multilinear mapping $\varphi : V_1 \times \ldots \times V_d \to U$ (U is an arbitrary vector space), there is a linear map $\Phi : \mathbf{V} \to U$ such that $\varphi(v^1, \ldots, v^d) = \Phi\left(v^1 \otimes \ldots \otimes v^d\right)$ for all $v^j \in V_j$. A constructive proof is given in [20].

$$\begin{array}{ccc} V_1 \times \ldots \times V_d & \underset{\varphi}{\to} & U \\ \otimes\downarrow & \Phi\nearrow & \\ {}_a\bigotimes_{j=1}^d V_j & & \end{array}$$

For infinite dimensions, one has to distinguish between algebraic and topological tensor spaces. The above definition of ${}_a\bigotimes_{j=1}^d V_j$ yields the algebraic tensor space (indicated by the left index a). *Algebraic* tensors are *finite* linear combinations of elementary tensors.

Assume now that V_j are normed spaces (norm denoted by $\|\cdot\|_j$). We may equip the algebraic tensor space by some norm $\|\cdot\|$ such that the tensor product is continuous, i.e. $\|v^1 \otimes \ldots \otimes v^d\| \leq C \prod_{j=1}^d \|v^j\|_j$. The completion of the normed tensor space $({}_a\bigotimes_{j=1}^d V_j, \|\cdot\|)$ yields the topological tensor space ${}_{\|\cdot\|}\bigotimes_{j=1}^d V_j$. We emphasise that choice of $\|\cdot\|$ is not fixed by the norms $\|v^j\|_j$. A natural condition is

$$\|v^1 \otimes \ldots \otimes v^d\| = \prod_{j=1}^d \|v^j\|_j \quad \text{for all } v^j \in V_j.$$

In this case, $\|\cdot\|$ is called a *crossnorm*. Since this definition involves only elementary tensors, it does not uniquely define the norm for all tensors. The situation improves for Hilbert spaces $(V_j, \langle\cdot,\cdot\rangle_j)$. Also here $_a\bigotimes_{j=1}^d V_j$ may be equipped with different scalar products $\langle\cdot,\cdot\rangle$, but there is one *canonical scalar product* of the tensor space defined by $\langle v^1 \otimes \ldots \otimes v^d, w^1 \otimes \ldots \otimes w^d\rangle = \prod_{j=1}^d \langle v^j, w^j\rangle_j$. The corresponding norm is a crossnorm.

12.2 Subspace Approximation and Tucker Format

We consider the finite-dimensional tensor space

$$\mathbb{R}^{\mathbf{I}} = \bigotimes_{j=1}^d V_j = \bigotimes_{j=1}^d \mathbb{R}^{I_j} \qquad \text{with } \#I_j = n_j\ ,\ j = 1,\ldots,d\ .$$

As mentioned above, we identify a tensor $\mathbf{u} \in \mathbb{R}^{\mathbf{I}}$ with the d-variate function $\mathbf{x} = (x_1,\ldots,x_d) \mapsto \mathbf{u}[\mathbf{x}] = \mathbf{u}[x_1,\ldots,x_d] \in \mathbb{R}$ depending on discrete variables $x_j \in I_j$ (usually called indices). $\mathbb{R}^{\mathbf{I}}$ is equipped with the inner product

$$\langle \mathbf{u}, \mathbf{v}\rangle := \sum_{\mathbf{x}\in\times_{j=1}^d I_j} \mathbf{u}(\mathbf{x})\mathbf{v}(\mathbf{x}) = \sum_{x_1\in I_1} \cdots \sum_{x_d\in I_d} \mathbf{u}[x_1,\ldots,x_d]\mathbf{v}[x_1,\ldots,x_d]$$

and the ℓ_2-norm $\|\mathbf{u}\| = \sqrt{\langle \mathbf{u}, \mathbf{u}\rangle}$.

For each direction $j = 1,\ldots,d$ let $\{\mathbf{b}_{x_j}^j : x_j \in I_j\}$ be a basis of V_j. Then, any tensor $\mathbf{u} \in \mathbb{R}^{\mathbf{I}} = \bigotimes_{j=1}^d V_j$ can be expanded w.r.t. the tensor product basis $\mathbf{b}_{x_1}^1 \otimes \cdots \otimes \mathbf{b}_{x_d}^d$:

$$\mathbf{u} = \sum_{x_1\in I_1} \ldots \sum_{x_d\in I_d} \mathbf{c}[x_1,\ldots,x_d]\ \mathbf{b}_{x_1}^1 \otimes \cdots \otimes \mathbf{b}_{x_d}^d\ .$$

A tensor is said to be represented in *Tucker format* with representation rank $\mathbf{s} = (s_1,\ldots,s_d)$ if

$$\mathbf{u} = \sum\nolimits_{k_1=1}^{s_1} \cdots \sum\nolimits_{k_d=1}^{s_d} \mathbf{c}[k_1,\ldots,k_d]\ \mathbf{b}_{k_1}^1 \otimes \cdots \otimes \mathbf{b}_{k_d}^d. \tag{12.2}$$

The coefficients $\mathbf{c}[\cdots]$ form the so-called core tensor $\mathbf{c} \in \mathbb{R}^{s_1} \otimes \cdots \otimes \mathbb{R}^{s_d}$. Obviously, $s_j \le n_j := \#I_j$ holds. Setting $U_j := \operatorname{span}\{\mathbf{b}_k^j : 1 \le k \le s_k\}$, we obtain subspaces satisfying $\mathbf{u} \in U_1 \otimes \cdots \otimes U_d$. In the next subsection we look for an exact representation (12.2) with minimal representation ranks s_j and the corresponding minimal subspaces U_j.

12.2.1 Minimal Subspaces

Given an algebraic tensor $\mathbf{u} \in \bigotimes_{j=1}^{d} V_j = \mathbb{R}^{\mathbf{I}}$, there are uniquely defined subspaces $U_j^{\min}(\mathbf{u}) = U_j \subset V_j$ of minimal dimension r_j such that

$$\mathbf{u} \in \bigotimes_{j=1}^{d} U_j \,, \qquad U_j \subset V_j \,.$$

This statement also holds for infinite-dimensional vector spaces V_j (cf. [20]). The dimensions $r_j = \operatorname{rank}_j(\mathbf{u}) := \dim U_j$ define the integer tuple $\mathbf{r} = \mathbf{r}(\mathbf{u}) = (r_1, \ldots, r_d)$ which we call the *tensor subspace rank* or *Tucker rank*[2] $\operatorname{rank}_{\mathscr{T}}(\mathbf{u})$ of the tensor $\mathbf{u}$. Given any tuple $\mathbf{r} = (r_1, \ldots, r_d)$, we define the set

$$\mathscr{T}_{\mathbf{r}} := \left\{ \mathbf{u} : \dim(U_j^{\min}(\mathbf{u})) = r_j, 1 \le j \le d \right\}. \tag{12.3}$$

Analogously, $\mathscr{T}_{\le \mathbf{r}}$ is defined by $\dim(U_j^{\min}(\mathbf{u})) \le r_j$.

Each subspace $U_j^{\min}(\mathbf{u})$ can be spanned by some basis $\{\mathbf{b}_k^j : 1 \le k \le r_j\}$. We denote the entries of $\mathbf{b}_k^j$ by $(\mathbf{b}_k^j)_{x_j} =: \mathbf{b}^j[k, x_j]$. We often cast the basis into a tensor (matrix) $\mathbf{b}^j = (\mathbf{b}_k^j)_{k=1,\ldots,r_j} \in \mathbb{R}^{r_j \times n_j}$. Among all possible bases, the orthonormal ones are often convenient:

$$\langle \mathbf{b}_k^j, \mathbf{b}_{k'}^j \rangle = \sum\nolimits_{x_j=1}^{n_j} \mathbf{b}^j[k, x_j] \mathbf{b}^j[k', x_j] = \delta_{k,k'} \,. \tag{12.4}$$

12.2.2 Reconstruction

Given a tensor $\mathbf{u} \in \mathbb{R}^{\mathbf{I}}$ and a Tucker rank $\mathbf{r} \le \operatorname{rank}_{\mathscr{T}}(\mathbf{u})$, we are searching for r_j-dimensional subspaces $U_j \subset V_j$ $(1 \le j \le d)$, which fit best to $\mathbf{u}$. Each subspace U_j is defined by its basis $\{\mathbf{b}_k^j : k = 1, \ldots, r_j\}$ of size $r_j \le n_j$. These bases allow us to write $\mathbf{u}$ in the form

$$\mathbf{u} = \sum_{k_1=1}^{r_1} \cdots \sum_{k_d=1}^{r_d} \mathbf{c}[k_1, \ldots, k_d] \mathbf{b}_{k_1}^1 \otimes \cdots \otimes \mathbf{b}_{k_d}^d$$

or in terms of coefficients

$$\mathbf{u}[x_1, \ldots, x_d] = \sum_{k_1=1}^{r_1} \cdots \sum_{k_d=1}^{r_d} \mathbf{c}[k_1, \ldots, k_d] \mathbf{b}^1[k_1, x_1] \ldots \mathbf{b}^d[k_d, x_d] \,. \tag{12.5}$$

[2] Any $\mathbf{u}$ with representation rank $\mathbf{s} = (s_1, \ldots, s_d)$ in (12.2) has a Tucker rank $\mathbf{r} = (r_1, \ldots, r_d)$ satisfying $r_j \le s_j$.

For orthonormal basis vectors (cf. (12.4)), the core tensor $\mathbf{c}$ can be obtained by projection:

$$\mathbf{c}[k_1,\dots,k_d] = \sum_{x_1=1}^{n_1} \cdots \sum_{x_d=1}^{n_d} \mathbf{u}[x_1,\dots,x_d]\mathbf{b}^1[k_1,x_1]\dots\mathbf{b}^d[k_d,x_d]\,.$$

Next we derive how a basis vector of a minimal subspace can be computed. We start with the case $d = 2$. Second order tensors $(x_1,x_2) \mapsto \mathbf{u}[x_1,x_2]$ can be identified with matrices $\mathbf{U}_{x_1,x_2} := \mathbf{u}[x_1,x_2]$. Spectral theory can be applied to derive the famous result about the singular value decomposition (SVD; cf. [46]),

$$\mathbf{u}[x_1,x_2] = \sum_{k=1}^{r} \mathbf{b}^1[x_1,k]\,\sigma[k]\,\mathbf{b}^2[x_2,k]\,,\ \sigma_1 \ge \sigma_2 \ge \dots > 0\,, \tag{12.6}$$

where the vectors $(\mathbf{b}_k^j)_{k=1}^r$ are orthonormal (cf. (12.4)). The spaces $U_1 \subset V_1$ and $U_2 \subset V_2$ are well defined by the respective basis functions, and the core tensor is $\mathbf{c}[k_1,k_2] = \delta_{k_1,k_2}\sigma_{k_1},\ k_1,k_2 = 1,\dots,r_1 = r_2 = r$.

The generalisation to the tensor case uses the *matricisation* or *unfolding*, which maps the tensor $\mathbf{u}$ into the matrix $\mathbf{M}^j(\mathbf{u}) = \mathbf{M}^j$ with entries $(\mathbf{M}^j)_{x_j,(x_1,\dots,\not{x}_j,\dots,x_d)} := \mathbf{u}[x_1,\dots,x_d]$ [5]. Then the minimal subspaces $U_j = U_j^{\min}(\mathbf{u})$ are given by the range of the matrix $\mathbf{M}^j(\mathbf{u})$. The factorisation $\mathbf{M}^j := \sum_{k=1}^{r_j} \sigma_k^j \mathbf{b}_k^j \otimes \mathbf{v}_k$ by SVD,

$$\mathbf{M}^j_{x_j,(x_1,\dots,\not{x}_j,\dots,x_d)} := \mathbf{u}[x_1,\dots,x_d] = \sum_{k=1}^{r_j} \mathbf{b}^j[k,x_j]\sigma^j[k]\mathbf{v}[k,x_1,\dots,\not{x}_j\dots,x_d]$$

with $\sigma^j[1] \ge \sigma^j[2] \ge \dots > 0$ yields a particular basis of $U_j = U_j^{\min}(\mathbf{u})$, which is called the j-th *HOSVD basis* . We use the notation

$$\mathbf{b}^j := \mathrm{LSVD}\,\mathbf{M}^j(\mathbf{u})$$

for taking the basis from the left part of an SVD, a procedure frequently used in the sequel.

12.2.3 Approximation

The representation of a tensor by (12.5) involves a data size depending on the Tucker ranks, which now are denoted by R_j (instead of r_j). Often, we want to decrease the data size without perturbing the tensor too much. For instance, we aim at another representation of the form (12.5) with $r_j < R_j$. An equivalent formulation is the following. Let $\mathbf{u} \in \mathbf{U} := U_1 \otimes \cdots \otimes U_d$ with $\dim U_j = R_j$. Find smaller subspaces

$\hat{U}_j \subset U_j$, $\dim \hat{U}_j = r_j$, and the best approximation $\hat{\mathbf{u}}$ of $\mathbf{u}$ in $\hat{\mathbf{U}} := \bigotimes_j \hat{U}_j$. Obviously, $\hat{\mathbf{u}}$ is given by $\hat{\mathbf{u}} = \mathbf{P}\mathbf{u}$, where $\mathbf{P}$ is the orthogonal projection from $\mathbf{U}$ onto $\hat{\mathbf{U}}$. Furthermore, $\mathbf{P}$ is the Kronecker product[3] $P_1 \otimes \cdots \otimes P_d$ of the orthogonal projections $P_j : U_j \to \hat{U}_j$. Another expression for $\mathbf{P}$ is the product $\mathbf{P}_1 \cdot \ldots \cdot \mathbf{P}_d$, where $\mathbf{P}_j = I \otimes \cdots \otimes I \otimes P_j \otimes I \otimes \cdots \otimes I$.

The HOSVD bases defined above allow us to define an easy approximation procedure. We recall that $U_j = U_j^{\min}(\mathbf{u})$ is spanned by the HOSVD basis $\{\mathbf{b}_k^j : k = 1, \ldots, R_j\}$, where the basis vectors are ordered according to the size of the singular values. The smaller subspace is defined by $\hat{U}_j := \operatorname{span}\{\mathbf{b}_k^j : k = 1, \ldots, r_j\}$. The algorithmic realisation of the projection $\mathbf{P}$ from above is as follows. Represent $\mathbf{u}$ by

$$\mathbf{u} = \sum_{k_1=1}^{R_1} \cdots \sum_{k_d=1}^{R_d} \mathbf{c}[k_1, \ldots, k_d]\, \mathbf{b}_{k_1}^1 \otimes \cdots \otimes \mathbf{b}_{k_d}^d$$

with the HOSVD bases $\mathbf{b}^j := \operatorname{LSVD} \mathbf{M}^j(\mathbf{u})$. Then $\hat{\mathbf{u}} = \mathbf{P}\mathbf{u}$ is given by

$$\hat{\mathbf{u}} = \sum_{k_1=1}^{r_1} \cdots \sum_{k_d=1}^{r_d} \mathbf{c}[k_1, \ldots, k_d]\, \mathbf{b}_{k_1}^1 \otimes \cdots \otimes \mathbf{b}_{k_d}^d .$$

The error of this approximation can estimated by

$$\begin{aligned} \|\mathbf{u} - \hat{\mathbf{u}}\|^2 &= \| I - \mathbf{P}_1 \cdots \mathbf{P}_d \mathbf{u}\|^2 = \|\big((I - \mathbf{P}_1) + \mathbf{P}_1(I - \mathbf{P}_2) + \ldots\big)\mathbf{u}\|^2 \\ &\le \sum_{j=1}^{d} \|(I - \mathbf{P}_j)\mathbf{u}\|^2 = \sum_{j=1}^{d} \varepsilon_j^2 \le d \inf_{\mathbf{w} \in \mathcal{T}_{\le \mathbf{r}}} \|\mathbf{u} - \mathbf{w}\|^2 , \end{aligned} \tag{12.7}$$

where $\varepsilon_j^2 = \sum_{k_j > r_j} (\sigma^j[k_j])^2$. The infimum denotes the error of the best rank $\mathbf{r}$ approximation. In contrast to $d = 2$, where the SVD provides the best approximation (in ℓ_2-norm), we obtain only a quasi-optimal error estimate for the HOSVD approximation if $d > 2$ [5]. A recent result [26] states that finding the best rank $\mathbf{r}$ approximation in $d > 2$ is in general an NP hard problem. This negative results also holds for rank-1 approximation, i.e. $\mathbf{r} = (1, \ldots, 1)$.

12.3 Hierarchical Tensor Representations

The subspace concept introduced above can be interpreted as a generalisation of SVD to higher dimensions $d > 2$. It enjoys many important properties, but it does not prevent exponential scaling of the storage of the entries $\mathbf{c}[k_1, \ldots, k_d]$. The

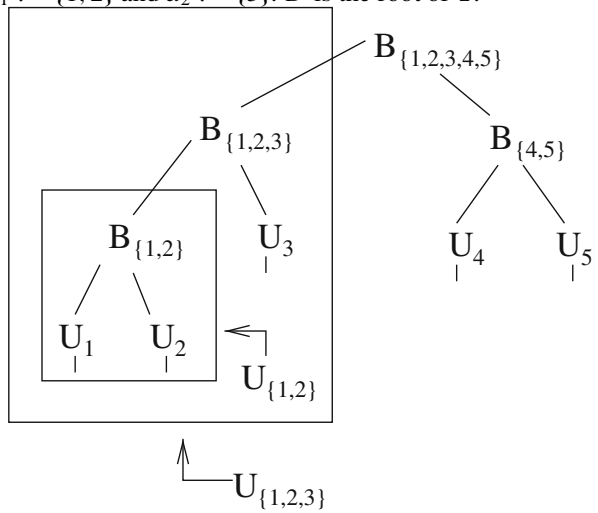
[3] The application of the Kronecker product $P_1 \otimes \cdots \otimes P_d$—as defined in the introduction—to an elementary tensor $v^1 \otimes \cdots \otimes v^d$ is given by $P_1 v^1 \otimes \cdots \otimes P_d v^d$.

reduction of degrees of freedom results only from the fact that n_j is replaced by $r_j \le n_j$. In particular, for $n_j = 2$ this concept is not helpful. The *Hierarchical Tucker format* (HT) in the form introduced in [24] extends the above idea of subspace approximation into a hierarchical or multi-level framework. Already earlier this tensor format has been proposed in the multi-configurational Hartree model [39,51] as well as in terms of tree tensor network states [49]. Following [20], we proceed in a hierarchical way.

We may start with defining subspaces $U_j \subset V_j$, $j = 1, \dots, d$, as introduced for the Tucker representation $U_j = \mathrm{span}\{\mathbf{b}_k^j : k = 1 \dots, r_j\}$.

For the representation of $\mathbf{u}$, we do not need, in general, the whole partial tensor space $U_1 \otimes U_2$, but only a subspace $U_{\{1,2\}} \subset U_1 \otimes U_2$ of dimension $r_{\{1,2\}} \le r_1 r_2$. We may continue, e.g. by building a subspace $U_{\{1,2,3\}} \subset U_{\{1,2\}} \otimes U_3 \subset U_1 \otimes U_2 \otimes U_3$, or $U_{\{1,2,3,4\}} \subset U_{\{1,2\}} \otimes U_{\{3,4\}}$ etc.

This approach can be cast into the framework described by a partition tree $\mathbb{T}$ with leaves $\{1\}, \dots \{d\}$, simply abbreviated by $1, \dots, d$, and vertices $\alpha \subset D := \{1, \dots, d\}$ corresponding to the partitions $\alpha = \alpha_1 \dot\cup \alpha_2$, e.g. $\alpha = \{1,2,3\} = \{1,2\} \cup \{3\}$, where $\alpha_1 := \{1,2\}$ and $\alpha_2 := \{3\}$. D is the root of $\mathbb{T}$.



Let $\alpha_1, \alpha_2 \subset D$ be the two sons of $\alpha \subset D$, then the condition $U_\alpha \subset U_{\alpha_1} \otimes U_{\alpha_2}$ implies that U_α has basis vectors of the form

$$\mathbf{b}_{k_\alpha}^{\alpha} = \sum_{k_{\alpha_1}=1}^{r_{\alpha_1}} \sum_{k_{\alpha_2}=1}^{r_{\alpha_2}} \mathbf{u}^{\alpha}[k_\alpha, k_{\alpha_1}, k_{\alpha_2}]\, \mathbf{b}_{k_{\alpha_1}}^{\alpha_1} \otimes \mathbf{b}_{k_{\alpha_2}}^{\alpha_2} \,. \tag{12.8}$$

The tensors $\mathbf{u}^\alpha \in X_\alpha := \mathbb{R}^{r_\alpha} \otimes \mathbb{R}^{r_{\alpha_1}} \otimes \mathbb{R}^{r_{\alpha_R}}$ are called *transfer tensors*. Finally, the given tensor $\mathbf{u}$ is determined by $\mathbf{u} = \sum_{k=1}^{r_D} \mathbf{c}_k^D \mathbf{b}_k^D$. Usually, $r_D = 1$ holds for the root $\alpha = D$ since $U_D := \mathrm{span}(\mathbf{v})$ is a possible choice.

Given a partition tree $\mathbb{T}$, the tensor $\mathbf{u}$ is completely defined by the transfer tensors $\mathbf{u}^\alpha$ ($\#\alpha > 1$), the leaf bases $\mathbf{u}^{\{j\}}$ ($1 \le j \le d$) and the coefficient vector $\mathbf{c}^D \in \mathbb{R}^{r_D}$. Indeed $\mathbf{u} \in \mathbb{R}^{\mathbf{I}}$ can be reconstructed by applying (12.8) recursively. The reconstruction of the tensor from its components defines a multilinear mapping

$$\tau : X_{\mathbb{T}} := (\times_{\alpha\in\mathbb{T}} X_\alpha) \times \mathbb{R}^{r_D} \to \mathbb{R}^{\mathbf{I}}, \quad \mathbf{u} = \tau((\mathbf{u}^\alpha)_{\alpha\in\mathbb{T}}, \mathbf{c}^D)\,. \tag{12.9}$$

Note that the data size of the parametrisation is $\dim X_{\mathbb{T}} = \mathcal{O}(r^3 + nrd)$, where $r := \max\{r_\alpha : \alpha \in \mathbb{T}\}$, and $n = \max_j \dim V_j$.

The minimal ranks r_α, $\alpha \in \mathbb{T}\backslash\{D\}$, are characterised by the matricisation $\mathbf{M}^\alpha(\mathbf{u})$ of the tensor $\mathbf{u}$. For a general subset $\alpha \subset D$, the entries of $\mathbf{M}^\alpha = \mathbf{M}^\alpha(\mathbf{u})$ are defined by $\mathbf{M}^\alpha[\mathbf{x}_\alpha, \mathbf{x}_{D\backslash\alpha}] = \mathbf{u}[x_1, \ldots, x_d]$, where $\mathbf{x}_\alpha$ is the index tuple $(x_k : k \in \alpha)$. Then $r_\alpha = \operatorname{rank}_\alpha(\mathbf{u}) := \dim \mathbf{M}^\alpha$ holds as in the Tucker case. Similarly, there are minimal subspaces $U_\alpha^{\min}(\mathbf{u})$, whose bases can be used in (12.8), provided that the relations (12.8) hold with minimal ranks r_α.

The set of all tensors with fixed ranks r_α is denoted by

$$\mathcal{M}_{\mathbf{r}} := \{\mathbf{u} : r_\alpha = \operatorname{rank}_\alpha(\mathbf{u}) \text{ for all } \alpha \in \mathbb{T}\}\,, \text{ where } \mathbf{r} = (r_\alpha)_{\alpha\in\mathbb{T}}.$$

Similarly, $\mathcal{M}_{\le\mathbf{r}}$ is defined by $r_\alpha \le \operatorname{rank}_\alpha(\mathbf{u})$.

We remark that the bases in (12.8) can be constructed to be orthonormal. Furthermore, the concept can be easily extended to a set of tensors $\mathbf{u}_k$, $k = 1, \ldots, K$. In the latter case, r_D may increase to K.

Given a tensor space $\bigotimes_{j=1}^d V_j$, there are various possibilities to build partition trees $\mathbb{T} = \mathbb{T}_{\mathbf{I}}$. The involved ranks $\{r_\alpha : \alpha \in \mathbb{T}\}$ are a subset of all ranks $\{r_\alpha : \alpha \subset D\}$ and the data size $\dim X_{\mathbb{T}}$ depends severely on the choice of $\mathbb{T}$. The appropriate tree $\mathbb{T}_{\mathbf{I}}$ depends on the individual tensor $\mathbf{u}$. Ballani–Grasedyck [3] have developed helpful methods improving the partitioning of trees.

12.3.1 Matrix Product Representation

We highlight a particular case, namely *tensor trains*. This *tensor train (TT) format* corresponds to an unbalanced tree with nodes $\alpha = \{1, \ldots, j\}$, $j = 1, \ldots, d-1$, and root $\alpha_D = \{1, \ldots, d\}$, where one chooses $U_{\{j\}} := V_j$ and $U_{\{1,\ldots,j+1\}} \subset U_{\{1,\ldots,j\}} \otimes U_{\{j+1\}}$. Throughout this chapter let us abbreviate $\alpha = \{1, \ldots, j\}$ simply by $\alpha := j$, without any ambiguity. The TT tensors or tensor trains were developed independently by the authors in [41, 43]. Later it turned out that this tensor representation has been known in quantum physics as *matrix product states*, see e.g. [49] for a recent survey. The transfer tensors $\mathbf{u}^{\{1,2,\ldots,j\}} =: \mathbf{u}_j$ are then of the form $\mathbf{u}_j \in \mathbb{R}^{r_{j-1}\times n_j \times r_j}$. Applying the recursive construction introduced above, the tensor can be written entrywise by

$$\mathbf{u}[x_1,\dots,x_d] = \sum_{k_1=1}^{r_1} \dots \sum_{k_{d-1}=1}^{r_{d-1}} \mathbf{u}_1[x_1,k_1]\mathbf{u}_2[k_1,x_2,k_2]\dots\mathbf{u}_d[k_{d-1},x_d].$$

Defining $r_0 = r_d := 1$ and introducing matrices $\mathbf{U}_j[x_j] \in \mathbb{R}^{r_{j-1}\times r_j}$ by

$$\big(\mathbf{U}_j[x_j]\big)_{k_{j-1},k_j} = \mathbf{u}_j[k_{j-1},x_j,k_j], \quad 1 \le j \le d\ ,$$

we can represent the tensor entries by matrix products

$$\mathbf{u}[\mathbf{x}] = \mathbf{u}[x_1,\dots,x_d] = \mathbf{U}_1[x_1]\cdots\mathbf{U}_j[x_j]\cdots\mathbf{U}_d[x_d].$$

As usual, orthogonal basis vectors are the preferred choice.

12.3.2 Approximation and Tensor Compression

The HOSVD can also be used for approximating tensors by hierarchical tensors of lower rank. For a brief presentation, we start from an HT representation of a tensor $\mathbf{u}$ by subspaces U_α of dimension $r_\alpha = \operatorname{rank}_\alpha(\mathbf{u})$ and seek for smaller subspaces $\hat{U}_\alpha \subset U_\alpha$ of given dimension $\hat{r}_\alpha = \dim \hat{U}_\alpha < \dim U_\alpha = r_\alpha$. For all vertices α with sons α_1, α_2, we have to require $\hat{U}_\alpha \subset \hat{U}_{\alpha_1} \otimes \hat{U}_{\alpha_2}$.

First we describe the truncation procedure from a theoretical point of view. For all $\alpha \in \mathbb{T}$ we determine the HOSVD basis $\mathbf{b}^\alpha = \operatorname{LSVD}\mathbf{M}^\alpha(\mathbf{u})$ together with the singular value $\sigma^\alpha[1] \ge \dots \ge \sigma^\alpha[r_\alpha] > 0$. Set $U'_\alpha := \operatorname{span}\{\mathbf{b}^\alpha_k : 1 \le k \le \hat{r}_\alpha\}$. Define P_α as the orthogonal projection from $V_\alpha := \bigotimes_{j\in\alpha} V_j$ onto U'_α and set $\mathbf{P}_\alpha := P_\alpha \otimes I_{\alpha^c} : \bigotimes_{j=1}^d V_j \to U'_\alpha \otimes V_{\alpha^c}$, where $\alpha^c := D\setminus\alpha$ is the complement of α. Then the product $\mathbf{P} := \mathbf{P}_{\alpha_1}\mathbf{P}_{\alpha_2}\cdots$ runs over all $\alpha \in \mathbb{T}$ and is ordered so that a father α_i and its son α_j have indices $i > j$. Then $\hat{\mathbf{u}} := \mathbf{P}\mathbf{u}$ is the desired truncated tensor.

The basis vectors $\mathbf{b}^\alpha_k$ do not appear in the algorithm, but only the transfer tensors $\mathbf{u}^\alpha$. We assume that the basis $\{\mathbf{b}^\alpha_k\}_k$ is orthonormal, otherwise it can be orthonormalised. Then the transformation from the orthonormal bases into the HOSVD bases can be performed by $\mathcal{O}(r^4 + nr^2d)$ operations (cf. [20]). This yields new transfer tensors which again are denoted by $\mathbf{u}^\alpha \in \mathbb{R}^{r_\alpha} \otimes \mathbb{R}^{r_{\alpha_1}} \otimes \mathbb{R}^{r_{\alpha_2}}$ (α_1, α_2 sons). The new quantities $\hat{\mathbf{u}}^\alpha \in \mathbb{R}^{\hat{r}_\alpha} \otimes \mathbb{R}^{\hat{r}_{\alpha_1}} \otimes \mathbb{R}^{\hat{r}_{\alpha_2}}$ corresponding to $\hat{\mathbf{u}} = \mathbf{P}\mathbf{u}$ are obtained by restriction to the smaller size, i.e. $\hat{\mathbf{u}}^\alpha[k_\alpha,k_{\alpha_1},k_{\alpha_2}] := \mathbf{u}^\alpha[k_\alpha,k_{\alpha_1},k_{\alpha_2}]$ for all $1 \le k_\alpha \le \hat{r}_\alpha$, $1 \le k_{\alpha_1} \le \hat{r}_{\alpha_1}$, $1 \le k_{\alpha_2} \le \hat{r}_{\alpha_2}$. The truncation error is similar as in (12.7) (cf. [14, 24]).

Theorem 12.1 (quasi-optimality). *The approximation* $\hat{\mathbf{u}}$ *leads to the error*

$$\inf_{\mathbf{v}\in\mathcal{M}_\mathbf{r}} \|\mathbf{v}-\mathbf{u}\| \le \|\hat{\mathbf{u}}-\mathbf{u}\| \le \sqrt{\sum_\alpha \sum_{k_\alpha > \hat{r}_\alpha} (\sigma^\alpha[k_\alpha])^2} \le \sqrt{2d-3} \inf_{\mathbf{v}\in\mathcal{M}_\mathbf{r}} \|\mathbf{v}-\mathbf{u}\|.$$

This result implies the quasi-optimality of HOSVD.

The quantity $2d-3$ is related to the number of orthogonal projections. If $\hat{r}_\alpha = r_\alpha$ for some $\alpha \in \mathbb{T}$, no projection is needed. Since the later TT format avoids proper subspaces of V_j, the bound becomes $\sqrt{d-1}$.

The described truncation is characterised by first computing the HOSVD bases and then performing the truncation. Another variant starts from the root $\alpha = D$, determines the HOSVD basis in α and performs the truncation at α, before the same procedure is repeated at the sons of α (cf. [20]).

Instead of fixing the ranks $\hat{r}_\alpha$, we can prescribe a tolerance $\varepsilon > 0$ and choose $\hat{r}_\alpha$ such that $[\sum_\alpha \sum_{k_\alpha > \hat{r}_\alpha} (\sigma^\alpha[k_\alpha])^2]^{1/2} \le \varepsilon$. This adaptive procedure allows us to perform operations with exact error bounds.

The theorem above can be used to characterise classes for which the best rank $\mathbf{r}$ approximation converges with an algebraic rate. Consider the tensor space $\mathcal{H}^d = \bigotimes_{j=1}^d L^2(I_j)$ for $I_j \subset \mathbb{R}$. A completion w.r.t. the $L^2(\mathbf{I})$ norm with $\mathbf{I} = I_1 \times \ldots \times I_d$ yields $\mathcal{H}^d = L^2(\mathbf{I})$. Let $\mathbf{u} \in \mathcal{H}^d$. Because of the infinite-dimensional setting, $\mathbf{M}^\alpha(\mathbf{u})$ ($\alpha \in \mathbb{T}$) is not a matrix, but a Hilbert–Schmidt operator with an infinite singular value decomposition. In particular, there is a sequence $\boldsymbol{\sigma}^\alpha := (\sigma^\alpha[k])_{k\in\mathbb{N}}$ of singular values. Theoretically, the HOSVD truncation of tensors is defined as in the finite-dimensional case. The only difference is that we replace the rank $r_\alpha = \infty$ by a finite rank $\hat{r}_\alpha < \infty$.

The decay behaviour of $\boldsymbol{\sigma}^\alpha$ can be quantified by introducing the *Schatten classes* $L_{*,p}$ ($0 < p < 2$) given by

$$\|\mathbf{M}_\alpha(\mathbf{u})\|_{*,p} := \|\boldsymbol{\sigma}_\alpha\|_{\ell_p} = \Big[\sum_{k\in\mathbb{N}} (\sigma^\alpha[k])^p\Big]^{1/p} \text{ and } \|\mathbf{u}\|_{*,p} = \sup_{\alpha\in\mathbb{T}\setminus\{D\}} \|\mathbf{M}^\alpha(\mathbf{u})\|_{*,p}.$$

For $\mathbf{u} \in L_{*,p}$ we can estimate the error of the HOSVD truncation.

Theorem 12.2 ([47]). *Let $\mathbf{u} \in L_{*,p}$ for $p < 2$. Then $\mathbf{u} \in \mathcal{H}^d$ can be approximated by a tensor $\hat{\mathbf{u}}$ of multi-linear rank $\mathbf{r} = (r_\alpha)_{\alpha\in\mathbb{T}}$ with an error bound*

$$\|\mathbf{u} - \hat{\mathbf{u}}\| \le C\big(\min\{r_\alpha : \alpha \in \mathbb{T}\}\big)^{-\tau} \sqrt{d}\, \|\mathbf{u}\|_{*,p} \quad \textit{with } \tau = \frac{1}{p} - \frac{1}{2}.$$

It has been shown that e.g. mixed Sobolev classes are contained in the Schatten classes [47]. For a more precise formulation and a discussion concerning the required degrees of freedom we refer to [47].

The HOSVD truncation ensures an error bound w.r.t. the Euclidean norm or, more precisely, with respect to the canonical scalar product induced by the scalar products of the spaces V_j. In the case of $V_j = L^2(I_j)$, this is the $L^2(\mathbf{I})$ norm on $\mathbf{I} = I_1 \times \ldots \times I_d$ (as, e.g. in Theorem 12.2). The arising L^2 estimate makes sense as long as the truncated tensor $\hat{\mathbf{u}}$ is used, e.g., in a scalar product. However, if we want to evaluate the function $\hat{\mathbf{u}}$ at a certain point $\mathbf{x} \in \mathbf{I}$, we need an L^∞

estimate of the error $\delta\mathbf{u} := \hat{\mathbf{u}} - \mathbf{u}$, which, unfortunately, cannot[4] be obtained from the L^2 estimate. Nevertheless, $\delta\mathbf{u}$ behaves well also with respect to stronger norms, provided that $\mathbf{u}$ is smooth. The Sobolev semi-norm $|\cdot|_m$ can be defined by $[\sum_{j=1}^{d} \|\partial^m \mathbf{u}/\partial^m x_j\|^2_{L^2(\mathbf{I})}]^{1/2}$ involving only uni-directional derivatives. Now the smoothness of f may be characterised by $|f|_m \le C$. Then this smoothness is inherited by the HOSVD approximation $\hat{\mathbf{u}}$; more precisely, also $|\hat{\mathbf{u}}|_m \le C$ is valid. In particular, $|\delta\mathbf{u}|_m \le 2\,|\mathbf{u}|_m$ describes that also the error is smooth. Stronger norms than L^2 can be estimated by interpolation inequalities, the so-called Gagliardo–Nirenberg inequalities. If $m > d/2$, the uniform norm for $\mathbf{I} = \mathbb{R}^d$ or $\mathbf{I} = [0,\infty)^d$ is bounded by

$$\|\delta\mathbf{u}\|_{L^\infty} \le c_m\, |\mathbf{u}|_m^{d/(2m)}\, |\delta\mathbf{u}|_{L^2}^{1-d/(2m)}.$$

For a proof, more details and generalisations we refer to [21].

12.4 Hierarchical Tensors as Differentiable Manifolds

The central aim of this chapter is to remove the redundancy in the parametrisation of our admissible set $\mathscr{M}_{\mathbf{r}}$ (the set of tensors of given rank tuple $\mathbf{r}$).

Redundancy in the parametrisation can cause serious difficulties in optimisation. In particular, for model reduction in dynamical systems it should be avoided completely. For example, the matrix product representation in the TT format is not unique. The same holds for the general hierarchical tensor representation, where basis transformations change the parameters of the representation, but not the represented tensor. In the case of a TT tensor, basis transformations are described by regular matrices $\mathbf{G}_j$ in

$$\mathbf{u}[\mathbf{x}] = \mathbf{U}_1[x_1]\mathbf{G}_1\mathbf{G}_1^{-1}\mathbf{U}_2[x_2]\mathbf{G}_2\mathbf{G}_2^{-1}\cdots\mathbf{G}_{d-1}\mathbf{G}_{d-1}^{-1}\mathbf{U}_d[x_d] = \tilde{\mathbf{U}}_1[x_1]\cdots\tilde{\mathbf{U}}_d[x_d]$$

yielding two different representations of the same tensor $\mathbf{u}$ by either $\mathbf{U}_j$ or $\tilde{\mathbf{U}}_j$.

Let us consider the space of parameters $X_{\mathbb{T}}$ (cf. (12.9)) and a single vector $\mathscr{U} := ((\mathbf{u}^\alpha)_{\alpha\in\mathbb{T}}, \mathbf{c}^D) \in X_{\mathbb{T}}$ from the parametrisation space. We define the action of $\mathscr{G} := (\mathbf{G}_\alpha)_{\alpha\in\mathbb{T}}$ on $\mathscr{U}$ by $\mathscr{G}\mathscr{U} = \hat{\mathscr{U}} := ((\hat{\mathbf{u}}^\alpha)_{\alpha\in\mathbb{T}}, \hat{\mathbf{c}}^D)$, for $\alpha \in \mathbb{T}$

$$\hat{\mathbf{u}}^\alpha[k_\alpha, k_{\alpha_1}, k_{\alpha_2}] := \sum_{j_\alpha, j_{\alpha_1}, j_{\alpha_2}} \mathbf{u}_\alpha[j_\alpha, j_{\alpha_1}, j_{\alpha_2}]\mathbf{G}_\alpha[j_\alpha, k_\alpha]\mathbf{G}_{\alpha_1}^{-1}[j_{\alpha_1}, k_{\alpha_1}]\mathbf{G}_{\alpha_2}^{-1}[j_{\alpha_2}, k_{\alpha_2}]$$

where α_1, α_2 are the sons of α. On the leaves $\alpha = \{j\}$ we define

[4] Also in the finite-dimensional case, where all norms are equivalent, the corresponding equivalence constant is too huge for practical purposes.

$$\hat{\mathbf{u}}^{\alpha}[k_\alpha, x_\alpha] := \sum\nolimits_{j_\alpha=1}^{r_\alpha} \mathbf{u}_\alpha[j_\alpha, x_\alpha]\mathbf{G}_\alpha[j_\alpha, k_\alpha].$$

We set $\hat{\mathbf{c}}^{D}[k_D] := \sum_{j_D=1}^{r_D} \hat{\mathbf{c}}^{D}[j_D]\mathbf{G}_D^{-1}[j_D, k_D]$ at the root $\alpha = D$ with sons α_1, α_2. Note that the set $\mathscr{G}$ of all operations $\mathbf{G}_\alpha$ constitutes a group. Having observed that the tensor $\mathbf{u}$ remains fixed under this transformation of the component tensors, we identify two representations $\mathscr{U}_1$ and $\mathscr{U}_2$, if there exists G such that $\mathscr{U}_2 = G\mathscr{U}_1$,

$$\mathscr{U}_1 \sim \mathscr{U}_2 \text{ if and only if there exits } G \in \mathscr{G} \text{ with } \mathscr{U}_2 = G\mathscr{U}_1 \,.$$

Standard differential geometry asserts that the equivalence classes $[\mathscr{U}] := \{\mathscr{U}_1 \sim \mathscr{U}\}$ form a smooth embedded submanifold in $X_{\mathbb{T}}$. Moreover, the quotient manifold $X_{\mathbb{T}}/\mathscr{G}$ is isomorphic to an embedded submanifold in the parameter space. Since all representations in the parameter space define the same tensor, $X_{\mathbb{T}}/\mathscr{G}$ is isomorphic to an embedded submanifold $\mathscr{M}_{\mathbf{r}} \subset \mathbb{R}^{\mathbf{I}}$ in the ambient tensor space $\mathbb{R}^{\mathbf{I}}$ (cf. [27, 39, 40, 50]).

Having this principle in mind, several authors have considered the differential geometry of rank-r matrices, e.g. [1, 8, 33] among many others, and of tensors from the Tucker set $\mathscr{T}_{\mathbf{r}}$ from (12.3) (cf. [33]). It can be extended to hierarchical tensor representations, following the philosophy explained in the previous chapters. Next, we present important results about the construction of tangent spaces, because they are required for computations.

For practical computations we need that the *tangent space* $\mathscr{T}_{\mathbf{u}}$ at $\mathbf{u} \in \mathscr{M}_{\mathbf{r}}$, i.e. the space of all tangent directions, can be computed by the Leibniz rule.

For example consider the curve $\mathbf{u}(t) \in \mathscr{M}_{\mathbf{r}}$ with $\mathbf{u}(0) =: \mathbf{u}$ in the tensor train format. Then $\delta\mathbf{u} := \frac{d}{dt}\mathbf{u}(0)$ is of the form

$$\delta\mathbf{u}[\mathbf{x}] = \delta U_1[x_1]U_2[x_2]\cdots U_d[x_d] + \ \ldots\ + U_1[x_1]\cdots U_{d-1}[x_{d-1}]\delta U_d[x_d] \in \mathscr{T}_{\mathbf{u}}\,.$$

Due to redundancy, the δU_j, $j = 1, \ldots, d$, are not uniquely defined. In particular, the δU_j, $j = 1, \ldots, d-1$, can be defined uniquely by imposing *gauge conditions* [27] $\delta U_j \perp U_j$, $j = 1, \ldots, d-1$, in the sense

$$\sum_{k_{j-1}=1}^{r_{j-1}} \sum_{x_j=1}^{n_j} U_j[k_{j-1}, x_j, k_j]\delta U_j[k_{j-1}, x_j, k'_j] = 0 \quad \text{for all } k_j, k'_j \in I_j\,.$$

We notice that the root tensor $U_d \in \mathbb{R}^{r_{d-1}\times n_d}$ is from the manifold of full rank (i.e. rank r_{d-1}) matrices, which is an open and dense subset in $\mathbb{R}^{r_{d-1}\times n_d}$.

For general hierarchical tensors associated to a tree $\mathbb{T}$, we start with $U_\alpha \subset U_{\alpha_1} \otimes U_{\alpha_2}$ of dimension r_α. Then there exists a complementary space Y_α with $Y_\alpha \oplus U_\alpha = U_{\alpha_1} \otimes U_{\alpha_2}$. Y_α is spanned by tensors

$$\mathbf{y}^{\alpha}_{j_\alpha} = \sum\nolimits_{k_{\alpha_1}} \sum\nolimits_{k_{\alpha_2}} \delta\mathbf{u}^{\alpha}[j_\alpha, k_{\alpha_1}, k_{\alpha_2}]\mathbf{b}^{\alpha_1}_{k_{\alpha_1}} \otimes \mathbf{b}^{\alpha_2}_{k_{\alpha_2}}$$

satisfying $\mathbf{y}^\alpha_{k_\alpha} \perp U_\alpha$. If all bases are chosen to be orthogonal, the $\delta\mathbf{u}^\alpha$ must satisfy the following condition:

$$\sum\nolimits_{k_{\alpha_1}} \sum\nolimits_{k_{\alpha_2}} \delta\mathbf{u}^\alpha[j_\alpha, k_{\alpha_1}, k_{\alpha_2}]\mathbf{u}^\alpha[k_\alpha, k_{\alpha_1}, k_{\alpha_2}] = 0 \quad \text{for all } k_\alpha, j_\alpha\,. \tag{12.10}$$

For general hierarchical tensors on a tree $\mathbb{T}$, let us define the subspace

$$W_\alpha := \{\delta\mathbf{u}^\alpha \in X_\alpha = \mathbb{R}^{r_\alpha \times r_{\alpha_1} \times r_{\alpha_2}} : \delta\mathbf{u}^\alpha \text{ satisfies (12.10)}\}\,, \tag{12.11}$$

provided that $\alpha \neq D$ is not a leaf. For $\alpha = D$ we set $W_D := X_D = \mathbb{R}^{r_{\alpha_1} \times r_{\alpha_2}}$, i.e. there is no gauge condition involved. For a leaf α, we need the gauge condition, and W_α is defined analogously to (12.11).

Given $\mathbf{u} \in \mathscr{M}_\mathbf{r}$ with components $\mathbf{u}^\alpha$ ($\alpha \in \mathbb{T}$), $\mathscr{U} = \left(\mathbf{u}^\alpha\right)_{\alpha\in\mathbb{T}} \in X_\mathbb{T}$ together with the parametrisation $\mathbf{u} = \tau(\mathscr{U})$, and $\mathbf{w}^\alpha \in \mathbb{R}^{r_\alpha \times r_{\alpha_1} \times r_{\alpha_2}}$ for some $\alpha \in \mathbb{T}$, we have $(\mathbf{u}^\beta)_{\beta\in\mathbb{T}} := \mathbf{X}_{\mathbf{w}^\alpha}(\mathscr{U})$, where the operator $\mathbf{X}_{\mathbf{w}^\alpha}(\mathscr{U})$ replaces $\mathbf{u}^\alpha$ by $\mathbf{w}^\alpha$ and leaves the other components unchanged, i.e. $\left(\mathbf{X}_{\mathbf{w}^\alpha}(\mathscr{U})\right)_\gamma = \mathbf{u}^\gamma$ for $\gamma \in \mathbb{T}\backslash\{\alpha\}$. Next we define the map $\mathbf{E}_\alpha$ extending the component tensor $\mathbf{w}^\alpha$ to the larger ambient space:

$$\mathbf{E}_\alpha : \mathbb{R}^{r_\alpha \times r_{\alpha_1} \times r_{\alpha_2}} \to \mathbb{R}^\mathbf{I}\,,\ \mathbf{E}_\alpha\mathbf{w}^\alpha = \tau(\mathbf{X}_{\mathbf{w}^\alpha}(\mathscr{U})) = \tau\left(\mathbf{X}_{\mathbf{w}^\alpha}\left((\mathbf{u}^\alpha)_{\alpha\in\mathbb{T}}\right)\right).$$

It turns out that a generic tensor $\delta\mathbf{u} \in \mathscr{T}_\mathbf{u}$ in the tangent space of $\mathbf{u} \in \mathscr{M}_\mathbf{r}$ is of the form

$$\delta\mathbf{u} = \sum\nolimits_{\alpha\in\mathbb{T}} \mathbf{E}^\alpha\,\delta\mathbf{u}^\alpha\,, \quad \delta\mathbf{u}^\alpha \in W_\alpha\,.$$

By the assumptions, the hierarchical rank of $\mathbf{u}$ is the tuple $\mathbf{r}$, the operators $\mathbf{E}_\alpha^\mathsf{T}\mathbf{E}_\alpha$, $\alpha \in \mathbb{T}\backslash\{D\}$, are invertible, and the bases $\mathbf{b}^\alpha$, $\alpha \neq D$, are orthonormal bases. Then,

$$\mathbf{E}_\alpha^+ := \left(\mathbf{E}_\alpha^\mathsf{T}\mathbf{E}_\alpha\right)^{-1}\mathbf{E}_\alpha^\mathsf{T} : \mathbf{R}^\mathbf{I} \to \mathbb{R}^{r_\alpha \times r_{\alpha_1} \times r_{\alpha_2}}$$

is a Moore–Penrose inverse of $\mathbf{E}_\alpha$. For the root $\alpha = D$, we obtain the identity: $\mathbf{E}_D^\mathsf{T}\mathbf{E}_D = \mathbf{I}$. (The case that α is a leaf can be formulated by obvious modifications.) Let $\Pi_{W_\alpha} : \mathbb{R}^{r_\alpha \times r_{\alpha_1} \times r_{\alpha_2}} \to W_\alpha$ be the orthogonal projection onto the component tensors satisfying the gauge condition. For the leaves we need obvious modifications. With these operators at hand, for $\mathbf{v} \in \mathbb{R}^\mathbf{I}$, the operator

$$P_{\mathscr{T}_\mathbf{u}}\mathbf{v} := \sum_{\alpha\in\mathbb{T}\backslash D} \mathbf{E}_\alpha\Pi_{W_\alpha}\mathbf{E}_\alpha^+\mathbf{v} + \mathbf{E}_D\mathbf{E}_D^\mathsf{T}\mathbf{v} =: \sum_{\alpha\in\mathbb{T}} \mathbf{E}_\alpha\,\delta\mathbf{u}^\alpha \ \in \mathscr{T}_\mathbf{u} \tag{12.12}$$

defines the orthogonal projection onto the tangent space $\mathscr{T}_\mathbf{u}$.

Remark: We have noticed that no gauge condition is imposed on the root component $\delta\mathbf{u}^D$, i.e. $W_D = X_D$. The gauge conditions (12.10) imply that the tensors $\mathbf{E}_\alpha\,\delta\mathbf{u}^\alpha$ are pairwise orthogonal. Furthermore, the tensor $\mathbf{u} = \mathbf{E}_D\mathbf{u}^D \in \mathscr{T}_\mathbf{u}$ itself is also included in the tangent space. Easily, it can be shown that a tangent

vector $\delta\mathbf{u}$ has a hierarchical rank $\mathbf{s}$ where $s_\alpha \leq 2r_\alpha$. Estimates of the Lipschitz continuity of $\mathbf{u} \mapsto P_{\mathscr{T}_\mathbf{u}}$ are upper bounds for the curvature at $\mathbf{u}$ and are given in [40]. The operators $\mathbf{E}_\alpha^\mathsf{T}$ are not difficult to implement, since they require only the computation of scalar product of tensors. Furthermore, the inverse $(\mathbf{E}_\alpha^\mathsf{T}\mathbf{E}_\alpha)^{-1}$ is applied only to the small parameter spaces W_α. This makes the projection onto the tangent space a flexible and efficient numerical tool, allowing the application of differential geometrical tools [1], see Sect. 12.5.

Remark 12.1 (closedness). The manifold $\mathscr{M}_\mathbf{r}$ is an open set. It has been shown in [12] that the closure of $\mathscr{M}_\mathbf{r}$ is $\mathscr{M}_{\leq\mathbf{r}}$, the set of all tensors with ranks $r'_i \leq r_i$.

This result is based on the observation that the matrix rank is an upper semi-continuous function [20]. For infinite-dimensional spaces it is important that even weak closedness holds. The singular points of the manifold $\mathscr{M}_\mathbf{r}$ are exactly those where $\mathbf{r}$ is not the actual rank. In the considerations above the complementary space $Y_\alpha \perp U_\alpha$ plays a crucial role. This concept can be easily extended to Hilbert spaces, but the arguments do not apply easily to Banach spaces. A deep functional analytic framework for the differential geometry of hierarchical tensors in Banach spaces has been developed in [13].

12.5 Numerical Methods

12.5.1 Formats and Representation

Here we describe how we can perform operations within the hierarchical tensor representation. Let $\mathbf{X} = \bigotimes_{j=1}^d X_j$, $\mathbf{Y} = \bigotimes_{j=1}^d Y_j$, and $\mathbf{Z} = \bigotimes_{j=1}^d Z_j$ be three tensor spaces. We consider binary operations $\diamond : \mathbf{X} \times \mathbf{Y} \to \mathbf{Z}$ with the property

$$\left(\bigotimes\nolimits_{j=1}^d x^j\right) \diamond \left(\bigotimes\nolimits_{j=1}^d y^j\right) = \bigotimes\nolimits_{j=1}^d \left(x^j \diamond y^j\right) \tag{12.13}$$

with equally denoted bilinear operations $\diamond : X_j \times Y_j \to Z_j$ for $1 \leq j \leq d$. Examples of such operations are the following ones.

1. *Hadamard product*. Set $\mathbf{X} = \mathbf{Y} = \mathbf{Z}$. The entry-wise multiplication for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^\mathbf{I}$ is defined by $(\mathbf{x} \circ \mathbf{y})[k_1, \ldots, k_d] := \mathbf{x}[k_1, \ldots, k_d]\mathbf{y}[k_1, \ldots, k_d]$. In the case of functions, it is the usual pointwise multiplication.
2. *Convolution*. For $X_j = \mathbb{R}^{n_j}$, $Y_j = \mathbb{R}^{m_j}$, and $Y_j = \mathbb{R}^{n_j+m_j-1}$ the convolution of two tensors $\mathbf{x} \in \mathbf{X}$ and $\mathbf{y} \in \mathbf{Y}$ is defined by $\mathbf{z} = \mathbf{x} \star \mathbf{y}$ with

$$\mathbf{z}[\mathbf{k}] := \sum\nolimits_\mathbf{l} \mathbf{x}[\mathbf{l}]\mathbf{y}[\mathbf{k} - \mathbf{l}], \tag{12.14}$$

where the sum is taken over all multi-indices $\mathbf{l} \in \mathbb{Z}^d$ so that $0 \le l_j \le n_j - 1$, $0 \le k_j - l_j \le m_j - 1$, while $\mathbf{k}$ satisfies $0 \le k_j \le n_j + m_j - 1$. Other variants use periodicity. Also for function spaces, the convolution is well known.

3. *Matrix-vector multiplication*. Matrix spaces $X_j = \mathbb{R}^{n_j \times m_j}$ lead to the space $\mathbf{X}$ of Kronecker matrices mapping $\mathbf{Y}$ onto $\mathbf{Z}$, where $Y_j = \mathbb{R}^{m_j}$, $Z_j = \mathbb{R}^{n_j}$.
4. *Matrix-matrix multiplication*. Set $X_j = \mathbb{R}^{n_j \times m_j}$, $Y_j = \mathbb{R}^{m_j \times \ell_j}$, $Z_j = \mathbb{R}^{n_j \times \ell_j}$.
5. *Scalar product*. Also the scalar product falls into this category, if we set $X_j = Y_j$ and $Z_j = \mathbb{R}$ (note that $\bigotimes_{j=1}^d \mathbb{R}$ can be identified with $\mathbb{R}$), but for these trivial spaces the algorithm can be implemented more directly.

We require the operations $\diamond : X_j \times Y_j \to Z_j$ to be practically implemented. Let two tensors $\mathbf{x} \in \mathscr{M}_{\mathbf{r}}(\mathbf{X})$ and $\mathbf{y} \in \mathscr{M}_{\mathbf{s}}(\mathbf{Y})$ be given in the hierarchical format with the same tree $\mathbb{T}$ and with the respective rank tuples $\mathbf{r}$ and $\mathbf{s}$. Their representation at the root $\alpha = D$ is given by $\mathbf{x} = \sum_k \mathbf{c}^D_{x,k} \mathbf{b}^D_{x,k}$ and $\mathbf{y} = \sum_\ell \mathbf{c}^D_{y,\ell} \mathbf{b}^D_{y,\ell}$ (the additional indices x or y refer to the two parameter tuples of $\mathbf{x} \in \mathscr{M}_{\mathbf{r}}(\mathbf{X})$ and $\mathbf{y} \in \mathscr{M}_{\mathbf{s}}(\mathbf{Y})$). By linearity, $\mathbf{x} \diamond \mathbf{y}$ is determined if all expressions $\mathbf{b}^D_{x,k} \diamond \mathbf{b}^D_{y,\ell}$ can be determined. For any $\alpha \in \mathbb{T}$ with sons α_1 and α_2 (in particular for $\alpha = D$ from above) the characteristic relation (12.8) shows that

$$\mathbf{b}^\alpha_{x,k} \diamond \mathbf{b}^\alpha_{y,\ell} = \left[\sum_{k_{\alpha_1}=1}^{r_{\alpha_1}} \sum_{k_{\alpha_2}=1}^{r_{\alpha_2}} \mathbf{u}^\alpha_x[k_\alpha, k_{\alpha_1}, k_{\alpha_2}]\, \mathbf{b}^{\alpha_1}_{x,k_{\alpha_1}} \otimes \mathbf{b}^{\alpha_2}_{x,k_{\alpha_2}} \right] \diamond \left[\sum_{\ell_{\alpha_1}=1}^{r_{\alpha_1}} \sum_{\ell_{\alpha_2}=1}^{r_{\alpha_2}} \mathbf{u}^\alpha_y[\ell_\alpha, \ell_{\alpha_1}, \ell_{\alpha_2}]\, \mathbf{b}^{\alpha_1}_{y,\ell_{\alpha_1}} \otimes \mathbf{b}^{\alpha_2}_{y,\ell_{\alpha_2}} \right]$$

$$= \sum \mathbf{u}^\alpha_x[k_\alpha, k_{\alpha_1}, k_{\alpha_2}] \mathbf{u}^\alpha_y[\ell_\alpha, \ell_{\alpha_1}, \ell_{\alpha_2}] \left(\mathbf{b}^{\alpha_1}_{x,k_{\alpha_1}} \diamond \mathbf{b}^{\alpha_1}_{y,\ell_{\alpha_1}} \right) \otimes \left(\mathbf{b}^{\alpha_2}_{x,k_{\alpha_2}} \diamond \mathbf{b}^{\alpha_2}_{y,\ell_{\alpha_2}} \right)$$

with summation over $k_{\alpha_1}, k_{\alpha_2}, \ell_{\alpha_1}, \ell_{\alpha_2}$. Define $\boldsymbol{\beta}^\alpha$ as the tuple of all $\mathbf{b}^\alpha_{x,k} \diamond \mathbf{b}^\alpha_{y,\ell}$. The index set consists of all pairs (k, ℓ) and has the size $r^\alpha_z := r^\alpha_x r^\alpha_y$. Note that $\boldsymbol{\beta}^\alpha$ is not explicitly evaluated. Instead, we need transfer tensors with respect to a frame. Here we choose $\boldsymbol{\beta}^\alpha$ as a frame spanning the involved subspace (note that the vectors $\mathbf{b}^\alpha_{x,k} \diamond \mathbf{b}^\alpha_{y,\ell}$ need not be linearly independent). The previous equation is only needed to obtain the definition of the transfer tensors with respect to the frames $\boldsymbol{\beta}^\alpha, \boldsymbol{\beta}^{\alpha_1}, \boldsymbol{\beta}^{\alpha_2}$:

$$\mathbf{u}^\alpha_y[(k_\alpha, \ell_\alpha), (k_{\alpha_1}, \ell_{\alpha_1}), (k_{\alpha_2}, \ell_{\alpha_2})] := \mathbf{u}^\alpha_x[k_\alpha, k_{\alpha_1}, k_{\alpha_2}] \mathbf{u}^\alpha_y[\ell_\alpha, \ell_{\alpha_1}, \ell_{\alpha_2}].$$

By assumption, we are able to evaluate the products $\boldsymbol{\beta}^\alpha$ for all leaves $\alpha = \{j\}$. In a next step, the frames are transformed into orthonormal bases. In the case of a proper frame, this procedure leads to a smaller representation rank r^α_z. Second, a HOSVD truncation can be applied to remove the negligible components.

A similar approach is used to compute the sum $\mathbf{u} := \mathbf{v} + \mathbf{w}$ of $\mathbf{v} \in \mathscr{M}_{\mathbf{r}_v}$ and $\mathbf{w} \in \mathscr{M}_{\mathbf{r}_w}$. Combining the bases $\mathbf{b}^\alpha_v$ and $\mathbf{b}^\alpha_w$ we obtain a frame $\mathbf{b}^\alpha_u$ of size $r^\alpha_u = r^\alpha_v + r^\alpha_w$. Now both $\mathbf{v}$ and $\mathbf{w}$ can be represented simultaneously. In particular, $\mathbf{v} = \sum_k \mathbf{c}^D_{v,k} \mathbf{b}^D_{u,k}$

and $\mathbf{w} = \sum_k \mathbf{c}^D_{w,k}\mathbf{b}^D_{u,k}$ implies $\mathbf{u} = \sum_k (\mathbf{c}^D_{v,k} + \mathbf{c}^D_{w,k})\mathbf{b}^D_{u,k}$, i.e. $\mathbf{c}^D_{u,k} := \mathbf{c}^D_{v,k} + \mathbf{c}^D_{w,k}$. As above, an orthonormalisation and truncation can follow. The precise cost of the various operations is described in [20, §13]. Concerning software realising these operations we refer to [35, 41].

12.5.2 Iterative Thresholding

For simplicity of exposition, let us consider a numerical scheme which produces a sequence of tensor $\mathbf{u}_n$, $n \in \mathbb{N}$, defined by

$$\mathbf{u}_{n+1} := \mathbf{u}_n + \mathbf{F}_n(\mathbf{u}_n) \in \mathbb{R}^{\mathbf{I}}.$$

This sequence may come from an (explicit) time stepping, from an iterative scheme solving a linear or even nonlinear equation. For example it may be a gradient or a Newton iteration. Suppose that the tensors are represented by some tensor format, e.g. $\mathbf{u}_n \in \mathcal{M}_{\le \mathbf{r}_n}$. As seen above, all operations let the representation rank increase. Therefore we have to expect $\mathbf{u}_{n+1} \in \mathcal{M}_{\le \mathbf{r}_{n+1}}$ with $\mathbf{r}_{n+1} \gg \mathbf{r}_n$. Hence, the computation is feasible only if we apply truncation to $\mathcal{M}_{\le \mathbf{r}}$ for some $\mathbf{r}$. In particular, this makes sense when the limit $\mathbf{u} = \lim \mathbf{u}_n$ can be well approximated in $\mathcal{M}_{\le \mathbf{r}}$. The resulting iteration becomes

$$\mathbf{y}_{n+1} = \mathbf{u}_n + \mathbf{F}_n(\mathbf{u}_n) \in \mathbb{R}^{\mathbf{I}},\ \mathbf{u}_{n+1} = HOSVD_{\mathbf{r}}(\mathbf{y}_{n+1}),$$

where $HOSVD_{\mathbf{r}}$ denotes the HOSVD truncation into $\mathcal{M}_{\le \mathbf{r}}$. In the case of an iteration with better than linear convergence, an analysis of the limiting behaviour is proved in [23]. The singular values obtained by $HOSVD_{\mathbf{r}}$ also allow for an adaptive choice of $\mathbf{r}$.

12.5.3 Riemannian Manifold Techniques

Instead of invoking $HOSVD_{\mathbf{r}}(\mathbf{y}_{n+1})$, the reduction step can be simplified by projecting $\mathbf{F}_n(\mathbf{u}_n)$ first to the tangent space $\mathcal{T}_{\mathbf{u}_n}$ at $\mathbf{u}_n$, i.e., $\mathbf{y}_{n+1} := \mathbf{u}_n + P_{\mathcal{T}_{\mathbf{u}_n}}\mathbf{F}_n(\mathbf{u}_n)$. The projection is defined in (12.12) and can be computed by standard techniques. However we need only the computation of the components $\delta\mathbf{u}^\alpha \in W_\alpha$, $\alpha \in \mathbb{T}$. Afterwards one has to project $\mathbf{y}_{n+1} \in \mathbf{u}_n + \mathcal{T}_{\mathbf{u}_n}$ to the manifold $\mathcal{M}_{\mathbf{r}}$, by using an appropriate *retraction* $R_{\mathbf{u}_n}(\mathbf{y}_{n+1} - \mathbf{u}_n)$,

$$\mathbf{y}_{n+1} := \mathbf{u}_n + P_{\mathcal{T}_{\mathbf{u}_n}}\mathbf{F}_n(\mathbf{u}_n) = \mathbf{u}_n + \boldsymbol{\xi}_n \in \mathbf{u}_n + \mathcal{T}_{\mathbf{u}_n} \tag{12.15}$$

$$\mathbf{u}_{n+1} := R_{\mathbf{u}_n}(\boldsymbol{\xi}_n) = R_{\mathbf{u}_n}(\mathbf{y}_{n+1} - \mathbf{u}_n)\,. \tag{12.16}$$

The retraction $R : (\mathbf{u}, \boldsymbol{\xi}) \mapsto R_{\mathbf{u}}(\boldsymbol{\xi}) \in \mathscr{M}_{\mathbf{r}}$ in (12.16) is a (local) mapping from the tangent bundle $\mathscr{T}(\mathscr{M}_{\mathbf{r}})$ to the manifold $\mathscr{M}_{\mathbf{r}}$, which has to satisfy a rigidity condition at $\boldsymbol{\xi}_n = \mathbf{0}$:

$$R_{\mathbf{u}_n}(\boldsymbol{\xi}_n) = \mathbf{u}_n + \boldsymbol{\xi}_n + \mathscr{O}(\|\boldsymbol{\xi}_n\|^2)\,.$$

A prototypical example for a retraction R is the *exponential map*. A retraction works as an approximate exponential map, but might be much simpler to implement. The concept of retractions has been introduced e.g. in [1] for optimisation on manifolds. We remark that the performance of the above algorithm depends crucially on the choice of the Riemannian metric and the retraction. One choice of a retraction is HOSVD, noting that $\mathbf{y}_n \in \mathscr{M}_{\leq 2\mathbf{r}}$.

Locally we observe the following error bound for the approximation of $\mathbf{v}_n := \mathbf{u}_n + \mathbf{F}_n(\mathbf{u}_n)$:

$$\|\mathbf{u}_{n+1} - \mathbf{v}_n\| \leq \min_{\mathbf{w}_n \in \mathscr{M}_{\leq r}} \|\mathbf{w}_n - \mathbf{v}_n\| + \mathscr{O}(\|\mathbf{u}_{n+1} - \mathbf{v}_n\|^2)$$

i.e. the best approximation plus a second order term.

12.5.4 Optimisation Problems

In the treatment of unconstrained optimisation problem one is looking for a minimiser $\mathbf{u}$ of a given cost functional $J : \mathbb{R}^{\mathbf{I}} \to \mathbb{R}$ [9–11]. We are seeking an appropriate approximation $\mathbf{u}_\varepsilon \in \mathscr{M}_{\leq \mathbf{r}}$ being a low rank hierarchical tensor. For this purpose we ask for a minimiser of the constrained problem

$$\mathbf{u}_\varepsilon := \operatorname{argmin}\{J(\mathbf{v}) : \mathbf{v} \in \mathscr{M}_{\leq \mathbf{r}}\}. \tag{12.17}$$

The first-order necessary condition for a minimiser of a cost functional $J : \mathbb{R}^{\mathbb{I}} \to \mathbb{R}$ constrained to the manifold $\mathscr{M}_{\mathbf{r}}$ can be formulated by (cf. [2, 25, 40])

$$\langle \nabla J(\mathbf{u}), \delta\mathbf{u} \rangle = 0, \quad \text{for all } \delta\mathbf{u} \in \mathscr{T}_{\mathbf{u}}\,. \tag{12.18}$$

Constrained optimisation problems can be considered similarly. Many different high-dimensional problems can be cast into the present variational framework. Local optimisation methods for the numerical treatment of these problems can be applied, in particular gradient or gradient like methods. For these methods one replaces the gradient by the Riemannian gradient, and arrives at a scheme of the form (12.15).

As a simple alternative, we mention an alternating directional search approach (ALS—alternating least squares method or alternating linear scheme) [28]. Given $\mathbf{u}^n = \left(\mathbf{U}_\alpha^n\right)_{\alpha \in \mathbb{T}}$, we fix all components, except one $\mathbf{U}_\alpha^n \in X_\alpha$ and optimise

$\operatorname{argmin}_{\mathbf{v}^\alpha \in X_\alpha} J \circ \tau\big(\mathbf{X}_{\mathbf{v}^\alpha}(\mathscr{U}_n)\big)$. Afterwards we update the tensor and bring the resulting tensor into a standard form, e.g. by orthogonalisation. Then we repeat the procedure optimising the other components. The optimisation has to be performed only on small subspaces. Let us highlight that the present parametrisation is linear, in contrast to multi-linear tensor parametrisation. Therefore a quadratic optimisation problems turns into a quadratic optimisation, and linear equations in the large ambient space turn into linear equations on small parameter spaces. This simple numerical technique is easy to implement and works quite efficiently. First convergence analysis has been developed [45], and further results are work in progress [48].

Local optimisation techniques can be applied to non-symmetric equations, and more generally to dynamical time-dependent problem in the framework of **Dirac–Frenkel Variational Principle** (see Chap. 19 in this volume). We consider the dynamical problem

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u} = \mathbf{F}(\mathbf{u}), \quad \mathbf{u}(0) = \mathbf{u}_0 \in \mathscr{M}_{\mathbf{r}}. \tag{12.19}$$

There may be various ways to approximate the trajectory of the initial value problem (12.19). The best approximation $\mathbf{u}_{\mathbf{r}}(t) := \operatorname{argmin}_{\mathbf{v}\in\mathscr{M}_{\mathbf{r}}} \|\mathbf{v}(t) - \mathbf{u}(t)\|$ is not feasible, because this requires the knowledge of exact trajectory $\mathbf{u}(t)$. The Dirac–Frenkel variational principle [39] determines the approximate trajectory on a given manifold $\mathbf{u}_{\mathbf{r}}(t) \in \mathscr{M}_{\mathbf{r}}$, which minimises

$$\|\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}(t) - \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}_{\mathbf{r}}(t)\| \to \min, \quad \mathbf{u}_{\mathbf{r}}(0) = \mathbf{u}(0).$$

This leads to the weak formulation

$$\langle \tfrac{\mathrm{d}}{\mathrm{d}t}\mathbf{u}_{\mathbf{r}} - \mathbf{F}(\mathbf{u}_{\mathbf{r}}), \delta\mathbf{u}\rangle = 0, \quad \text{for all } \delta\mathbf{u} \in \mathscr{T}_{\mathbf{u}_{\mathbf{r}}}. \tag{12.20}$$

In the case that the manifold is a closed linear space, the equations above are simply the corresponding Galerkin equations. Let us highlight that for the gradient in the limiting case $\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{u} = 0$, one obtains the first order condition (12.18). This approach applies also to non-variational problems, e.g. non-symmetric equations. The Dirac–Frenkel principle is well-known in molecular quantum dynamics (MCTDH) [39] for the Tucker format. For hierarchical tensors it has been formulated by [39, 51]. First convergence results have been established recently in [40]. A simple explicit Euler scheme time stepping [6, 7, 42] for the numerical treatment of (12.20) results again in (12.15).

Nevertheless, all local optimisation methods share the difficulty that they do not necessarily provide a global minimum. Let us remark that, in general, if the exact solution $\mathbf{u} := \operatorname{argmin}_{\mathbf{v}\in\mathbb{R}^{\mathbf{I}}} J(\mathbf{v})$ is not in $\mathscr{M}_{\leq\mathbf{r}}$, finding the solution of the constraint optimisation problem (12.17) is NP hard [26]. However, if $\mathbf{u} \notin \mathscr{M}_{\leq\mathbf{r}}$, we do not

need the exact minimiser of (12.17). We need only an appropriate approximation $\tilde{\mathbf{u}} \in \mathscr{M}_{\le \mathbf{r}}$. We are aiming that a quasi-optimal solution $\tilde{\mathbf{u}} \in \mathscr{M}_{\le \mathbf{r}}$, i.e.

$$\|\tilde{\mathbf{u}} - \mathbf{u}\| \le c \, \inf_{\mathbf{v} \in \mathscr{M}_{\le \mathbf{r}}} \|\mathbf{u} - \mathbf{v}\| \, .$$

This is much easier to obtain by local optimisation techniques, in particular, if combined with some a posteriori error control.

12.6 Tensorisation

In the sequel we will use the unbolded notation v and $v[k]$ for vectors in $V = \mathbb{R}^n$, and reserve the boldface notation for higher order tensors $\mathbf{u} \in \bigotimes_{i=1}^d V_i$. What we have in mind are mainly function based tensors. Given for example $f \in C^0([0,1])$, sampling f at a uniform grid yields

$$v[k] := f(2^{-d}k) \, , \quad k = 0, 1, \ldots, 2^d - 1 \, .$$

Alternatively, we may already consider $v[k] := \langle f, \phi_k \rangle$, $k = 0, 1, \ldots, 2^d - 1$, where $\phi_k(x) := 2^{d/2}\phi(2^d x - k)$. Here ϕ is a one-periodic function, e.g. a scaling function from a certain multi-resolution analysis.

Let us consider the vector $k \mapsto f[k]$, $k \in \Delta_d := \{0, \ldots, 2^d - 1\} \to \mathbb{R}$. If we represent k in the binary form

$$k := \kappa(\boldsymbol{\mu}) := \sum_{j=1}^{d} 2^{j-1}\mu_j \, , \quad \mu_j \in I_j := \{0, 1\} \, , \quad j = 1, \ldots, d, \tag{12.21}$$

then the inverse map $\kappa^{-1} : \mathbf{I} = \times_{j=1}^d I_j \to \Delta_d$ is a bijection. Any d-tuple $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$, which is a binary string of length d, defines an integer $\kappa(\boldsymbol{\mu}) \in \Delta_d = \{0, \ldots, 2^d - 1\}$. The tensor $\mathbf{v} \in \mathbb{R}^{\mathbf{I}} = \bigotimes_{j=1}^d \mathbb{R}^2$ of order d associated to the vector v is defined by

$$\boldsymbol{\mu} \mapsto \mathbf{v}[\boldsymbol{\mu}] := v[\kappa(\boldsymbol{\mu})] \, , \quad \boldsymbol{\mu} \in \mathbf{I} \, . \tag{12.22}$$

Therefore any vector $v \in \mathbb{R}^{2^d}$ can be converted isomorphically into a tensor $\mathbf{v} \in \mathbb{R}^{\mathbf{I}} = \bigotimes_{i=1}^d \mathbb{R}^2$. So far no information is lost and no reduction of data is achieved. The idea is to apply tensor product approximation in a hierarchical tensor format. Since the ordering of the binary indices $\mu_j \in \{0, 1\}$ resembles a multi-scale decomposition of Δ_j, the special TT format is relatively canonical. We refer to the definitions in Sect. 12.3.1.

To understand why this compression can be quite useful, we have to show that the required multi-linear rank $\mathbf{r} = (r_1, \ldots, r_{d-1})$ is small in many important cases. Indeed these ranks are the ranks of the corresponding matricisations $\mathbf{M}_j(\mathbf{v})$.

If a vector $v = (v[0], \ldots, v[n-1]) \in \mathbb{R}^n$, $n = 2^d$, has the tensorisation $\mathbf{v} \in \bigotimes_{j=1}^{d} \mathbb{R}^2$, its matricisation $\mathbf{M}_j(\mathbf{v})$, $1 \le j \le d-1$, yields

$$\mathbf{M}_j(\mathbf{v}) = \begin{bmatrix} v[0] & v[m] & \cdots & v[n-m] \\ v[1] & v[m+1] & \cdots & v[n-m+1] \\ \vdots & \vdots & & \vdots \\ v[m-1] & v[2m-1] & \cdots & v[n-1] \end{bmatrix} \text{ with } m := 2^j .$$

For $v[k] := f(2^{-d}k)$, the ℓ-th column ($0 \le \ell \le 2^j - 1$) samples the shifted functions $f_\ell(x) := f(x - 2^{-j}\ell)$, $0 \le x < 2^{-j}$.

If, e.g., f is $2^{-\ell}$ periodic, the j-th rank is equal to one. If f is a (piece wise) polynomial order $p-1$, the rank is bounded by p. More generally, if the linear spaces $S_p := \operatorname{span}\{g_1, \ldots, g_p\}$ are 2^{-j}-translation invariant, i.e. $f \in S_p \Rightarrow f(\cdot - 2^{-j}) \in S_p$, then the j-rank is bound by p. Typical examples of translation invariant spaces are spaces consisting of all solutions of homogenous constant coefficient differential equations

$$g^{(p-1)} + a_{p-2}g^{(p-2)} + \ldots + a_1 g' + a_0 = 0 .$$

Therefore the trigonometric polynomials $\{\sin 2\pi nx, \cos 2\pi nx : 1 \le n \le p/2\}$ are of rank at most $p+1$, while $\operatorname{span}\{e^{-\lambda x}\}$ is translation invariant for any λ, leading to rank 1. Whenever convergence is estimated by local polynomial approximation, the present approach works as well [15]. Typically, for piecewise analytic functions, the singular values of the matricisation are decaying exponentially. For example, Gaussian functions $e^{-\alpha x^2}$ can be approximated by small rank, widely independent of α. The tensorisation of vectors is heavily used in first numerical experiments with high-dimensional PDEs. There is a large collection of papers in this regard, which cannot be discussed completely in the present short survey. Vector tensorisation in the above form has been applied to Hartree–Fock computations, density functional theory and the treatment two-electron integrals by [29–31].

For matrices and higher-order tensors we can proceed by combining our previous tensor techniques with the binary index representation

$$\mathbf{v}[\mu_1 \ldots, \mu_{d_1}, \nu_1, \cdots, \nu_{d_2}] .$$

However, this is not the only way and not always recommended. Another ordering $\mathbf{v}[\mu_1, \nu_1 \ldots, \mu_{d_1}, \nu_{d_2}]$ may be preferred. This case corresponds to the tiling of the unit square $\Omega = [0,1]^2$ into congruent squares of length 2^{-j}. This way is even better suited to the treatment of linear operators (matrices) in $\mathbb{R}^{\mathbf{I}}$. For example, the identity matrix represented in the first case has maximal rank d, in the second case it has rank 1. Another well-know transformation, the Hadamard–Walsh transform has rank 1.

Another interesting transformation is the Fourier transform. It turns out that it is not of low rank. However, it can be performed as a product of low rank transformations, as used in the fast Fourier transform (FFT).

Let us remark that for matrices and high-order tensors there is often a more promising re-ordering of the binary indices. To find these representations, we need an algorithm which searches for a best ordering, or more generally for a best tree in the general hierarchical setting. In [3] an adaptive strategy for this purpose has been developed. The authors demonstrate on the example $f_\alpha(x, y) = 1/\sqrt{\varepsilon - (x - \theta y)^2},\ 0 \le x, y \le 1$, that for an arbitrary direction α the function f_α can be approximated with almost the same ranks.

The HOSVD provides a useful tool to analyse a given function with respect to tensorised representation, i.e. approximate the tensorised tensor train representation by low-rank TT tensors. In [32, 44] it has also shown that one can define complementary spaces with wavelet-like functions. Usually this part, if not neglected at all, is sparse in many cases. We do not pursue this direction further right now. The idea of transforming a vector into a tensor, or a low-order tensor into a high-order tensor, and combining this with the compression techniques of hierarchical tensors as described above, has not being completely elaborated so far. There is still much room left for further research in this direction.

Last but not least we would like to mention that with a different technique of vector tensorisation, particularly suited for fermions, the approximate solution of the electronic Schrödinger equation can be cast into a high-order tensor over $\mathbb{C}^2$ [37, 38].

The present calculus has many links to the numerical treatment of fermions in the Fock space and to basic techniques in quantum information theory. Both are dealing with tensor spaces $\bigotimes_{j=1}^{d} \mathbb{C}^2$. However, we do not consider the probabilistic interpretation of the tensors in the spirit of quantum mechanics. Moreover, we do not confine to the special case that the operators under consideration have to be unitary. Indeed the present approach is completely independent of quantum mechanics. The common features are space of high-order tensors and problems related to this. On the other hand, we are restricted to low-rank hierarchical representations due to practical limitations, since otherwise the complexity makes the computations intractable by our methods. In quantum mechanics the concept of hierarchical tensors existed for several decades, although not in this pure form and often covered under some other issues, like in early renormalisation group theory, such that there may be further results helpful for our purpose, and both communities can benefit from merging expertise.

12.6.1 Convolution

Let $n = 2^d$. The convolution of $v, w \in \mathbb{R}^n$ is defined as in (12.14). The result belongs to $\mathbb{R}^{2n-1}$, which can be embedded into $\mathbb{R}^{2n}$. If the convolution is based on the traditional fast Fourier transform, the required work is $O(n \log n)$. Using

the tensorisation, the data sizes of the corresponding tensors $\mathbf{v}, \mathbf{w} \in \bigotimes_{j=1}^{d} \mathbb{R}^2$ (or their approximations) are expected to be much smaller than $O(n)$, possibly even of the magnitude $O(\log n)$. Then an algorithm for $u = v \star w$ should require a work corresponding to the data sizes of $\mathbf{v}$ and $\mathbf{w}$. We have to define an operation $\mathbf{v} \star \mathbf{w}$ in such a way that $\mathbf{u} = \mathbf{v} \star \mathbf{w}$ holds.

The convolution of multivariate functions can be separated (see (12.14) and (12.13)). This leads to a simple algorithm for all tensor formats. In the case of tensorisation, the d-dimensionality is more artificial. In fact, for $\mathbf{v} = \bigotimes_{j=1}^{d} v^j$ and $\mathbf{w} = \bigotimes_{j=1}^{d} w^j$ with $v^j, w^j \in \mathbb{R}^2$, a statement of the form $\mathbf{v} \star \mathbf{w} = \bigotimes_{j=1}^{d} \left(v^j \star w^j\right)$ does not make sense since $v^j \star w^j$ is an element of $\mathbb{R}^3$. However, the embedding of $\mathbb{R}^2$ into ℓ_0 , the vector space of all sequences $(a_i)_{i=0}^{\infty}$ with finitely many $a_i \neq 0$, can be extended to an embedding of $\bigotimes_{j=1}^{d} \mathbb{R}^2$ into $\bigotimes_{j=1}^{d} \ell_0$. The interpretation of the convolution in $\bigotimes_{j=1}^{d} \ell_0$ together with the TT tensor format leads to an algorithm for $\mathbf{u} = \mathbf{v} \star \mathbf{w}$ whose cost is only related to the data sizes of $\mathbf{v}$ and $\mathbf{w}$ (details in [19]).

The hierarchical tensor representation has been introduced in the first funding period. During the past 5 years, 50 articles for journals and proceedings have been finished at the MPI Leipzig and are already published. During the same time 20 articles for journals and proceedings has been finished by the TU Berlin group about the present subject and related topics. Among them are eight joint and already published papers with the MPI group. Further papers are published by I. Oseledets and other authors from the Russian group. J. Ballani, S. Kühn, V. Khoromskaia, T. Rohwedder, and A. Uschmajew have finished their doctoral theses about tensor approximation and spectral problems for high-dimensional PDEs. For sake of brevity, we do not cite all of them. We have mentioned only a selection of those which are relevant for the present considerations. Prof. R. Schneider presented the material in a series of *John von Neumann Lectures* as *John von Neumann Guest Professor* at TU Munich in 2012, and thanks for the kind hospitality of the Mathematical Institute at TUM.

Prof. W. Hackbusch has published a recent monograph about Numerical Tensor Calculus [20], and a survey article for *Acta Numerica* is in print [22]. Our research is strongly related to the projects of L. Grasedyck and C. Lubich, hence we refer to their articles in this volume for further information.

References

1. Absil, P.-A., Mahony, R., Sepulchre, R.: Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton (2008)
2. Arnold, A., Jahnke, T.: On the approximation of high-dimensional differential equations in the hierarchical Tucker format. BIT (2013). doi:10.1007/s10543-013-0444-2
3. Ballani, J., Grasedyck, L.: Tree adaptive approximation in the hierarchical tensor format. SIAM J. Sci. Comput. **36**, A1415–A1431 (2014)

4. Beylkin, G., Mohlenkamp, M.J.: Algorithms for numerical analysis in high dimensions. SIAM J. Sci. Comput. **26**, 2133–2159 (2005)
5. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. SIAM J. Matrix Anal. Appl. **21**, 1253–1278 (2000)
6. Dolgov, S., Khoromskij, B.: Simultaneous state-time approximation of the chemical master equation using tensor product formats. NLAA, online (2014)
7. Dolgov, S., Khoromskij, B., Oseledets, I.V.: Fast solution of multi-dimensional parabolic problems in the TT/QTT formats with initial application to the Fokker-Planck equation. SIAM J. Sci. Comput. **34**, A3016–A3038 (2012)
8. Edelman, A., Arias, T.A., Smith, S.T.: The geometry of algorithms with orthogonality constraints. SIAM J. Matrix Anal. Appl. **20**, 303–353 (1998)
9. Espig, M., Hackbusch, W.: A regularised Newton method for the efficient approximation of tensors represented in the canonical tensor format. Numer. Math. **122**, 489–525 (2012)
10. Espig, M., Hackbusch, W., Handschuh, S., Schneider, R.: Optimization problems in contracted tensor networks. Comput. Vis. Sci. **14**, 271–285 (2012)
11. Espig, M., Hackbusch, W., Rohwedder, T., Schneider, R.: Variational calculus with sums of elementary tensors of fixed rank. Numer. Math. **122**, 469–488 (2012)
12. Falcó, A., Hackbusch, W.: On minimal subspaces in tensor representations. Found. Comput. Math. **12**, 765–803 (2012)
13. Falcó, A., Hackbusch, W., Nouy, A.: Geometric structures in tensor representations. Preprint 9/2013, Leipzig (2013)
14. Grasedyck, L.: Hierarchical singular value decomposition of tensors. SIAM J. Matrix Anal. Appl. **31**, 2029–2054 (2010)
15. Grasedyck, L.: Polynomial approximation in hierarchical Tucker format by vector-tensorization. SPP 1324 Preprint 43 (2010)
16. Grasedyck, L., Hackbusch, W.: An introduction to hierarchical ($\mathscr{H}$-)rank and TT-rank of tensors with examples. Comput. Methods Appl. Math. **11**, 291–304 (2011)
17. Grasedyck, L., Kressner, D., Tobler, C.: A literature survey of low-rank tensor approximation techniques. GAMM-Mitteilungen **36**(1), 53–78 (2013)
18. Greub, W.H.: Multilinear Algebra, 2nd edn. Springer, New York (1978)
19. Hackbusch, W.: Tensorisation of vectors and their efficient convolution. Numer. Math. **119**, 465–488 (2011)
20. Hackbusch, W.: Tensor Spaces and Numerical Tensor Calculus. Springer, Berlin (2012)
21. Hackbusch, W.: L^∞ estimation of tensor truncations. Numer. Math. **125**, 419–440 (2013)
22. Hackbusch, W.: Numerical tensor calculus. Acta Numer. **23**, 651–742 (2014)
23. Hackbusch, W., Khoromskij, B., Tyrtyshnikov, E.E.: Approximate iterations for structured matrices. Numer. Math. **109**, 365–383 (2008)
24. Hackbusch, W., Kühn, S.: A new scheme for the tensor representation. J. Fourier Anal. Appl. **15**, 706–722 (2009)
25. Haegeman, J., Osborne, T., Verstraete, F.: Post-matrix product state methods: to tangent space and beyond. Phys. Rev. B **88**, 075133 (2013)
26. Hillar, C.J., Lim, L.-H.: Most tensor problems are NP hard. J. ACM **60**(6), 1–39 (2013)
27. Holtz, S., Rohwedder, T., Schneider, R.: On manifolds of tensors of fixed TT-rank. Numer. Math. **120**, 701–731 (2012)
28. Holtz, S., Rohwedder, T., Schneider, R.: The alternating linear scheme for tensor optimization in the tensor train format. SIAM J. Sci. Comput. **34**, A683–A713 (2012)
29. Khoromskaia, V., Khoromskij, B.: Møller-Plesset (MP2) energy correction using tensor factorizations of the grid-based two-electron integrals. Comp. Phys. Comm. **185**, 2–10 (2014)
30. Khoromskaia, V., Khoromskij, B., Schneider, R.: QTT representation of the Hartree and exchange operators in electronic structure calculations. Comput. Methods Appl. Math. **11**, 327–341 (2011)
31. Khoromskaia, V., Khoromskij, B., Schneider, R.: Tensor-structured calculation of two-electron integrals in a general basis. SIAM J. Sci. Comput. **35**, A987–A1010 (2013)

32. Khoromskij, B., Oseledets, I.V.: Quantics-TT approximation of elliptic solution operators in higher dimensions. Russ. J. Numer. Anal. Math. Model. **26**, 303–322 (2011)
33. Koch, O., Lubich, C.: Dynamical low rank approximation. SIAM J. Matrix Anal. Appl. **29**, 434–454 (2007)
34. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**, 455–500 (2009)
35. Kressner, D., Tobler, C.: `htucker` – a MATLAB toolbox for tensors in hierarchical Tucker format. Technical report, MATHICSE, EPF Lausanne (2012)
36. Landsberg, J.M.: Tensors: Geometry and Applications. AMS, Providence (2012)
37. Legeza, O., Rohwedder, T., Schneider, R.: High dimensional methods in quantum chemistry. In: Encyclopedia of Applied and Computational Mathematics. Springer (to appear)
38. Legeza, O., Rohwedder, T., Schneider, R., Szalay, S.: Tensor product approximation (DMRG) and coupled cluster method in quantum chemistry. In: Bach, V., Delle, L. (eds.) Many-Electron Approaches in Physics, Chemistry and Mathematics: A Multidisciplinary View, Springer Verlag (2014). http://arxiv.org/abs/1310.2736
39. Lubich, C.: From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis. EMS, Zürich (2008)
40. Lubich, C., Rohwedder, T., Schneider, R., Vandereycken, B.: Dynamical approximation of hierarchical Tucker and tensor-train tensors. SIAM J. Matrix Anal. Appl. **34**, 470–494 (2013)
41. Oseledets, I.V.: Tensor-train decomposition. SIAM J. Sci. Comput. **33**, 2295–2317 (2011)
42. Oseledets, I.V., Khoromskij, B., Schneider, R.: Efficient time-stepping scheme for dynamics on TT manifolds. Preprint 24/2012, Leipzig (2012)
43. Oseledets, I.V., Tyrtyshnikov, E.E.: Breaking the curse of dimensionality, or how to use SVD in many dimensions. SIAM J. Sci. Comput. **31**, 3744–3759 (2009)
44. Oseledets, I.V., Tyrtyshnikov, E.E.: Algebraic wavelet transform via quantics tensor train decomposition. SIAM J. Sci. Comput. **33**(3), 1315–1328 (2011)
45. Rohwedder, T., Uschmajew, A.: On local convergence of alternating schemes for optimization of convex problems in the tensor train format. SIAM J. Numer. Anal. **51**(2), 1134–1162 (2013)
46. Schmidt, E.: Zur Theorie der linearen und nichtlinearen Integralgleichungen. I. Teil: Entwicklung willkürlicher Funktionen nach Systemen vorgeschriebener. Math. Ann. **63**, 433–476 (1907)
47. Schneider, R., Uschmajew, A.: Approximation rates for the hierarchical tensor format in periodic Sobolev spaces. J. Complex. **30**(2), 56–71 (2014)
48. Schneider, R., Uschmajew, A.: Convergence of gradient-related line-search methods on closed sets via Lojasiewicz inequality (in preparation)
49. Schollwöck, U.: The density-matrix renormalization group in the age of matrix product states. Ann. Phys. **326**, 96–192 (2011)
50. Uschmajew, A., Vandereycken, B.: The geometry of algorithms using hierarchical tensors. Linear Algebra Appl. **439**(1), 133–166 (2013)
51. Wang, H., Thoss, M.: Multilayer formulation of the multiconfiguration time-dependent Hartree theory. J. Chem. Phys. **119**, 1289–1299 (2003)

Chapter 13
Nonlinear Eigenproblems in Data Analysis: Balanced Graph Cuts and the RatioDCA-Prox

Leonardo Jost, Simon Setzer, and Matthias Hein

Abstract It has been recently shown that a large class of balanced graph cuts allows for an exact relaxation into a nonlinear eigenproblem. We review briefly some of these results and propose a family of algorithms to compute nonlinear eigenvectors which encompasses previous work as special cases. We provide a detailed analysis of the properties and the convergence behavior of these algorithms and then discuss their application in the area of balanced graph cuts.

13.1 Introduction

Spectral clustering is one of the standard methods for graph-based clustering [12]. It is based on the spectral relaxation of the so called normalized cut, which is one of the most popular criteria for balanced graph cuts. While the spectral relaxation is known to be loose [7], tighter relaxations based on the graph p-Laplacian have been proposed in [4]. Exact relaxations for the Cheeger cut based on the nonlinear eigenproblem of the graph 1-Laplacian have been proposed in [8, 11]. In [9] the general balanced graph cut problem of an undirected, weighted graph (V, E) is considered. Let $n = |V|$ and denote the weight matrix of the graph by $W = (w_{ij})_{i,j=1}^{n}$, then the general balanced graph cut criterion can be written as

$$\arg\min_{A \subset V} \frac{\mathrm{cut}(A, \overline{A})}{\hat{S}(A)},$$

where $\overline{A} = V \setminus A$, $\mathrm{cut}(A, \overline{A}) = \sum_{i \in A, j \in \overline{A}} w_{ij}$, and $\hat{S}: 2^V \to \mathbb{R}_+$ is a symmetric and nonnegative balancing function. Exact relaxations of such balanced graph cuts

L. Jost (✉) • M. Hein
University of Saarland, Postfach 15 11 50, 66041 Saarbrücken, Germany
e-mail: leo@santorin.cs.uni-sb.de; hein@math.uni-sb.de

S. Setzer
University of Saarland, Building E1.7, 66041 Saarbrücken, Germany
e-mail: simon.setzer@gmail.com

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_13

and relations to corresponding nonlinear eigenproblems are discussed in [9] and are briefly reviewed in Sect. 13.2. A further generalization to hypergraphs has been established in [10].

There exist different approaches to minimize the exact continuous relaxations. However, in all cases the problem boils down to the minimization of a ratio of a convex and a difference of convex functions. The two lines of work of [2,3] and [8,9] have developed different algorithms for this problem, which have been compared in [2]. We show that both types of algorithms are special cases of our new algorithm RatioDCA-prox introduced in Sect. 13.3.1. We provide a unified analysis of the properties and the convergence behavior of RatioDCA-prox. Moreover, in Sect. 13.4 we prove stronger convergence results when the RatioDCA-prox is applied to the balanced graph cut problem or, more generally, problems where one minimizes nonnegative ratios of Lovasz extensions of set functions. Further, we discuss the choice of the relaxation of the balancing function in [9] and show that from a theoretical perspective the Lovasz extension is optimal which is supported by the numerical results in Sect. 13.5.

13.2 Exact Relaxation of Balanced Graph Cuts

A key element for the exact continuous relaxation of balanced graph cuts is the Lovasz extension of a function on the power set 2^V to $\mathbb{R}^V$.

Definition 13.1. Let $\hat{S} : 2^V \to \mathbb{R}$ be a set function with $\hat{S}(\emptyset) = 0$. Let $f \in \mathbb{R}^V$, let V be ordered such that $f_1 \leq f_2 \leq \ldots \leq f_n$ and define $C_i = \{j \in V \mid j > i\}$. Then, the Lovasz extension $S : \mathbb{R}^V \to \mathbb{R}$ of $\hat{S}$ is given by

$$S(f) = \sum_{i=1}^{n} f_i \Big(\hat{S}(C_{i-1}) - \hat{S}(C_i) \Big) = \sum_{i=1}^{n-1} \hat{S}(C_i)(f_{i+1} - f_i) + f_1 \hat{S}(V).$$

Note that for the characteristic function of a set $C \subset V$, we have $S(\mathbf{1}_C) = \hat{S}(C)$.

The Lovasz extension is convex if and only if $\hat{S}$ is submodular [1] and every Lovasz extension can be written as a difference of convex functions [9]. Moreover, the Lovasz extension of a symmetric set function is positively one-homogeneous[1] and preserves non-negativity, that is $S(f) \geq 0$, $\forall f \in \mathbb{R}^V$ if $\hat{S}(A) \geq 0$, $\forall A \subset V$. It it well known, see e.g. [10], that the Lovasz extension of the submodular cut function, $\hat{R}(A) = \mathrm{cut}(A, \overline{A})$, yields the total variation on a graph,

[1] A function $A: \mathbb{R}^n \to \mathbb{R}$ is (positively) p-homogeneous if $A(vx) = v^p A(x)$ for all $v \in \mathbb{R}$ $(v \geq 0)$. In the following we will call functions just homogeneous when referring to positive homogeneity.

$$R(f) = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} |f_i - f_j|. \tag{13.1}$$

Theorem 13.1 shows exact continuous relaxations of balanced graph cuts [9]. A more general version for the class of constrained fractional set programs is given in [5].

Theorem 13.1. *Let $G = (V, E)$ be an undirected, weighted graph and $S : V \to \mathbb{R}$ and let $\hat{S} : 2^V \to \mathbb{R}$ be symmetric with $\hat{S}(\emptyset) = 0$, then*

$$\min_{f \in \mathbb{R}^V} \frac{\frac{1}{2} \sum_{i,j=1}^{n} w_{ij} |f_i - f_j|}{S(f)} = \min_{A \subset V} \frac{\mathrm{cut}(A, \overline{A})}{\hat{S}(A)},$$

if either one of the following two conditions holds

1. *S is one-homogeneous, even, convex and $S(f + \alpha \mathbf{1}) = S(f)$ for all $f \in \mathbb{R}^V$, $\alpha \in \mathbb{R}$ and $\hat{S}$ is defined as $\hat{S}(A) := S(\mathbf{1}_A)$ for all $A \subset V$.*
2. *S is the Lovasz extension of the non-negative, symmetric set function $\hat{S}$ with $\hat{S}(\emptyset) = 0$.*

Let $f \in \mathbb{R}^V$ and denote by $C_t := \{i \in V \mid f_i > t\}$, then it holds under both conditions,

$$\min_{t \in \mathbb{R}} \frac{\mathrm{cut}(C_t, \overline{C_t})}{\hat{S}(C_t)} \leq \frac{\frac{1}{2} \sum_{i,j=1}^{n} w_{ij} |f_i - f_j|}{S(f)}.$$

We observe that the exact continuous relaxation corresponds to a minimization problem of a ratio of non-negative, one-homogeneous functions, where the enumerator is convex and the denominator can be written as a difference of convex functions.

13.3 Minimization of Ratios of Non-negative Differences of Convex Functions via the RatioDCA-Prox

We consider in this paper continuous optimization problems of the form

$$\min_{f \in \mathbb{R}^V} F(f), \quad \text{where} \quad F(f) = \frac{R(f)}{S(f)} = \frac{R_1(f) - R_2(f)}{S_1(f) - S_2(f)}, \tag{13.2}$$

where R_1, R_2, S_1, S_2 are convex and one-homogeneous and $R(f) = R_1(f) - R_2(f)$ and $S(f) = S_1(f) - S_2(f)$ are non-negative. Thus we are minimizing a non-negative ratio of d.c. (difference of convex) functions. As discussed above the exact continuous relaxation of Theorem 13.1 leads exactly to such a problem, where

$R_2(f) = 0$ and $R_1(f) = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}|f_i - f_j|$. Different choices of balancing functions lead to different functions S.

While [2, 3, 8] consider only algorithms for the minimization of ratios of convex functions, in [9] the RatioDCA has been proposed for the minimization of problems of type (13.2). The generalized version RatioDCA-prox is a family of algorithms which contains the work of [2, 3, 8, 9] as special cases and allows us to treat the minimization problem (13.2) in a unified manner.

13.3.1 The RatioDCA-Prox Algorithm

The RatioDCA-prox algorithm for minimization of (13.2) is given in Algorithm 8. In each step one has to solve the convex optimization problem

$$\min_{G(u)\leq 1} \Phi_{f^k}^{c^k}(u), \tag{13.3}$$

which we denote as the *inner problem* in the following with

$$\Phi_{f^k}^{c^k}(u) := R_1(u) - \langle u, r_2(f^k)\rangle + \lambda^k \left(S_2(u) - \langle u, s_1(f^k)\rangle\right) - c^k \langle u, g(f^k)\rangle$$

and $c^k \geq 0$. As the constraint set we can choose any set containing a neighborhood of 0, such that the inner problem is bounded from below, i.e. any nonnegative convex p-homogeneous ($p \geq 1$) function G. Although a slightly more general formulation is possible, we choose the constraint set to be compact, i.e. $G(f) = 0 \Leftrightarrow f = 0$. Moreover, $s_1(f^k) \in \partial S_1(f^k)$, $r_2(f^k) \in \partial R_2(f^k)$, $g(f^k) \in \partial G(f^k)$, where $\partial S_1, \partial R_2, \partial G$ are the subdifferentials. Note that for any p-homogeneous function A we have the generalized Euler identity [14, Theorem 2.1] that is $\langle f, a(f)\rangle = p\, A(f)$ for all $a(f) \in \partial A(f)$.

Clearly $\Phi_{f^k}^{c^k}$ is also one-homogeneous and with the Euler identity we get $\Phi_{f^k}^{c^k}(f^k) = -c^k\, pG(f^k) \leq 0$ so we can always find minimizers at the boundary.

Algorithm 8 RatioDCA-prox – Minimization of a ratio of non-negative, one-homogeneous d.c. functions

1: **Initialization:** $f^0 =$ random with $G(f^0) = 1$, $\lambda^0 = F(f^0)$
2: **repeat**
3: find $s_1(f^k) \in \partial S_1(f^k)$, $r_2(f^k) \in \partial R_2(f^k)$, $g(f^k) \in \partial G(f^k)$
4: find $f^{k+1} \in \arg\min_{G(u)\leq 1} \Phi_{f^k}^{c^k}(u)$
5: $\lambda^{k+1} = F(f^{k+1})$
6: **until** $f^{k+1} \in \arg\min_{G(u)\leq 1} \Phi_{f^{k+1}}^{c^{k+1}}(u)$

The difference to the RatioDCA in [9] is the additional proximal term $-c^k \langle u, g(f^k)\rangle$ in $\Phi_{f^k}^{c^k}(u)$ and the choice of G. It is interesting to note that this

term can be derived by applying the RatioDCA to a different d.c. decomposition of F. Let us write F as

$$F = \frac{R_1' - R_2'}{S_1' - S_2'} = \frac{(R_1 + c_R G) - (R_2 + c_R G)}{(S_1 + c_S G) - (S_2 + c_S G)} \tag{13.4}$$

with arbitrary $c_R, c_S \geq 0$. If we now define $c^k := c_R + \lambda^k c_S$, the function to be minimized in the inner problem of the RatioDCA reads

$$\Phi'_{f^k}(u) = R_1'(u) - \langle u, r_2'(f^k)\rangle + \lambda^k \left(S_2'(u) - \langle u, s_1'(f^k)\rangle\right) = \Phi^{c_k}_{f^k}(u) + c^k G(u),$$

which is not necessarily one-homogeneous anymore. The following lemma implies that the minimizers of the inner problem of RatioDCA-prox and of RatioDCA applied to the d.c.-decomposition (13.4) can be chosen to be the same.

Lemma 13.1. *For* $G(f^k) = 1$ *we have* $\arg\min_{G(u)\leq 1} \Phi'_{f^k}(u) \supseteq \arg\min_{G(u)\leq 1} \Phi^{c^k}_{f^k}(u)$. *Moreover,*

1. *If* $p > 1, c^k > 0$ *then* $\arg\min_u \Phi'_{f^k}(u) \supseteq \nu \cdot \arg\min_{G(u)\leq 1} \Phi^{c^k}_{f^k}(u)$ *for some* $\nu \geq 1$,
2. *If* $f^k \in \arg\min_{G(u)\leq 1} \Phi^{c^k}_{f^k}(u)$ *then* $\arg\min_u \Phi'_{f^k}(u) \supseteq \arg\min_{G(u)\leq 1} \Phi^{c^k}_{f^k}(u)$.

Proof. For fixed $\xi \geq 0$ it follows from the one-homogeneity of $\Phi^{c^k}_{f^k}$ that any minimizer of $\arg\min_{G(u)=\xi} \Phi'_{f^k}(u)$ is a multiple of one $f^{k+1} \in \arg\min_{G(u)\leq 1} \Phi^{c^k}_{f^k}(u)$, so let us look at νf^{k+1} with $G(f^{k+1}) = 1$. We get from the homogeneity of $\Phi^{c^k}_{f^k}$ and G for $\nu > 0$ that

$$\frac{\partial}{\partial \nu}\left(\Phi'_{f^k}(\nu f^{k+1})\right) = \Phi^{c^k}_{f^k}(f^{k+1}) + c^k p \nu^{p-1} \leq c^k p(\nu^{p-1} - 1),$$

which is non-positive for $\nu \in (0, 1]$ and with $\Phi'_{f^k}(0) = 0 \geq \Phi'_{f^k}(f^k) = c^k(1-p)$ it follows that a minimum is attained at $\nu \geq 1$. If $p > 1, c^k > 0$ then the global optimum of Φ'_{f^k} exists and by the previous arguments is attained at multiples of $f^{k+1} \in \arg\min_{G(u)\leq 1} \Phi^{c^k}_{f^k}(u)$. If $f^k \in \arg\min_{G(u)\leq 1} \Phi^{c^k}_{f^k}(u)$ then also the global optimum of Φ'_{f^k} exists and the claim follows since $\nu = 1$ is a minimizer of $\Phi'_{f^k}(\nu f^k) = -\nu c^k p + \nu^p c^k$. □

Note that $G(f^k) = 1$ is no restriction since we get from the one-homogeneity of $\Phi^{c^k}_{f^k}$ that $G(f^k) = 1$ for all k. The following lemma verifies the intuition that the strength of the proximal term of RatioDCA-prox controls in some sense how close successive iterates are.

Lemma 13.2. *Let* $f_1^{k+1} \in \arg\min_{G(u)\leq 1} \Phi_{f^k}^{c^k}(u)$, *and* $f_2^{k+1} \in \arg\min_{G(u)\leq 1} \Phi_{f^k}^{d^k}(u)$. *If* $c^k \leq d^k$ *then* $\left\langle f_1^{k+1}, g(f^k)\right\rangle \leq \left\langle f_2^{k+1}, g(f^k)\right\rangle$.

Proof. This follows from

$$\begin{aligned}
\Phi_{f^k}^{d^k}(f_2^{k+1}) &\leq \Phi_{f^k}^{d^k}(f_1^{k+1}) = \Phi_{f^k}^{c^k}(f_1^{k+1}) + (c^k - d^k)\left\langle f_1^{k+1}, g(f^k)\right\rangle \\
&\leq \Phi_{f^k}^{c^k}(f_2^{k+1}) + (c^k - d^k)\left\langle f_1^{k+1}, g(f^k)\right\rangle \\
&= \Phi_{f^k}^{d^k}(f_2^{k+1}) + (d^k - c^k)\left\langle f_2^{k+1}, g(f^k)\right\rangle \\
&\quad + (c^k - d^k)\left\langle f_1^{k+1}, g(f^k)\right\rangle.
\end{aligned}$$

□

Remark 13.1. As all proofs can be split up into the individual steps we may choose different functions G in every step of the algorithm. Moreover, it will not be necessary that f^{k+1} is an exact minimizer of the inner problem, but we will only use that $\Phi_{f^k}^{c^k}(f^{k+1}) < \Phi_{f^k}^{c^k}(f^k)$.

13.3.2 Special Cases

It is easy to see that we get for $c^k = 0$ and $G = \|\cdot\|_2$ the RatioDCA [9] as a special case of the RatioDCA-prox. Moreover, Lemma 13.1 shows that the RatioDCA-prox corresponds to the RatioDCA with a general constraint set for the d.c. decomposition of the ratio F given in (13.4).

If we apply RatioDCA-prox to the ratio cut problem, where $\hat{S}(C) = |C||\overline{C}|$, then $R(u) = R_1(u) = \frac{1}{2}\sum_{i,j=1}^n w_{ij}|u_i - u_j|$ and [3] chose $S(u) = S_1(u) = \|u - \text{mean}(u)\mathbf{1}\|_1$. The following lemma shows that for a particular choice of G and c^k, RatioDCA-prox and algorithm 1 of [3], which calculates iterates $\tilde{f}^{k+1}$ for $v^k \in \partial S(f^k)$ by

$$\begin{aligned}
h^{k+1} &= \arg\min_u \left\{ \frac{1}{2}\sum_{i,j} w_{ij}|u_i - u_j| + \frac{\lambda^k}{2c}\|u - (\tilde{f}^k + cv^k)\|_2^2 \right\}, \\
\tilde{f}^{k+1} &= h^{k+1}/\left\|h^{k+1}\right\|_2,
\end{aligned}$$

produce the same sequence if given the same initialization.

Lemma 13.3. *If* $f^0 = \tilde{f}^0$, $\text{mean}(f^0) = 0$, $c > 0$ *and one uses the same subgradients in each step then, for the sequence* $\tilde{f}^k$ *produced by algorithm 1 of [3] and* f^k *produced by RatioDCA-prox with* $c^k = \frac{\lambda^k}{2c}$ *and* $G(u) = \|u\|_2^2$, *we have* $\tilde{f}^k = f^k$ *for all* k.

Proof. If $f^k = \tilde{f}^k$ and we choose $v^k := s_1(f^k) = s_1(\tilde{f}^k) = (\mathbb{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T)\,\mathrm{sign}(f^k - \mathrm{mean}(f^k))$. For RatioDCA-prox we get f^{k+1} by

$$f^{k+1} = \arg\min_{\|u\|_2^2 \leq 1} \Phi_{f^k}^{c^k}(u)$$

and for the algorithm 1 of [3]

$$\begin{aligned} h^{k+1} &= \arg\min_u \left\{ R(u) + \frac{\lambda^k}{2c}\left(\|u\|_2^2 - 2\left\langle u, f^k\right\rangle - 2\left\langle u, cv^k\right\rangle\right)\right\} \\ &= \arg\min_u \left\{ \Phi_{f^k}^{c^k}(u) + \frac{\lambda^k}{2c}\|u\|_2^2 \right\} \end{aligned}$$

Finally, $\tilde{f}^{k+1} = h^{k+1} / \left\|h^{k+1}\right\|_2$ and application of Lemma 13.1 then shows that $\tilde{f}^{k+1} = f^{k+1}$. As $\|.\|_2^2$ is strictly convex, the minimizers are unique. □

Analogously, the algorithm presented in [2] is a special case of RatioDCA-prox applied to the ratio cheeger cut where $R(u) = R_1(u) = \frac{1}{2}\sum_{i,j} w_{ij}|u_i - u_j|$ and $S(u) = S_1(u) = \sum_i |u_i - \mathrm{median}(u)|$.

13.3.3 Monotony and Convergence

In this section we show that the sequence $F(f^k)$ produced by RatioDCA-prox is monotonically decreasing similar to the RatioDCA of [9] and, additionally, we can show a convergence property, which generalizes the results of [2, 3].

Proposition 13.1. *For every nonnegative sequence c^k any sequence f^k produced by RatioDCA-prox satisfies $F(f^{k+1}) < F(f^k)$ for all $k \geq 0$ or the sequence terminates. Moreover, we get that $c^k\left\langle f^{k+1} - f^k, g(f^k)\right\rangle \to 0$.*

Proof. If the sequence does not terminate then $\Phi_{f^k}^{c^k}(f^{k+1}) < \Phi_{f^k}^{c^k}(f^k)$ and it follows

$$\begin{aligned} &R(f^{k+1}) - \lambda^k S(f^{k+1}) - c^k\left\langle f^{k+1}, g(f^k)\right\rangle \leq \Phi_{f^k}^{c^k}(f^{k+1}) \\ &< \Phi_{f^k}^{c^k}(f^k) = -c^k\left\langle f^k, g(f^k)\right\rangle, \end{aligned}$$

where we used that for any one-homogeneous convex function A we have for all $f, g \in \mathbb{R}^V$ and all $a \in \partial A(g)$

$$A(f) \geq A(g) + \langle f - g, a\rangle = \langle f, a\rangle .$$

Adding $c^k \left\langle f^{k+1}, g(f^k)\right\rangle$ gives

$$R(f^{k+1}) - \lambda^k S(f^{k+1}) < c^k \left\langle f^{k+1}, g(f^k)\right\rangle - c^k \left\langle f^k, g(f^k)\right\rangle \leq 0 \tag{13.5}$$

where we used that since G is convex

$$\left\langle f^{k+1}, g(f^k)\right\rangle - \left\langle f^k, g(f^k)\right\rangle \leq G(f^{k+1}) - G(f^k) = 0.$$

Dividing (13.5) by $S(f^{k+1})$ gives $F(f^{k+1}) < F(f^k)$. As the sequence $F(f^k)$ is bounded from below and monotonically decreasing and thus converging and $S(f^{k+1})$ is bounded on the constraint set, we get the convergence result from

$$\lambda^{k+1} S(f^{k+1}) - \lambda^k S(f^{k+1}) \leq c^k \left\langle f^{k+1} - f^k, g(f^k)\right\rangle \leq 0.$$

□

If we choose $G(u) = \frac{1}{2}\|u\|_2^2$ we get $g(f^k) = f^k$ and if c^k is bounded from below $\|f^{k+1} - f^k\|_2 \to 0$ as in the case of [2, 3] but we can show, that this convergence holds for any strictly convex function G.

Proposition 13.2. *If G is strictly convex and $c^k \geq \gamma > 0$ for all k, then any sequence f^k produced by RatioDCA-prox fulfills $\|f^{k+1} - f^k\|_2 \to 0$.*

Proof. As in the proof of Proposition 13.1, we have $\left\langle g(f^k), f^{k+1} - f^k\right\rangle \leq 0$ and $G(f^{k+1}) = G(f^k) = 1$. Suppose $f^{k+1} \in G_\varepsilon := \{u | G(u) = 1, \|u - f^k\| \geq \varepsilon\}$. If $\left\langle g(f^k), f^{k+1} - f^k\right\rangle = 0$, then the first order condition yields for $0 < t < 1$

$$G(f^k + t(f^{k+1} - f^k)) \geq G(f^k) + \left\langle g(f^k), t(f^{k+1} - f^k)\right\rangle = G(f^k) = 1,$$

which is a contradiction to the strict convexity of G as for $0 < t < 1$,

$$G(f^k + t(f^{k+1} - f^k)) < (1-t)G(f^k) + tG(f^{k+1}) = 1.$$

Thus with the compactness of G_ε we get

$$\left\langle g(f^k), f^{k+1} - f^k\right\rangle \leq \max_{u \in G_\varepsilon} \left\langle g(f^k), u - f^k\right\rangle =: \delta < 0.$$

However, with $c^k \geq \gamma > 0$ for all k this contradicts for k large enough the result $\left\langle f^{k+1} - f^k, g(f^k)\right\rangle \to 0$ as $k \to \infty$ of Proposition 13.1. Thus under the stated conditions $\|f^{k+1} - f^k\|_2 \to 0$ as $k \to \infty$. □

While the previous result does not establish convergence of the sequence, it establishes that the set of accumulation points has to be connected.

As we are interested in minimizing the ratio F we want to find vectors f with $S(f) \neq 0$

Lemma 13.4. *If $S(f^0) \neq 0$ then every vector in the sequence f^k produced by RatioDCA-prox fulfills $S(f^k) \neq 0$.*

Proof. As R_2 and S_1 are one-homogeneous and $G(f^k) = 1$, we have for any vector h with $S(h) = 0$ and $G(h) = 1$,

$$\begin{aligned}\Phi^{c^k}_{f^k}(h) &\geq R_1(h) - R_2(h) + \lambda^k(S_2(h) - S_1(h)) - c^k \left\langle h, g(f^k)\right\rangle \\ &\geq R(h) - c^k \left\langle f^k, g(f^k)\right\rangle \geq -c^k \left\langle f^k, g(f^k)\right\rangle = \Phi^{c^k}_{f^k}(f^k)\end{aligned}$$

where we have used that $\left\langle g(f^k), h\right\rangle \leq G(h) - G(f^k) + \left\langle f^k, g(f^k)\right\rangle = \left\langle f^k, g(f^k)\right\rangle$. Further, if f^k is a minimizer then the algorithm terminates. □

13.3.4 Choice of the Constraint Set and the Proximal Term

While the iterates f^k and thus the final result of RatioDCA and RatioDCA-prox differ in general, the following lemma shows that termination of RatioDCA implies termination of RatioDCA-prox and under some conditions also the reverse implication holds true. Thus switching from RatioDCA to RatioDCA-prox at termination does not allow to get further descent.

Lemma 13.5. *Let f_2^k, $\|f_2^k\|_2 = 1$, $f_1^k = \frac{f_2^k}{G(f_2^k)^{\frac{1}{p}}}$, $c^k \geq 0$, $s_1(f_2^k) = s_1(f_1^k), r_2(f_2^k) = r_2(f_1^k)$ as in the algorithm RatioDCA-prox and*

$$\Omega_1 = \operatorname*{arg\,min}_{G(u)\leq 1} \Phi^{c^k}_{f_1^k}(u), \quad \text{and} \quad \Omega_2 = \operatorname*{arg\,min}_{\|u\|_2\leq 1} \Phi^0_{f_2^k}(u).$$

Then the following implications hold:

1. *If $f_2^k \in \Omega_2$ then $f_1^k \in \Omega_1$.*
2. *If $f_1^k \in \Omega_1$ and either $\partial G(f_1^k) = \{g(f_1^k)\}$ or $c^k = 0$ then $f_2^k \in \Omega_2$.*

Proof. If $f_2^k \in \Omega_2$ then $\Phi^0_{f_2^k}(f_2^k) = 0$. As $\Phi^0_{f_2^k}$ is one-homogeneous, f_2^k is also a global minimizer and thus for all $u \in \mathbb{R}^V$ with $G(u) \leq 1$, $\Phi^{c^k}_{f_1^k}(u) = \Phi^0_{f_1^k}(u) - c^k \left\langle g(f_1^k), u\right\rangle \geq -c^k \left\langle g(f_1^k), u\right\rangle \geq -c^k p$. As $\left\langle g(f_1^k), f_1^k\right\rangle = p$, f_1^k is a minimizer which proves the first part.
On the other hand if

$$f_1^k \in \operatorname*{arg\,min}_{G(u)\leq 1} \Phi^{c^k}_{f_1^k}(u),$$

then by Lemma 13.1 also

$$f_1^k \in \arg\min_u \left\{ \Phi^{c^k}_{f_1^k}(u) + c^k G(u) \right\}.$$

f_1^k being a global minimizer implies

$$0 \in \partial \left(\Phi^{c^k}_{f_1^k} + c^k G \right)(f_1^k) = \partial \Phi^0_{f_1^k}(f_1^k) - c^k g(f_1^k) + c^k \partial G(f_1^k) = \partial \Phi^0_{f_1^k}(f_1^k),$$

where we used that by assumption $c^k(g(f_1^k) - \partial G(f_1^k)) = 0$. Thus f_1^k is also a minimizer of $\Phi^0_{f_1^k}$ and the result follows with $\Phi^0_{f_1^k}(f_1^k) = \Phi^0_{f_1^k}\left(f_2^k\right) = 0$ and $\Phi^0_{f_1^k} = \Phi^0_{f_2^k}$. □

13.3.5 Nonlinear Eigenproblems

The sequence $F(f^k)$ is not only monotonically decreasing but we also show now that the sequence f^k converges to a generalized nonlinear eigenvector as introduced in [8].

Theorem 13.2. *Each cluster point f^* of the sequence f^k produced by RatioDCA-prox fulfills for a c^* and with $\lambda^* = \frac{R(f^*)}{S(f^*)} \in \left[0, F(f^0)\right]$*

$$0 \in \partial\big(R_1(f^*) + c^* G(f^*)\big) - \partial\big(R_2(f^*) + c^* G(f^*)\big) - \lambda^* \big(\partial S_1(f^*) - \partial S_2(f^*)\big).$$

If for every f with $G(f) = 1$ the subdifferential $\partial G(f)$ is unique or $c^k = 0$ for all k, then f^* is an eigenvector with eigenvalue λ^* in the sense that it fulfills

$$0 \in \partial R_1(f^*) - \partial R_2(f^*) - \lambda^* \big(\partial S_1(f^*) - \partial S_2(f^*)\big). \tag{13.6}$$

Proof. By Proposition 13.1 the sequence $F(f^k)$ is monotonically decreasing. By assumption $S = S_1 - S_2$ and $R = R_1 - R_2$ are nonnegative and hence F is bounded below by zero. Thus we have convergence towards a limit

$$\lambda^* = \lim_{k \to \infty} F(f^k).$$

Note that f^k is contained in a compact set, which implies that there exists a subsequence f^{k_j} converging to some element f^*. As the sequence $F(f^{k_j})$ is a subsequence of a convergent sequence, it has to converge towards the same limit, hence also

$$\lim_{j \to \infty} F(f^{k_j}) = \lambda^*.$$

Assume now that for all c it holds $\min_{G(u)\leq 1} \Phi^c_{f^*}(u) < \Phi^c_{f^*}(f^*)$. Then by Proposition 13.1, any vector $f^{(c)} \in \arg\min_{G(u)\leq 1} \Phi^c_{f^*}(u)$ satisfies

$$F(f^{(c)}) < \lambda^* = F(f^*)\,,$$

which is a contradiction to the fact that the sequence $F(f^k)$ has converged to λ^*. Thus there exists c^* such that $f^* \in \arg\min_{G(u)\leq 1} \left\{\Phi^{c^*}_{f^*}(u)\right\}$ and by Lemma 13.1 then $f^* \in \arg\min_u \left\{\Phi^{c^*}_{f^*}(u) + c^* G(u)\right\}$ and we get

$$0 \in \partial R_1(f^*) - r_2(f^*) + \lambda^* \left(\partial S_2(f^*) - s_1(f^*)\right) - c^* g(f^*) + c^* \partial G(f^*).$$

If $c^k = 0$ for all k then we only need to look at $c^* = 0$. In this case or if we get from $G(f^*) = 1$ that $\partial G(f^*) = \{g(f^*)\}$ it follows that

$$0 \in \partial R_1(f^*) - r_2(f^*) + \lambda^* \left(\partial S_2(f^*) - s_1(f^*)\right)$$

which then implies that f^* is an eigenvector of F with eigenvalue λ^*. □

Remark 13.2. Equation (13.6) is a necessary condition for f^* being a critical point of F. If R_2, S_1 are continuously differentiable at f^*, it is also sufficient. The necessity of (13.6) follows from [6, Proposition 2.3.14]. If R_2, S_1 are continuously differentiable at f^* then we get from [6, Propositions 2.3.6 and 2.3.14] that $0 \in \partial F(f^*)$ and f^* is a critical point of F.

13.4 The RatioDCA-Prox for Ratios of Lovasz Extensions: Application to Balanced Graph Cuts

A large class of combinatorial problems [5, 9] allows for an exact continuous relaxation which results in a minimization problem of a non-negative ratio of Lovasz extensions as introduced in Sect. 13.1. In this paper, we restrict ourselves to balanced graph cuts even though most statements can be immediately generalized to the class of problems considered in [5].

We first collect some important properties of Lovasz extensions before we prove stronger results for the RatioDCA-prox when applied to minimize a non-negative ratio of Lovasz extensions.

13.4.1 Properties of the Lovasz Extension

The following lemma is a reformulation of [1, Proposition 4.2(c)] for our purposes:

Lemma 13.6. *Let $\hat{S}$ be a submodular function with $\hat{S}(\emptyset) = \hat{S}(V) = 0$. If S is the Lovasz extension of $\hat{S}$ then*

$$\langle \partial S(f), \mathbf{1}_{C_i} \rangle = S(\mathbf{1}_{C_i}) = \hat{S}(C_i)$$

for all sets $C_i = \{j \in V | f_j > f_i\}$.

Proof. Let wlog f be in increasing order $f_1 \leq f_2 \leq \cdots \leq f_n$. With $f = \sum_{i=1}^{n-1} \mathbf{1}_{C_i}(f_{i+1} - f_i) + \mathbf{1}_V \cdot f_1$ we get

$$\sum_{i=1}^{n} \hat{S}(C_i)(f_{i+1} - f_i) = S(f) = \langle \partial S(f), f \rangle = \sum_{i=1}^{n-1} \langle \partial S(f), \mathbf{1}_{C_i} \rangle (f_{i+1} - f_i).$$

Since $\hat{S}$ is submodular S is convex and thus $\langle \partial S(f), \mathbf{1}_{C_i} \rangle \leq S(\mathbf{1}_{C_i}) = \hat{S}(C_i)$, but because $f_{i+1} - f_i \geq 0$ this holds with equality in all cases. □

More generally this also holds if $\hat{S}$ is not submodular:

Lemma 13.7. *Let $\hat{S}$ be a set function with $\hat{S}(\emptyset) = \hat{S}(V) = 0$. If S is the Lovasz extension of $\hat{S}$ then*

$$\langle \partial S(f), \mathbf{1}_{C_i} \rangle = \hat{S}(C_i)$$

for all sets $C_i = \{j \in V | f_j > f_i\}$.

Proof. $\hat{S}$ can be written as the difference of two submodular set functions $\hat{S} = \hat{S}_1 - \hat{S}_2$ and the Lovasz extension S of $\hat{S}$ is the difference of the corresponding Lovasz extensions S_1 and S_2. We get $\partial S(f) \subseteq \partial S_1(f) - \partial S_2(f)$ [6, Propositions 2.3.1 and 2.3.3] and both S_1 and S_2 fulfill the conditions of Lemma 13.6. Thus

$$\begin{aligned} \langle \partial S(f), \mathbf{1}_{C_i} \rangle &\subseteq \langle \partial S_1(f) - \partial S_2(f), \mathbf{1}_{C_i} \rangle = \langle \partial S_1(f), \mathbf{1}_{C_i} \rangle - \langle \partial S_2(f), \mathbf{1}_{C_i} \rangle \\ &= S_1(\mathbf{1}_{C_i}) - S_2(\mathbf{1}_{C_i}) = S(\mathbf{1}_{C_i}) \end{aligned}$$

and the claim follows since $\partial S(f)$ is nonempty [6, Proposition 2.1.2]. □

Also Lovasz extensions are maximal in the considered class of functions:

Lemma 13.8. *Let $\hat{S}$ be a symmetric set function with $\hat{S}(\emptyset) = 0$, S_L its Lovasz extension and S any extension fulfilling the properties of Theorem 13.1, that is S is one-homogeneous, even, convex and $S(f + \alpha \mathbf{1}) = S(f)$ for all $f \in \mathbb{R}^V$, $\alpha \in \mathbb{R}$ and $\hat{S}(A) := S(\mathbf{1}_A)$ for all $A \subset V$. Then $S_L(f) \geq S(f)$ for all $f \in \mathbb{R}^V$.*

Proof. By Lemma 13.7 and using the convexity and one-homogeneity of S we get

$$S_L(f) = \sum_{i=1}^{n-1} \hat{S}(C_i)(f_{i+1} - f_i) = \sum_{i=1}^{n-1} S(\mathbf{1}_{C_i})(f_{i+1} - f_i)$$
$$\geq \sum_{i=1}^{n-1} \langle \partial S(f), \mathbf{1}_{C_i} \rangle (f_{i+1} - f_i) = \langle \partial S(f), f \rangle = S(f)$$

□

Remark 13.3. By [9, Lemma 3.1] any function S fulfilling the properties of the lemma can be rewritten by $S(f) = \sup_{u \in U} \langle u, f \rangle$ where $U \subset \mathbb{R}^n$ is a closed symmetric convex set and $\langle u, \mathbf{1} \rangle = 0$ for all $u \in U$. The previous lemma implies that for a given set function $\hat{S}(C)$ the set U is maximal for the Lovasz extension S_L. In turn this implies that the subdifferential of S_L is maximal everywhere and thus should be used in the RatioDCA-prox. In [3, 9] the authors use for the balancing function $\hat{S}(C) = |C||\overline{C}|$ instead of the Lovasz extension $S_L(f) = \frac{1}{2}\sum_{i,j=1}^{n} |f_i - f_j|$ the convex function $S(f) = \|f - \text{mean}(f)\mathbf{1}\|_1$ which fulfills the properties of the previous lemma. In Sect. 13.5 we show that using the Lovasz extension leads almost always to better balanced graph cuts.

13.4.2 The RatioDCA-Prox for Balanced Graph Cuts

Applied to balanced graph cuts we can show the following "improvement theorem" generalizing the result of [9] for our algorithm. It implies that we can use the result of any other graph partitioning method as initialization and in particular, we can always improve the result of spectral clustering.

Theorem 13.3. *Let $(A, \overline{A})$ be a given partition of V and let $S : V \to \mathbb{R}_+$ satisfy one of the conditions stated in Theorem 13.1. If one uses as initialization of RatioDCA-prox $f^0 = \mathbf{1}_A$, then either the algorithm terminates after one step or it yields an f^1 which after optimal thresholding as in Theorem 13.1 gives a partition $(B, \overline{B})$ which satisfies*

$$\frac{\text{cut}(B, \overline{B})}{\hat{S}(B)} < \frac{\text{cut}(A, \overline{A})}{\hat{S}(A)}.$$

Proof. This follows in the same way from Proposition 13.1 as in [9, Theorem 4.2]. □

In the case that we have Lovasz extensions we can show that accumulation points are directly related to the optimal sets:

Theorem 13.4. *If R_2 and S_1 are Lovasz-extensions of the corresponding set functions then every accumulation point f^* of RatioDCA-prox with $c^k = 0$ fulfills*

$F(f^*) = F(\mathbf{1}_{C^*})$ where C^* is the set we get from optimal thresholding of f^*. If also R_1 and S_2 are the Lovasz-extensions then $f^* = \sum_{i=1}^m \alpha_i \mathbf{1}_{C_i} + b\mathbf{1}_V$ with $\alpha_i > 0$, $C_i = \{j \in V \mid f_j^* > f_i^*\}$, $b \in \mathbb{R}$, and

$$\frac{\hat{R}(C_i)}{\hat{S}(C_i)} = \lambda^* = \frac{R(f^*)}{S(f^*)}, \quad i = 1, \ldots, m.$$

If λ^* is only attained for one set C^* then $f^* = \mathbf{1}_{C^*}$ is the only accumulation point.

Proof. In the proof of Theorem 13.2 it has been shown that from f^* no further descent is possible. Assume $F(f^*) > F(\mathbf{1}_{C^*})$. Then

$$\begin{aligned}\Phi^0_{f^*}(\mathbf{1}_{C^*}) &= R_1(\mathbf{1}_{C^*}) - \langle r_2(f^*), \mathbf{1}_{C^*}\rangle + \lambda^*(S_2(\mathbf{1}_{C^*}) - \langle s_1(f^*), \mathbf{1}_{C^*}\rangle)\\ &= R(\mathbf{1}_{C^*}) - \lambda^* S(\mathbf{1}_{C^*})\\ &< R(\mathbf{1}_{C^*}) - F(\mathbf{1}_{C^*})S(\mathbf{1}_{C^*}) = 0 = \Phi^0_{f^*}(f^*)\end{aligned}$$

which leads to a contradiction. Thus the first claim follows from Theorem 13.1. If also R_1 and S_2 are the Lovasz-extensions then for $f^* = \sum_{i=1}^{n-1} \alpha_i \mathbf{1}_{C_i} + \mathbf{1}_V \cdot \min_j f_j^*$ we get by Lemma 13.6 and the definition of the Lovasz extension that

$$0 = \Phi^0_{f^*}(f^*) = \sum_{i=1}^n \alpha_i \Phi^0_{f^*}(\mathbf{1}_{C_i})$$

and if for one $\alpha_i > 0$ we have $\frac{\hat{R}(C_i)}{\hat{S}(C_i)} > \lambda^*$ then $\Phi^0_{f^*}(\mathbf{1}_{C_i}) > 0$ and we get $\Phi^0_{f^*}(\mathbf{1}_{C^*}) < 0 = \Phi^0_{f^*}(f^*)$ which again is a contradiction. □

Remark 13.4. By Lemma 13.5 this also holds for $c^k > 0$ if G is differentiable at the boundary.

If we have Lovasz extensions we can also use the reduced version of the RatioDCA-prox with $c^k = 0$ to guarantee termination. We are thus in the striking situation that in general we can guarantee stronger convergence properties if $c^k \geq \gamma > 0$ for all k by Proposition 13.2 but an even stronger property such as finite convergence can only be proven when $c^k = 0$.

Theorem 13.5. *Let $c^k = 0$ and S_1, R_2 be Lovasz extensions in the RatioDCA-prox. Further, let C_k^* be the set obtained by optimal thresholding of f^k. If in step 5 of RatioDCA-prox we choose, $\lambda^k = F(\mathbf{1}_{C_k^*})$, and in step 4 choose $f^{k+1} = \mathbf{1}^* := \frac{\mathbf{1}_{C_k^*}}{G(\mathbf{1}_{C_k^*})^{\frac{1}{p}}}$ if $\mathbf{1}^* \in \arg\min_{G(u)\leq 1} \Phi^{c^k}_{f^k}$, then the RatioDCA-prox terminates in finitely many steps.*

Proof. With $c^k = 0$ and using Lemma 13.7 and as R_1, S_2 are convex and one-homogeneous, we get

$$
\begin{aligned}
&R(f^{k+1}) - F(\mathbf{1}_{C_k^*})S(f^{k+1}) \leq \Phi_{f^k}^{c^k}(f^{k+1}) \\
&\leq \Phi_{f^k}^{c^k}(\mathbf{1}^*) = R_1(\mathbf{1}^*) - \langle r_2(f^k), \mathbf{1}^* \rangle + F(\mathbf{1}_{C_k^*})\big(S_2(\mathbf{1}^*) - \langle s_1(f^k), \mathbf{1}^* \rangle\big) \\
&= R(\mathbf{1}^*) - F(\mathbf{1}_{C_k^*})S(\mathbf{1}^*) = 0
\end{aligned}
$$

and thus $F(\mathbf{1}_{C_{k+1}^*}) \leq F(f^{k+1}) \leq F(\mathbf{1}_{C_k^*})$ and equality in the second inequality only holds if $f^{k+1} = \mathbf{1}^*$, but then in the next step we either get strict improvement or the sequence terminates. As there are only finitely many different cuts, RatioDCA-prox has to terminate in finitely many steps. □

13.5 Experiments

The convex inner problem in Eq. (13.3) is solved using the primal dual hybrid gradient method (PDHG) as in [9]. In the first iterations the problem is not solved to high accuracy as all results in this paper only rely on the fact that either the algorithm terminates or

$$\phi_{f^k}^{c^k}(f^{k+1}) < \phi_{f^k}^{c^k}(f^k).$$

13.5.1 Influence of the Proximal Term

First, we study the influence of different values of c^k in the RatioDCA-prox algorithm. We choose $G = \|\cdot\|_2^2$ and choose different values for c^k.

We compare the algorithms on the wing graph from [13] (62,032 vertices, 243,088 edges) and a graph built from the two-moons dataset (2,000 vertices, 33,466 edges) as described in [4].

In Table 13.1 we have plotted the resulting ratio cheeger cuts (RCC) of ten different choices of $c^k = c \cdot \lambda^k$ for RatioDCA-prox. In all cases we use one initialization with the second eigenvector of the standard graph Laplacian and 99 initializations with random vectors, which are the same for all algorithms. As one is interested in the best result and how often this can be achieved, we report the best, average and top10 performance. For both graphs there is no clear trend that a particular choice of the proximal term improves or worsens the results compared to $c^k = 0$ which corresponds to the RatioDCA. This confirms the reported results of [2] where also no clear difference between $c^k = 0$ and the general case has been observed.

Table 13.1 Displayed are the averages of all, the 10 best and the best cuts for different values of $c^k = c\lambda^k$ on wing (top) and two-moons (bottom)

Graph\c	0	0.1	0.25	0.5	0.75	1	1.5	2	3	4
Wing										
Avg	2.6683	2.6624	2.6765	2.6643	2.6602	2.6595	2.6566	2.6565	**2.6548**	2.6573
Top 10 avg	2.5554	2.5519	2.5625	2.5533	2.5549	**2.5514**	2.5523	2.5605	2.5555	2.5523
Best cut	2.545	**2.5439**	2.5532	2.5487	2.5451	2.5471	2.5448	2.5539	2.5472	2.5472
Two-moons										
Avg	2.4872	2.4855	2.5017	2.5158	**2.4569**	2.4851	2.4848	2.7868	3.028	2.929
Top 10 avg	**2.448**	2.4485	2.4481	2.4487	2.4484	2.4484	2.4481	2.4492	2.4491	2.4483
Best cut	2.4447	2.4473	2.4472	2.4461	2.4457	2.4476	2.4465	2.4482	2.4478	**2.4441**

Table 13.2 For each graph it is shown how many times for the 11 initializations the RatioDCA-prox with the Lovasz extension performs better/equal/worse than the previously used continuous extension and the ratio of the best solutions of Lovasz vs continuous extension is shown (<100 % means that the Lovasz extension produced a better ratio cut)

Graph	two-moons	whitaker3	uk	4elt	fe_4elt	3elt	crack
Better/equal/worse	11/0/0	11/0/0	0/11/0	10/0/1	0/11/0	10/0/1	10/0/1
Ratio of best cuts (%)	99.41	99.95	100	99.98	100	99.97	99.83

13.5.2 Comparing the Lovasz Extension to Other Extensions

In previous work [3, 9] on the ratio cut with the balancing function $\hat{S}(C) = |C||\overline{C}|$ not the Lovasz extension $S_L(f) = \frac{1}{2}\sum_{i,j=1}^{n} |f_i - f_j|$ has been used but the function $S(f) = \|f - \text{mean}(f)\mathbf{1}\|_1$. As discussed in Sect. 13.4, this should lead to worse performance in the algorithm as the subdifferential of S_L is maximal. In Table 13.2 we compare both extensions with the RatioDCA-prox with $c^k = 0$ and $G(u) = \|u\|_2^2$ on seven different graphs [13]. One initialization is done with the second eigenvector of the standard graph laplacian and the same 10 random initializations are used for both extensions.

While the differences in the best found cut are minor, using the Lovasz extension for the balancing function leads consistently to better results.

References

1. Bach, F.: Learning with submodular functions: a convex optimization perspective. Found. Trends Mach. Learn. **6**(2–3), 145–373 (2008)
2. Bresson, X., Laurent, T., Uminsky, D., von Brecht, J.H.: Convergence and energy landscape for cheeger cut clustering. In: Advances in Neural Information Processing Systems 25 (NIPS), pp. 1394–1402. Curran Associates, Red Hook (2012)
3. Bresson, X., Laurent, T., Uminsky, D., von Brecht, J.H.: Convergence of a steepest descent algorithm for ratio cut clustering (2012). ArXiv:1204.6545v1

4. Bühler, T., Hein, M.: Spectral clustering based on the graph p-Laplacian. In: Proceedings of the 26th International Conference on Machine Learning (ICML), Montreal, pp. 81–88 (2009)
5. Bühler, T., Rangapuram, S., Setzer, S., Hein, M.: Constrained fractional set programs and their application in local clustering and community detection. In: Proceedings of the 30th International Conference on Machine Learning (ICML), Atlanta, pp. 624–632 (2013)
6. Clarke, F.: Optimization and Nonsmooth Analysis. Wiley, New York (1983)
7. Guattery, S., Miller, G.L.: On the quality of spectral separators. SIAM J. Matrix Anal. Appl. **19**, 701–719 (1998)
8. Hein, M., Bühler, T.: An inverse power method for nonlinear eigenproblems with applications in 1-spectral clustering and sparse PCA. In: Advances in Neural Information Processing Systems 23 (NIPS), pp. 847–855. Curran Associates, Red Hook (2010)
9. Hein, M., Setzer, S.: Beyond spectral clustering – tight relaxations of balanced graph cuts. In: Advances in Neural Information Processing Systems 24 (NIPS), pp. 2366–2374. Neural Information Processing Systems/Curran Associates, La Jolla/Red Hook (2011)
10. Hein, M., Setzer, S., Jost, L., Rangapuram, S.: The total variation on hypergraphs – learning on hypergraphs revisited. In: Advances in Neural Information Processing Systems 26 (NIPS), pp. 2427–2435 (2013)
11. Szlam, A., Bresson, X.: Total variation and Cheeger cuts. In: Proceedings of the 27th International Conference on Machine Learning (ICML), Haifa, pp. 1039–1046 (2010)
12. von Luxburg, U.: A tutorial on spectral clustering. Stat. Comput. **17**, 395–416 (2007)
13. Walshaw, C.: The graph partitioning archive (2004). Staffweb.cms.gre.ac.uk/~c.walshaw/partition/
14. Yang, F., Wei, Z.: Generalized Euler identity for subdifferentials of homogeneous functions and applications. J. Math. Anal. Appl. **337**, 516–523 (2008)

Chapter 14
Adaptive Approximation Algorithms for Sparse Data Representation

Mijail Guillemard, Dennis Heinen, Armin Iske, Sara Krause-Solberg, and Gerlind Plonka

Abstract We survey our latest results on the development and analysis of adaptive approximation algorithms for sparse data representation, where special emphasis is placed on the Easy Path Wavelet Transform (EPWT), nonlinear dimensionality reduction (NDR) methods, and their application to signal separation and detection.

14.1 Introduction

During the last few years there has been an increasing interest in efficient (i.e., sparse) representation and denoising of high-dimensional signals. We have focussed our research on the development and analysis of adaptive approximation algorithms for high-dimensional signals, especially (a) scattered data denoising by wavelet transforms; (b) nonlinear dimensionality reduction relying on geometrical and topological concepts. This contribution reviews our recent research results on (a) and (b).

For (a), we present a general framework for the *Easy Path Wavelet Transform* (EPWT) for sparse representation and denoising of scattered data taken from high-dimensional signals (in Sect. 14.2). As regards (b), we continue our research on nonlinear dimensionality reduction (NDR) methods (cf. Sect. 14.3), where we combine recent NDR methods with non-negative matrix factorization (NNMF), for the purpose of separating sources from a mixture of signals without a prior knowledge about the mixing process. More details on dimensionality reduction

M. Guillemard
Technical University of Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: guillemard@math.tu-berlin.de

D. Heinen • G. Plonka
University of Göttingen, Lotzestr. 16-18, 37083 Göttingen, Germany
e-mail: d.heinen@math.uni-goettingen.de; plonka@math.uni-goettingen.de

A. Iske (✉) • S. Krause-Solberg
University of Hamburg, Bundesstrasse 55, 20146 Hamburg, Germany
e-mail: armin.iske@uni-hamburg.de; sara.krause-solberg@uni-hamburg.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_14

and NNMF, along with our recent results on signal separation, are discussed in Sect. 14.4.

The presented results are based on our papers [7–9, 11, 13, 17–21, 25] and have been achieved in the project "Adaptive approximation algorithms for sparse data representation" of the German Research Foundation's priority program DFG-SPP 1324.

14.2 The Easy Path Wavelet Transform

Let Ω be a connected domain in $\mathbb{R}^d$ and let Γ be a large finite set of points in Ω. We let $h_\Gamma := \max_{y\in\Omega} \min_{x\in\Gamma} \|y - x\|_2$ be the *fill distance* of Γ in Ω and its *grid distance* is $g_\Gamma := \min_{x,x'\in\Gamma, x\neq x'} \|x - x'\|_2$. We say that the set Γ is *quasi-uniform*, if $h_\Gamma < 2g_\Gamma$. Further, let $f : \Omega \to \mathbb{R}$ be a piecewise smooth function that is sampled at Γ, i.e., the values $f(x)$, $x \in \Gamma$, are given. We are now interested in an efficient approximation of f using a meshless multiscale approach called Easy Path Wavelet Transform (EPWT). For applications, we usually assume that Γ approximates a smooth manifold in $\mathbb{R}^d$. For example, our approach covers the efficient approximation of digital images, see [16, 20], where Γ is chosen to be a set of regular grid points in a rectangle Ω, and the approximation of piecewise smooth functions on the sphere, see [18], where $\Omega = \mathbb{S}^2$ and Γ is a suitably chosen quasi-uniform point set on the sphere $\mathbb{S}^2$.

Similar approaches have also been proposed for generalizing the wavelet transform to data defined on weighted graphs, see [22]. In this section, we extend the EPWT proposed in [16, 18, 24] to the case of high-dimensional data approximation.

14.2.1 The General EPWT Algorithm for Sparse Approximation

Let us shortly recall the notions of a biorthogonal wavelet filter bank of perfect reconstruction. To this end, let φ be a sufficiently smooth, compactly supported, one-dimensional scaling function, $\tilde{\varphi}$ the corresponding biorthogonal compactly supported scaling function, and ψ, $\tilde{\psi}$ the corresponding pair of biorthogonal compactly supported wavelets, see, e.g., [2, 15]. These functions provide us with a filter bank of perfect reconstruction with sequences $(h_n)_{n\in\mathbb{Z}}$, $(\tilde{h}_n)_{n\in\mathbb{Z}}$ of low-pass filter coefficients and $(g_n)_{n\in\mathbb{Z}}$, $(\tilde{g}_n)_{n\in\mathbb{Z}}$ of high-pass filter coefficients.

Assume that the number $N = |\Gamma|$ of given points $x \in \Gamma \subset \mathbb{R}^d$ is a power of 2, $N = 2^J$, where $J \gg 1$. We denote $\Gamma^J := \Gamma$ and its elements by $x_k^J = x_k$, $k = 1, \ldots, N$, i.e., we fix some ordering of the points in Γ^J. Now the EPWT works as follows. In a first step, we seek a suitable permutation p^J of the indices of the points in Γ^J by determining a path of length N through all points x_k^J such that consecutive data points $(x^J_{p^J(k)}, f(x^J_{p^J(k)}))$ and $(x^J_{p^J(k+1)}, f(x^J_{p^J(k+1)}))$ in the path

Algorithm 9 Decomposition

Let $\Gamma = \{x_1, \ldots, x_N\} = \{x_1^J, \ldots, x_N^J\} = \Gamma^J \subset \mathbb{R}^d$ be a given point set. Let $f_k^J := f(x_k)$, for $k = 1, \ldots, N$, where $N = 2^J$. Choose a biorthogonal wavelet filterbank with decomposition filters $\tilde{h}, \tilde{g}$, and reconstruction filters h, g, where $\sum_{k\in\mathbb{Z}} \tilde{h}(k) = \sqrt{2}$, and a low-pass filter $\tilde{h}_p$, where $\sum_{k\in\mathbb{Z}} \tilde{h}_p(k) = 1$.
Iteration: Perform the following 4 steps for $\ell = J, J-1 \ldots, J-L+1$ with $L < J$:

1. Find a suitable path vector $p^\ell \in \mathbb{N}^{2^\ell}$ consisting of a permutation of the indices of the points in Γ^ℓ that describes a fixed order of points $(x^\ell_{p^\ell(k)}, f^\ell_{p^\ell(k)})$, $k = 1, \ldots, 2^\ell$.
2. Apply the (periodic) low-pass filter $\tilde{h}$ to $(f^\ell_{p^\ell(k)})_{k=1}^{2^\ell}$ followed by downsampling by two to obtain the low-pass data $(f_k^{\ell-1})_{k=1}^{2^{\ell-1}}$. Apply the (periodic) high-pass filter $\tilde{g}$ to $(f^\ell_{p^\ell(k)})_{k=1}^{2^\ell}$ followed by downsampling by two to obtain the vector of wavelet coefficients $(d_k^{\ell-1})_{k=1}^{2^{\ell-1}}$.
3. Apply the low-pass filter $\tilde{h}_p$ to point vector $(x^\ell_{p^\ell(k)})_{k=1}^{2^\ell}$ (component-wise) followed by downsampling by two to obtain a new vector of scattered points $(x_k^{\ell-1})_{k=1}^{2^{\ell-1}}$.
Determine the new point set $\Gamma^{\ell-1} := \{x_1^{\ell-1}, \ldots, x_{2^{\ell-1}}^{\ell-1}\}$.
4. Apply a hard-threshold operator T_θ to the wavelet vector $(d_k^{\ell-1})_{k=1}^{2^{\ell-1}}$ to find

$$\tilde{d}_k^{\ell-1} = T_\theta(d_k^{\ell-1}) = \begin{cases} d_k^{\ell-1} & \text{if } |d_k^{\ell-1}| \geq \theta, \\ 0 & \text{if } |d_k^{\ell-1}| < \theta, \end{cases}$$

with a predefined threshold parameter $\theta > 0$.

Output: low-pass function values $(f_k^{J-L})_{k=1}^{2^{J-L}}$, thresholded high-pass function values $(\tilde{d}_k^\ell)_{k=1}^{2^\ell}$, $\ell = J-1, \ldots, J-L$, path vectors p^ℓ, $\ell = J, \ldots, J-L+1$.

strongly "correlate". In the second step, we apply the one-dimensional wavelet filter bank to the sequence of functions values $(f(x^J_{p^J(k)}))_{k=1}^N$, and simultaneously a low-pass filter to the points $(x^J_{p^J(k)})_{k=1}^N$, where we consider each of the d components separately. The significant high-pass coefficients corresponding to the function values will be stored. The $N/2$ low-pass data will be processed further at the next level of the EPWT. Particularly, we denote the set of the $N/2$ points obtained by low-pass filtering and downsampling of $(x^J_{p^J(k)})_{k=1}^N$ by Γ^{J-1}, and relate the low-pass function coefficients to these points. Again, we start with seeking a permutation p^{J-1} of the indices of the points in Γ^{J-1} to obtain an appropriate ordering of the data and apply the one-dimensional wavelet filter bank to the ordered low-pass function data. We iterate this procedure and obtain a sparse representation of the original data by applying a hard thresholding procedure to the high-pass coefficients of the function value components. The complete procedure is summarized by Algorithm 9.

By construction many high pass values d_k^ℓ will vanish. An optimal storage of the path vectors p^ℓ depends on the original distribution of the points x_k^J and on the applied filter $\tilde{h}_p$. Employing a "lazy" filter, we have $x_k^\ell := x^{\ell+1}_{p^{\ell+1}(2k)}$, such that at each level the new point set is just a subset of that of the preceding level of half cardinality.

Algorithm 10 Reconstruction

Reconstruct values $f(x_k) = f(x_k^J)$ by applying the following iteration, where $(\tilde{f}_k^{J-L})_{k=1}^{2^{J-L}} := (f_k^{J-L})_{k=1}^{2^{J-L}}$.
Iteration: Perform the following three steps for $\ell = J-L, J-L+1, \ldots, J-1$:

1. Apply an upsampling by two and then the low-pass filter h to $(\tilde{f}_k^{\ell})_{k=1}^{2^{\ell}}$.
2. Apply an upsampling by two and then the high-pass filter g to $(\tilde{d}_k^{\ell})_{k=1}^{2^{\ell}}$.
3. Add the results of the previous two steps to obtain $(\tilde{f}_{p^{\ell+1}(k)}^{\ell+1})_{k=1}^{2^{\ell+1}}$, and invert permutation $p^{\ell+1}$.

Output: $(\tilde{f}_k^J)_{k=1}^N$, the approximated function values at scattered points $x_k \in \Gamma$.

14.2.2 Construction of Path Vectors

The main challenge for the application of the EPWT to sparse data representation is to construct path vectors through the point sets Γ^ℓ, $\ell = J, \ldots, J-L+1$. This step is crucial for the performance of the data compression. The path construction is based on determining a suitable correlation measure that takes the local distance of the scattered points x_k^ℓ into account, on the one hand, and the difference of the corresponding low-pass values f_k^ℓ, on the other hand. In the following, we present some strategies for path construction and comment on their advantages and drawbacks.

14.2.2.1 Path Construction with Fixed Local Distances

One suitable strategy for path construction [16, 24] is based on a priori fixed local ε-neighborhoods of the points x_k^ℓ. In $\mathbb{R}^d$, we consider a neighborhood of the form

$$N_\varepsilon(x_k^\ell) = \{x \in \Gamma^\ell \setminus \{x_k^\ell\} : \|x_k^\ell - x\|_2 \leq \operatorname{mMn} 2^{(J-\ell)/d}\varepsilon\},$$

where $\varepsilon > 2^{J/d}\, g_\Gamma$ depends on the distribution of the original point set $\Gamma = \Gamma^J$. For example, starting with a regular rectangular grid in $\mathbb{R}^2$ with mesh size $g_\Gamma = 2^{-J/2}$ (with J even) in both directions, one may think about a constant ε with $\sqrt{2} \leq \varepsilon < 2$, such that each inner grid point has eight neighbors.

For path construction at level ℓ of the EPWT, we choose a first point $x^\ell \in \Gamma^\ell$ randomly, and put $x_{p^\ell(1)}^\ell := x^\ell$. Let now $P_j^\ell := \{x_{p^\ell(1)}^\ell, \ldots, x_{p^\ell(j)}^\ell\}$ be the set of points that have already been taken in the path. Now, we determine the $(j+1)$-th point by

$$x_{p^\ell(j+1)}^\ell := \operatorname*{argmin}_{x \in N_\varepsilon(x_{p^\ell(j)}^\ell) \setminus P_j^\ell} |f(x) - f(x_{p^\ell(j)}^\ell)|, \tag{14.1}$$

i.e., we choose the point x in the neighborhood of the point $x^\ell_{p^\ell(j)}$ with minimal absolute difference of the corresponding function values. This measure has been applied in the *rigorous EPWT* of [16, 18]. The advantage of fixing the local neighborhood in spatial domain lies in the reduced storage costs for the path vector that needs to be kept to ensure a reconstruction. The drawback of this measure is that the set of "admissible points" $N_\varepsilon(x^\ell_{p^\ell(j)}) \setminus P^\ell_j$ may be empty. In this case a different rule for finding the next path entry has to be applied.

A special measure occurs if one tries to mimic the one-dimensional wavelet transform. In order to exploit the piecewise smoothness of the function f to be approximated, one should prefer to construct path vectors, where locally three consecutive points $x^\ell_{p^\ell(j-1)}$, $x^\ell_{p^\ell(j)}$, $x^\ell_{p^\ell(j+1)}$ lie (almost) on a straight line. This consideration leads to the following measure: We fix a threshold μ for the function values. For finding the next point in the path, we compute

$$N_{\varepsilon,\mu}(x^\ell_{p^\ell(j)}) := \{x \in N_\varepsilon(x^\ell_{p^\ell(j)}) \setminus P^\ell_j : |f(x) - f(x^\ell_{p^\ell(j)})| \le \mu\}, \tag{14.2}$$

and then let

$$x^\ell_{p^\ell(j+1)} := \operatorname*{argmin}_{x \in N_{\varepsilon,\mu}(x^\ell_{p^\ell(j)})} \frac{\langle x^\ell_{p^\ell(j-1)} - x^\ell_{p^\ell(j)}, x^\ell_{p^\ell(j)} - x\rangle}{\|x^\ell_{p^\ell(j-1)} - x^\ell_{p^\ell(j)}\|_2 \, \|x^\ell_{p^\ell(j)} - x\|_2}, \tag{14.3}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual scalar product in $\mathbb{R}^d$. Note that in (14.3) the cosine of the angle between the vectors $x^\ell_{p^\ell(j-1)} - x^\ell_{p^\ell(j)}$ and $x^\ell_{p^\ell(j)} - x$ is minimized if $x^\ell_{p^\ell(j-1)}$, $x^\ell_{p^\ell(j)}$ and x are co-linear. This approach is taken in [16, 24] for images (called *relaxed EPWT*), and in [10] for scattered data denoising.

Remark 14.1. The idea to prefer path vectors, where the angles between three consecutive points in the path is as large as possible, can be theoretically validated in different ways. Assume that the given wavelet decomposition filter $\tilde{g} = (\tilde{g}_k)_{k\in\mathbb{Z}}$ in the filter bank satisfies the moment conditions $\sum_{k\in\mathbb{Z}} \tilde{g}_k = 0$ and $\sum_{k\in\mathbb{Z}} k\tilde{g}_k = 0$. Then we simply observe that for a constant function $f(x) = c$ for $x \in \Gamma$ and $c \in \mathbb{R}$ by

$$d^J_n = \sum_{k\in\mathbb{Z}} \tilde{g}_{k-2n+1} f(x^J_{p^J(k)}) = c \sum_{k\in\mathbb{Z}} \tilde{g}_{k-2n+1} = 0$$

all wavelet coefficients vanish, while for a linear function of the form $f(x) = a^T x + b$ with $a \in \mathbb{R}^d$ and $b \in \mathbb{R}$ we have

$$d^J_n = \sum_{k\in\mathbb{Z}} \tilde{g}_{k-2n+1} f(x^J_{p^J(k)}) = a^T \sum_{k\in\mathbb{Z}} \tilde{g}_{k-2n+1} x^J_{p^J(k)} + b \sum_{k\in\mathbb{Z}} \tilde{g}_{k-2n+1}.$$

Consequently, these coefficients only vanish, if the points in the sequence $(x^J_{p^J(k)})_{k\in\mathbb{Z}}$ are co-linear and equidistant, see [10]. A second validation for choosing the path vector using the criterion (14.3) is given by the so-called *path smoothness condition* in [17], see also Subsection 14.2.4, Remark 14.4.

Remark 14.2. Our numerical results in Sect. 14.2.5 show that the relaxed path construction proposed in (14.2)–(14.3) is far superior to the rigorous path construction (14.1), since it produces fewer "interruptions", i.e., cases where $N_\varepsilon(x_{p^\ell(j)})\setminus P^\ell_j = \emptyset$, and a new path entry needs to be taken that is no longer locally correlated to the preceding point, which is usually leading to large wavelet coefficients and a higher effort in path coding (see [16, 24]).

14.2.2.2 Path Construction with Global Distances

We want to present a second path construction using a global weight function. Considering the vectors $y^\ell_k = y(x^\ell_k) := ((x^\ell_k)^T, f^\ell_k)^T \in \mathbb{R}^{d+1}$ at each level, we define a symmetric weight matrix $W^\ell = (w(y^\ell_k, y^\ell_{k'}))^{2^\ell}_{k,k'=1}$, where the weight is written as

$$w(y^\ell_k, y^\ell_{k'}) = w_1(x^\ell_k, x^\ell_{k'}) \cdot w_2(f^\ell_k, f^\ell_{k'}).$$

Now the weights for the scattered points x^ℓ_k can be chosen differently from the weights for the (low-pass) function values f^ℓ_k. A possible weight function used already in the context of bilateral filtering [25] is

$$w(y^\ell_k, y^\ell_{k'}) = \exp\left(\frac{-\|x^\ell_k - x^\ell_{k'}\|_2^2}{2^{2(J-\ell)/d}\,\eta_1}\right) \cdot \exp\left(\frac{-|f^\ell_k - f^\ell_{k'}|^2}{2^{J-\ell}\,\eta_2}\right),$$

where η_1 and η_2 need to be chosen appropriately. The normalization constant $2^{2(J-\ell)/d}$ in the weight w_1 is due to the reduction of the points $x \in \Gamma^\ell$ by factor 2, at each level, so that the distances between the points grow. The normalization constant $2^{J-\ell}$ in the weight w_2 arises from the usual amplification of the low-pass coefficients in the wavelet transform with filters $\tilde{h}$ satisfying $\sum_{k\in\mathbb{Z}} \tilde{h}_k = \sqrt{2}$.

Having computed the weight matrix $W^\ell = (w(y^\ell_k, y^\ell_{k'}))^{2^\ell}_{k,k'=1}$, we simply compute the path vector as follows. We choose the first component $x^\ell_{p^\ell(1)}$ randomly from Γ^ℓ. Using again the notation $P^\ell_j := \{x^\ell_{p^\ell(1)}, \ldots, x^\ell_{p^\ell(j)}\}$ for the set of points in Γ^ℓ that are already contained in the path vector, we now determine the next point as

$$x^\ell_{p^\ell(j+1)} := \operatorname*{argmax}_{x\in\Gamma^\ell\setminus P^\ell_j} w(y(x), y(x^\ell_{p^\ell(j)})),$$

where uniqueness can be achieved by fixing a rule if the maximum is attained at more than one point. The advantage of this path construction is that no

"interruptions" occur. The essential drawback consists in higher storage costs for path vectors, where we can no longer rely on direct local neighborhood properties of consecutive points in the path vector. Further, computing the full weight matrix W^ℓ is very expensive. The costs can be reduced by cutting the spatial weight at a suitable distance defining

$$w_1(x_k^\ell, x_{k'}^\ell) = \begin{cases} \exp(-\|x_k^\ell - x_{k'}^\ell\|_2^2 / 2^{2(J-\ell)/d}\,\eta_1) & \text{for } \|x_k^\ell - x_{k'}^\ell\|_2 \le 2^{-\ell/d} D, \\ 0 & \text{for } \|x_k^\ell - x_{k'}^\ell\|_2 > 2^{-\ell/d} D, \end{cases} \tag{14.4}$$

with D chosen appropriately to ensure a sufficiently large spatial neighborhood.

Remark 14.3. This approach has been used in [10] for random path construction, where the compactly supported weight function $w_1(x_k^\ell, x_{k'}^\ell)$ above is employed. Taking the weight function

$$w_1(x_k^\ell, x_{k'}^\ell) = \begin{cases} 1 & \text{for } \|x_k^\ell - x_{k'}^\ell\|_2 \le 2^{-\ell/d} D, \\ 0 & \text{for } \|x_k^\ell - x_{k'}^\ell\|_2 > 2^{-\ell/d} D, \end{cases}$$

and $w_2(f_k^\ell, f_{k'}^\ell) = \exp\left(\frac{-|f_k^\ell - f_{k'}^\ell|^2}{2^{J-\ell}\eta_2}\right)$ we obtain a distance measure that is equivalent to (14.1).

14.2.3 EPWT for Scattered Data Denoising

The EPWT can also be used for denoising of scattered data. Let us again assume $\Gamma = \{x_1, \ldots, x_N\}$ are scattered points in $\mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be a smooth function sampled on $\Gamma \subset \Omega$. For the measured data $\tilde{f}(x_j)$, we suppose that

$$\tilde{f}(x_j) = f(x_j) + z_j,$$

where z_j denotes additive Gaussian noise with zero mean and an unknown variance σ^2. For the distribution of the points in Ω we assume quasi-uniformity as before.

We now apply the EPWT, Algorithms 9 and 10 in Sect. 14.2.1, for data denoising. Note that in case of noisy function values, the construction of path vectors (being based on the correlation of function values at points with small spatial distance) is now influenced by the noise. To improve the denoising performance, we have to resemble the "cycle spinning" method (see [3]) that works as follows. We apply the (tensor product) wavelet shrinkage not only to the image itself, but also to the images that are obtained by up to seven cyclic shifts in x- and y-direction. After un-shifting, one takes the average of the 64 reconstructed images, thereby greatly improving the denoising result.

Employing the EPWT algorithm, we use Algorithms 9 and 10, applying them 64 times using different starting values $x_{p^J(1)}$ as a first path component each time. For the path construction, we utilize one of the two methods described in Sect. 14.2.2. After reconstruction of the 64 data sets, we take the average in order to obtain the denoising result. Similarly as for wavelet denoising, the threshold parameter θ in Algorithm 9 needs to be selected carefully depending on the noise level.

In [10] we have employed two different path constructions for image denoising. The first one is very similar to the path construction in Sect. 14.2.2.1. The second one is based on a weight matrix resembling that in Sect. 14.2.2.2. Here, the next component in the path vector is chosen randomly according to a probability distribution based on the weight matrix.

For images, the proposed denoising procedure strongly outperforms the usual tensor-product wavelet shrinkage with cycle spinning, see [10]. Moreover, the procedure is not restricted to rectangular grids, but can be used in a much more general context for denoising of functions on manifolds. Numerical examples of the EPWT-based denoising scheme are given in Sect. 14.2.5.

14.2.4 Optimal Image Representation by the EPWT

In this subsection we restrict ourselves to the EPWT on digital images on a domain $\Omega = [0,1)^2$. For cartoon models, where the image is piecewise Hölder continuous or even Hölder smooth, we can prove that the EPWT leads to optimally sparse image representations, see [17, 19]. To explain this, let $F \in L^2(\Omega)$ be a piecewise Hölder continuous image. More precisely, let $\{\Omega_i\}_{1\le i\le K}$ be a finite set of regions forming a disjoint partition of Ω whose boundaries are continuous and of finite length. In each region Ω_i, F is assumed to be Hölder continuous of order $\alpha \in (0,1]$,

$$|F(x) - F(x+h)| \le C\,\|h\|_2^{\alpha}, \qquad x,\, x+h \in \Omega_i, \tag{14.5}$$

where $C > 0$ does not depend on i. For given samples $\{(F(2^{-J/2}n))\}_{n\in I_J}$, the function F can be approximated by the piecewise constant function

$$F^J(x) = \sum_{n\in I_J} F(2^{-J/2}n)\,\chi_{[0,1)^2}(2^{J/2}x - n), \qquad x \in [0,1)^2,$$

where the index set $I_J := \{n = (n_1, n_2) \in \mathbb{N}^2 : 0 \le n_1 \le 2^{J/2} - 1,\ 0 \le n_2 \le 2^{J/2} - 1\}$ is of cardinality 2^J. In this special case $\alpha \in (0,1]$ we can rely on the orthogonal Haar wavelet filter bank in Algorithms 9 and 10. An optimal image representation is strongly based on an appropriate path construction. As shown in [19], we need to satisfy the following two conditions.

Region condition. At each level ℓ of the EPWT, we need to choose the path vector, such that it contains at most $R_1 K$ discontinuities which are incurred by crossing over from one region Ω_i to another region, or by jumping within one region Ω_i. Here R_1 does not depend on J or ℓ, and K is the number of regions.

Diameter condition. At each level ℓ of the EPWT, we require

$$\|x_{p^\ell(k)} - x_{p^\ell(k+1)}\|_2 \leq D_1 2^{-\ell/2},$$

for almost all points $x^\ell_{p^\ell(k)}$, $k = 1, \ldots, 2^\ell - 1$, where D_1 does not depend on J or ℓ. The number of path components which do not satisfy the diameter condition is bounded by a constant being independent of ℓ and J.

The region condition suggests that for path construction, we should first collect all points that belong to one region Ω_i before transferring to the next region. The diameter condition ensures that the remaining points in Γ^ℓ are quasi-uniformly distributed at each level ℓ of the EPWT. Satisfying these two conditions for the path vectors, we have shown in [19], Corollary 3.1 that the M-term approximation F_M reconstructed from the M most significant EPWT wavelet coefficients, satisfies the *asymptotically optimal* error estimate

$$\|F - F_M\|_2^2 \leq \tilde{C}\, M^{-\alpha} \tag{14.6}$$

with a constant $\tilde{C}$ and the Hölder exponent $\alpha \in (0, 1]$ in (14.5).

Remark 14.4. Observe that at each level of the EPWT the path vector $(p^\ell(j))_{j=1}^{2^j}$ determines a planar curve that interpolates $f^\ell_{p^\ell(j)}$ at the points $x^\ell_{p^\ell(j)}$, $j = 1, \ldots, 2^\ell$. By definition, this curve is only piecewise linear. A generalization of the optimal M-term approximation result (14.6) for piecewise Hölder smooth images with Hölder exponent $\alpha > 1$ has been developed in [17]. In this case, one needs to generalize the idea of a piecewise linear path vector curve to a smooth path function that satisfies, besides the region condition and the diameter condition, a third condition called *path smoothness condition*, see [17]. More precisely, let us consider a domain $\Omega \subset [0, 1]^2$ with a sufficiently smooth boundary and a disjoint partition Ω_i of Ω with smooth boundaries of finite length. Further, instead of (14.5), we assume that $F \in L^2(\Omega)$ is a piecewise smooth bivariate function being Hölder smooth of order $\alpha > 1$ in each region Ω_i, $i = 1, \ldots, K$. In order to show the optimal error estimate (14.6) also for $\alpha > 1$, we need to employ a path function that approximates the values $f^\ell_{p^\ell(j)}$ at the points $x^\ell_{p^\ell(j)}$ being a planar curve that is not only piecewise smooth but smooth of order α inside a region Ω_i with suitably bounded derivatives, see [17], Section 3.2. Particularly, this condition suggests that one should avoid "small angles" in the path curve.

Fig. 14.1 *Top row:* Reconstruction by tensor-product wavelet compression using the 7–9 biorthogonal filter bank with 1,000 wavelet coefficients for test image `clock` (PSNR 29.93), 700 coeffs for `Lenna` (PSNR 24.28), and 200 coeffs for `sail` (PSNR 19.58). *Bottom row*: Reconstruction by EPWT wavelet transform using the 7–9 biorthogonal filter bank with 1,000 wavelet coefficients for `clock` (PSNR 33.55), 700 coeffs for `Lenna` (PSNR 30.46), 200 coeffs for `sail` (PSNR 27.19)

14.2.5 Numerical Results

We shortly illustrate the performance of the proposed EPWT algorithm for sparse date representation and data denoising. In Fig. 14.1, we illustrate the application of the EPWT for sparse image representation, see also [16, 24]. The three considered images are of size 256×256. In Algorithm 9, we have used the 7–9 biorthogonal filter bank for the function values, and the lazy filter bank for the grid points, i.e., at each level of the EPWT, we have kept only every other grid point. The path construction from Sect. 14.2.2.1 is taken, where in (14.2) the parameters $\varepsilon = \sqrt{2}$ and $\mu = 5$ are employed. The threshold parameter θ in Algorithm 9 is chosen, such that 1,000 most significant EPWT wavelet coefficients are kept for the clock image, 700 coefficients are kept for the Lenna image and 200 coefficients are kept for the sail image. Figure 14.1 shows the reconstructed images, where we compare the results of a tensor-product wavelet compression with the 7–9 biorthogonal filter bank with the results of the EPWT reconstruction, using the same number of wavelet coefficients for the reconstruction in both cases.

In a second example we study the denoising behavior of the EPWT approach as described in Sect. 14.2.3. In Fig. 14.2, we present the noisy pepper image with a PSNR of 19.97 and compare the denoising results of different methods. In particular,

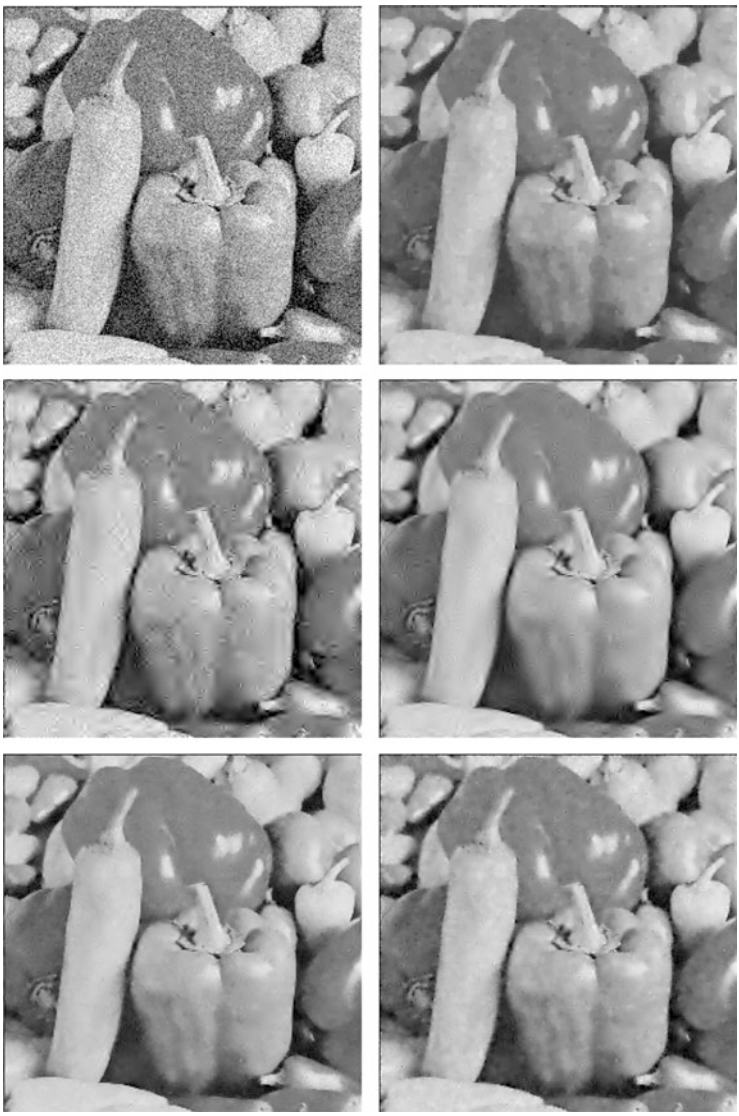


Fig. 14.2 *Top row:* `Peppers` with additive white Gaussian noise with $\sigma = 25$ (PSNR 19.97) and reconstruction by the Four-Pixel Scheme [28] (PSNR 28.26), *Mid row*: Reconstruction by 2d tensor product wavelet transform using the 7–9 biorthogonal filter bank without (PSNR 24.91) and with cycle spinning (PSNR 28.11) *Bottom row*: Reconstruction by our approach described in Sect. 14.2.3 using a relaxed path construction with fixed local distances in (14.2), (PSNR 29.01) and a random path construction based on (14.4) (PSNR 27.96)

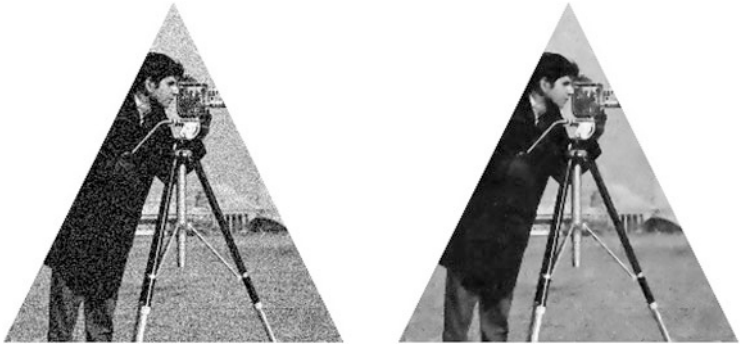


Fig. 14.3 `Cameraman`. Data with additive white Gaussian noise with $\sigma = 25$ (PSNR 19.98), and EPWT reconstruction using the approach in Sect. 14.2.3 (PSNR 26.31)

we have used the four-pixel denoising scheme based on anisotropic diffusion by Welk et al. [28] with 76 iterations and step size 0.001 providing a PSNR of 28.26. Further, we apply the 7–9 wavelet shrinkage with a PSNR of 24.91 and the 7–9 wavelet shrinkage with cycle spinning using 64 shifts of the image and yielding the PSNR 28.11. Our EPWT denoising approach employing a relaxed path construction as described in Sect. 14.2.2.1 achieves a PSNR of 29.01 while a random path construction based on the ideas in Sect. 14.2.2.2 yields the PSNR 27.96. Note that the repeated application of the EPWT shrinkage method can be done in a parallel process. While our proposed EPWT denoising is (due to the path constructions) more expensive than the tensor-product wavelet shrinkage its application is not restricted to rectangular regular grids.

The third example shows the EPWT denoising to a triangular domain taking the approach in Sect. 14.2.3, see Fig. 14.3. We use the 7–9 biorthogonal filter bank for the function values, the lazy filter bank for the grid points, and the path construction from Sect. 14.2.2.1 with $\varepsilon = 1.3$, $\mu = 89$ and threshold $\theta = 89$.

14.3 Dimensionality Reduction on High-Dimensional Signal Data

To explain basic concepts on dimensionality reduction, we regard *point cloud data* as a finite family of vectors

$$X = \{x_i\}_{i=1}^m \subset \mathbb{R}^n$$

contained in an n-dimensional Euclidean space. The fundamental assumption is that X lies in $\mathscr{M}$, a low dimensional (topological) *space* embedded in $\mathbb{R}^n$. Therefore,

$X \subset \mathscr{M} \subset \mathbb{R}^n$ with $p := \dim(\mathscr{M}) \ll n$. Another ingredient is a parameter domain Ω for $\mathscr{M}$, where Ω is assumed to be embedded in a low dimensional space $\mathbb{R}^d$ with $p \leq d < n$. Moreover, we assume the existence of a homeomorphism (diffeomorphism)

$$\mathscr{A} : \Omega \to \mathscr{M},$$

so that Ω is a homeomorphic (diffeomorphic) copy of $\mathscr{M}$. This concept can then be used for signal analysis in a low dimensional environment. In practice, we can only approximate Ω by a projection

$$P : \mathscr{M} \to \Omega',$$

where Ω' is a homeomorphic copy of Ω. The low dimensional structure representing X is the reduced data $Y = \{y_i\}_{i=1}^m \subset \Omega' \subset \mathbb{R}^d$, according to the following diagram.

$$\begin{array}{ccccc} X & \hookrightarrow & \mathscr{M} & \hookrightarrow & \mathbb{R}^n \\ \downarrow{\scriptstyle P|_X} & & \downarrow{\scriptstyle P} & & \\ Y & \hookrightarrow & \Omega' & \hookrightarrow & \mathbb{R}^d \end{array}$$

Principal component analysis (PCA) is a classical linear projection method. Dimensionality reduction by PCA can be described as an eigenvalue problem, so that PCA can be applied by using the *singular value decomposition* (SVD). More precisely, in PCA we consider centered data X (i.e., X has zero mean) in matrix form $X \in \mathbb{R}^{n\times m}$. Now the concept of PCA is to construct a linear projection $P : \mathbb{R}^n \to \mathbb{R}^n$, for $\operatorname{rank}(P) = p < n$, with minimal error $\operatorname{err}(P, X) = \sum_{k=1}^m \|x_k - P(x_k)\|$, or, equivalently, with maximal variance $\operatorname{var}(P, X) = \sum_{k=1}^m \|P(x_k)\|^2$. These conditions can in turn be reformulated as an eigenvalue problem, where the p largest eigenvalues of the covariance matrix $XX^T \in \mathbb{R}^{n\times n}$ are sought, cf. [14].

Another classical linear dimensionally reduction method is *multidimensional scaling* (MDS), which is also relying on an eigendecomposition of data $X \in \mathbb{R}^{n\times m}$. In contrast to PCA, the MDS method constructs a low dimensional configuration of X without using an explicit projection map. More precisely, on input matrix $X \in \mathbb{R}^{n\times m}$, MDS works with the distance matrix $D = (d_{ij})_{i,j=1,\cdots,m}$, of the points in X to compute an optimal configuration of points $Y = (y_1, \cdots, y_m) \in \mathbb{R}^{p\times m}$, with $p \leq n$, minimizing the error $\operatorname{err}(Y, D) = \sum_{i,j=1}^m (d_{ij} - \|y_i - y_j\|)^2$. In other words, the low dimensional configuration of points Y preserves the distances of the higher dimensional dataset X approximately.

In the construction of *nonlinear* dimensionality reduction (NDR) methods, we are especially interested in their interaction with signal processing tools, e.g., convolution transforms. When applying signal transforms to the dataset X, one important task is the analysis of the incurred geometrical deformation. To this end,

we propose the concept of *modulation maps* and *modulation manifolds* for the construction of particular datasets which are relevant in signal processing and NDR, especially since we are interested in numerical methods for analyzing geometrical properties of the modulation manifolds, with a particular focus on their scalar and mean curvature.

We define a modulation manifold by employing a homeomorphism (or diffeomorphism) $\mathscr{A} : \Omega \to \mathscr{M}$, for a specific manifold Ω, as used in signal processing. The basic objective is to understand how the geometry of Ω is distorted when we transform Ω using a modulation map $\mathscr{A}$. More explicitly, let $\{\phi_k\}_{k=1}^d \subset \mathscr{H}$ be a set of vectors in an Euclidean space $\mathscr{H}$, and $\{s_k : \Omega \to \mathscr{C}_{\mathscr{H}}(\mathscr{H})\}_{k=1}^d$ a family of smooth maps from a manifold Ω to $\mathscr{C}_{\mathscr{H}}(\mathscr{H})$ (the continuous functions from $\mathscr{H}$ into $\mathscr{H}$). We say that a manifold $\mathscr{M} \subset \mathscr{H}$ is a $\{\phi_k\}_{k=1}^d$-*modulated manifold* if

$$\mathscr{M} = \left\{ \sum_{k=1}^{d} s_k(\alpha)\phi_k,\ \alpha \in \Omega \right\}.$$

In this case, the map $\mathscr{A} : \Omega \to \mathscr{M}, \alpha \mapsto \sum_{k=1}^d s_k(\alpha)\phi_k$, is called *modulation map*.

To make one prototypical example (cf. [7]), we regard a map of the form

$$\mathscr{A}(\alpha)(t_i) = \sum_{k=1}^{d} \phi_k(\alpha_k t_i), \quad \alpha = (\alpha_1, \ldots, \alpha_d) \in \Omega, \quad \{t_i\}_{i=1}^n \subset [0,1],$$

for a set of band-limited functions $\{\phi_k\}_{k=1}^d$ in combination with a finite set of uniform samples $\{t_i\}_{i=1}^n \subset [0,1]$.

Now we use the same notation for the band-limited functions ϕ_k and the above mentioned vector of sampling values $\{\phi_k(t_i)\}_{i=1}^n$, as this is justified by the *Whittaker-Shannon interpolation formula* as follows.

As the support of the band-limited functions ϕ_k is located in $[0,1]$, the Whittaker-Shannon interpolation formula allows us to reconstruct each ϕ_k exactly from the finite samples $(\phi_k(t_i))_{i=1}^n \in \mathbb{R}^n$. This in turn gives a one-to-one relation between the band-limited functions $\phi_k : [0,1] \to \mathbb{R}$ and the vectors $(\phi_k(t_i))_{i=1}^n \in \mathbb{R}^n$. Note that the maps $s_k(\alpha)$ are in our example given by $s_k(\alpha)\phi_k(t_i) = \phi_k(\alpha_k t_i)$. In other words, we use the (continuous) map $s_k(\alpha)$, $f(t) \mapsto f(\alpha_k t)$, as the scaling by factor α_k, being the k-th coordinate of vector $\alpha \in \Omega \subset \mathbb{R}^d$.

To explain our analysis of the geometric distortions incurred by $\mathscr{A}$, we restrict ourselves to the case $d = 3$ and $\Omega \subset \mathbb{R}^3$ with $\dim(\Omega) = 2$. We compute the scalar curvature of $\mathscr{M}$ from the parametrization of Ω and the modulation map $\mathscr{A}$ by the following algorithm [7].

Algorithm 11

On input parametrization $\alpha = (\alpha_j(\theta_1, \theta_2))_{j=1}^d$ of Ω and band-limited functions $\{\phi_j\}_{j=1}^d$ that are generating the map $\mathscr{A}$, perform the following steps.

(1) Compute the Jacobian matrices J_α;

(2) Compute the metric tensor $g_{ij} = \sum_{\ell=1}^{n} t_\ell^2 \sum_{r,q=1}^{d} \left(\frac{d\phi_r}{dt}(\alpha_r t_\ell) \frac{d\phi_q}{dt}(\alpha_q t_\ell) \frac{\partial \alpha_r}{\partial \theta_i} \frac{\partial \alpha_q}{\partial \theta_j} \right)$;

(3) Compute the Christoffel symbols $\Gamma_{ij}^k = \frac{1}{2} \sum_{\ell=1}^{p} \left(\frac{\partial g_{j\ell}}{\partial x_i} + \frac{\partial g_{i\ell}}{\partial x_j} - \frac{\partial g_{ij}}{\partial x_\ell} \right) g^{\ell k}$;

(4) Compute the tensors $R^\ell{}_{ijk} = \sum_{h=1}^{p} (\Gamma_{jk}^h \Gamma_{ih}^\ell - \Gamma_{ik}^h \Gamma_{jh}^\ell) + \frac{\partial \Gamma_{jk}^\ell}{\partial x_i} - \frac{\partial \Gamma_{ik}^\ell}{\partial x_j}$;

(5) Compute the scalar curvature $S = \sum_{i,j=1}^{p} g^{ij} R_{ij}$, where $R_{ij} = \sum_{k,\ell=1}^{p} g^{k\ell} R_{kij}^k$.

Output: The scalar curvature S of $\mathscr{M} = \mathscr{A}(\Omega)$.

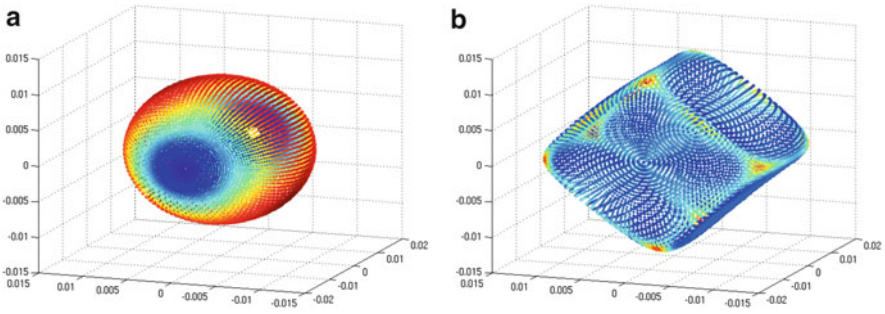


Fig. 14.4 (**a**) A sphere Ω whose colors represent the scalar curvature of $\mathscr{M} = \mathscr{A}(\Omega)$, (**b**) PCA projection of $\mathscr{M} = \mathscr{A}(\Omega)$ with Gaussian curvature represented by colors

For further details concerning the construction of Algorithm 11, we refer to [7].

14.4 Audio Signal Separation and Signal Detection

In many relevant applications of signal processing there is an increasing demand for effective methods to estimate the components from a mixture of acoustic signals. In recent years, different decomposition techniques were developed to do so, including *independent subspace analysis* (ISA), based on *independent component analysis* (ICA), see [1, 5, 26], and *non-negative matrix factorization* (NNMF), see [6, 23, 27]. The computational complexity of these methods, however, may be very large, in particular for real-time computations on audio signals. In signal separation, dimensionality reduction methods are used to first reduce the

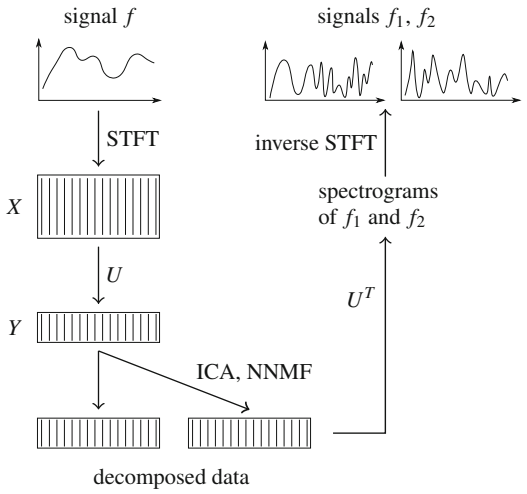


Fig. 14.5 Signal separation with dimensionality reduction

dimension of the data obtained from a time-frequency transform, e.g., *short time Fourier transform* (STFT), before the reduced data is decomposed into different components, each assigned to one of the source signals. For the application of dimensionality reduction in combination with NNMF, however, *non-negative* dimensionality reduction methods are essentially required to guarantee non-negative output data from non-negative input data (e.g., a non-negative spectrogram from the STFT). For the special case of PCA, a suitable rotation map is constructed in [12] for the purpose of back-projecting the reduced data to the positive orthant of the Cartesian coordinate system, where the sought rotation is given by the solution of a constraint optimization problem in a linear subspace of orthogonal matrices.

In this section, we evaluate different decomposition methods for signal separation in combination with the non-negative PCA projection from [12]. The basic steps of our method are illustrated in Fig. 14.5.

To explain how we use PCA, let $U \in \mathbb{R}^{D\times d}$ be an orthogonal projection, satisfying $Y = U^T X$, being obtained by the solution of the minimization problem

$$\min_{\tilde{U}^T\tilde{U}=I} \sum_{k=1}^{n} \left\| x_k - \tilde{U}\tilde{U}^T x_k \right\|_2 . \tag{14.7}$$

The solution of (14.7) is given by the maximizer of the variance $\mathrm{var}(Y)$ of Y, as given by the trace of YY^T. This observation allows us to reformulate the minimization problem in (14.7) as an equivalent maximization problem,

$$\max_{\tilde{U}^T\tilde{U}=I} \mathrm{tr}(\tilde{U}^T XX^T \tilde{U}), \tag{14.8}$$

where the maximizer U of $\mathrm{var}(Y)$ is given by a matrix U whose d columns contain the eigenvectors of the d largest eigenvalues of the covariance matrix XX^T.

For further processing the data in a subsequent decomposition by NNMF, the data matrix Y is essentially required to be *non-negative*. Note, however, that even if the data matrix X (obtained e.g., by STFT) may be non-negative, this is not necessarily the case for the components of the reduced data matrix Y. Therefore, we reformulate the maximization problem in (14.8) by adding a non-negativity constraint:

$$\max_{\substack{\tilde{U}^T\tilde{U}=I \\ \tilde{U}^T X\geq 0}} \mathrm{tr}(\tilde{U}^T XX^T\tilde{U}). \tag{14.9}$$

Note that this additional restriction transforms the simple PCA problem (14.8) into a much more difficult non-convex optimization problem (14.9) with many local solutions, for which (in general) none of the solutions is known analytically.

We tackle this fundamental problem as follows. We make use of the fact that the input data set X is *non-negative*, before it is projected onto a *linear* subspace, with the perception that there exists a rotation of the low-dimensional data set Y into the non-negative orthant. Indeed, as proven in [12], such a rotation map exists, which motivates us to split the *non-negative PCA* (NNPCA) problem (14.9) into a PCA part and a rotation part. This, in turn, gives rise to seek for a general construction of a rotation matrix W satisfying $WU^TX \geq 0$.

To further explain our splitting approach, recall that we already know the solution U of the PCA part. Since the rotation matrix W is orthogonal, it does not affect the value of the NNPCA cost functional. Now, in order to determine the rotation matrix W, we consider solving an auxiliary optimization problem on the set of orthogonal matrices $SO(d)$, i.e., we minimize the cost functional

$$J(\tilde{W}) = \frac{1}{2}\sum_{i,j}\left[\left(\tilde{W}U^TX\right)_-\right]_{ij}^2 \qquad \text{where } [Z_-]_{ij} = \begin{cases} z_{ij} & \text{if } z_{ij} < 0, \\ 0 & \text{otherwise,} \end{cases} \tag{14.10}$$

as this was proposed in [21] in the context of ICA. However, we cannot solve this optimization problem directly by an additive update algorithm, since the set of rotation matrices $SO(d)$ is not invariant under additions. But an elegant way to minimize the cost functional J in (14.10) uses the Lie-group structure of $SO(d)$ to transfer the problem into an optimization problem on the Lie-algebra of skew-symmetric matrices $\mathfrak{so}(d)$. Due to the vector space property of $\mathfrak{so}(d)$, standard methods can be applied to find the minimum (see [9, 11, 21] for details).

14.4.1 Decomposition Techniques

There are different methods for the decomposition of the (reduced) spectrogram Y. Among them, independent component analysis (ICA) and non-negative matrix

factorization (NNMF) are commonly used. In either case, for the application of ICA or NNMF, we assume the input data Y to be a linear mixture of source terms s_i, i.e.,

$$Y = AS, \tag{14.11}$$

where $A \in \mathbb{R}^{d\times r}$ and $S \in \mathbb{R}^{r\times n}$ are unknown. For the estimation of A and S we need specific additional assumptions to balance the disproportion of equations and unknowns in the factorization problem (14.11).

14.4.1.1 Independent Component Analysis (ICA)

The basic assumption of ICA is that the source signals are statistically independent. Furthermore, the data matrix Y is assumed to result from n realizations of a d-dimensional random vector. In order to estimate S, a random variable $\mathscr{S}$ is constructed, whose n realizations yield the columns of the source matrix S. The components of $\mathscr{S}$ are chosen to be as stochastically independent as possible, where the stochastical independence can be measured by the *Kullback-Leibler distance* [4].

In practice, the number of sources is usually unknown. Therefore, we may detect more independent components than the true number of sources. In this case, two or more of the separated components belong to the same source. Thus, the sources are combinations of the independent components. In a subsequent step, the sources are grouped into independent subspaces, each corresponding to one source. Finally, the sources are reconstructed from these multi-component subspaces [1]. This procedure is called *independent subspace analysis* (ISA). The main difficulty of ISA is to identify components belonging to the same multi-component subspace.

14.4.1.2 Non-negative Matrix Factorization (NNMF)

The factorization of the given data Y into a mixing matrix A and the source signals (source components) S, i.e., $Y = AS$, could be done by matrix factorization. The data we use for signal separation are obtained by taking the modulus of the signal's STFT, and so the input data is non-negative. Since the source components are assumed to be spectrograms, too, we assume them to be non-negative as well. Therefore, non-negative matrix factorizations (NNMF) are suitable tools for decomposition.

There are different NNMF algorithms available, all of which are relying on the non-negativity $Y, A, S \geq 0$, where different measures $d(Y, AS)$ for the reconstruction error were proposed [6, 23, 27]. We consider using the generalized *Kullback-Leibler distance* (proposed in [13] and used for decomposing signal data in [27]):

$$d(Y, AS) = \sum_{i,j} Y_{ij} \log \frac{Y_{ij}}{(AS)_{ij}} - Y_{ij} + (AS)_{ij}.$$

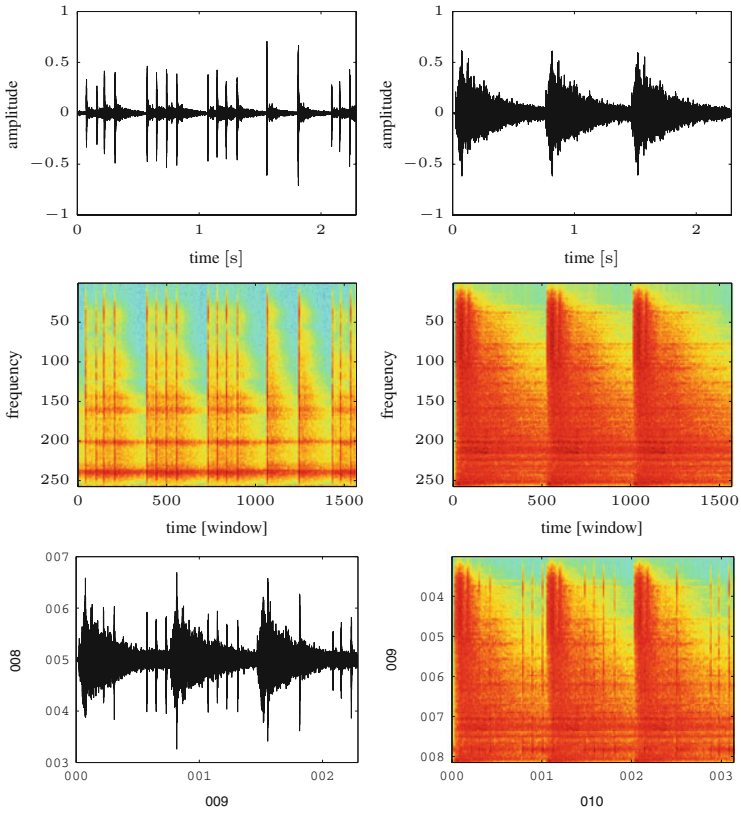


Fig. 14.6 Two acoustic signals: castanets f_1 (*top left*), cymbal f_2 (*top right*), and corresponding spectrograms (*second row*). Signal $f = f_1 + f_2$ and spectrogram (*third row*)

14.4.2 Numerical Results

We present one numerical example comparing the decomposition strategies ICA and NNMF. We consider a mixture $f = f_1 + f_2$ of acoustic transient signals, where f_1 is a sequence of castanets and f_2 a cymbal signal, shown in Fig. 14.6, where also the combination $f = f_1 + f_2$ of the two signals is displayed. The spectrograms in these figures are generated with an STFT using a Hamm-window. Since f_2 is a high-energy signal, f has a complex frequency characteristic. Therefore, the extraction of the castanets signal f_1, being active only at a few time steps, is a challenging task.

The obtained separations, resulting from the two different decomposition methods using NNPCA and PCA, respectively, are displayed in Fig. 14.7. Note that both methods, NNMF and ICA, achieve to reproduce the characteristic peaks of the castanets quite well. However, in the case of NNMF strong artifacts of the castanets are visible in the cymbal signal, whereas the separation by ICA is almost perfect.

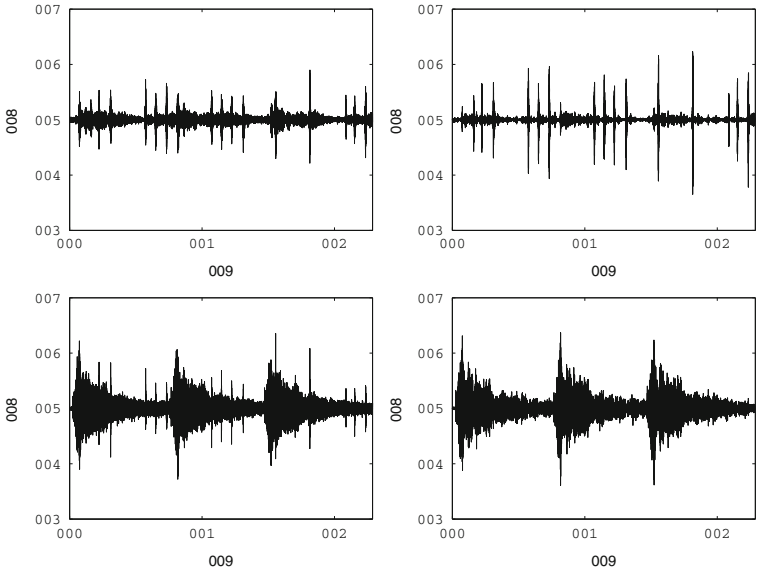


Fig. 14.7 Signal separation by NNPCA & NNMF (*left column*); PCA & ICA (*right column*)

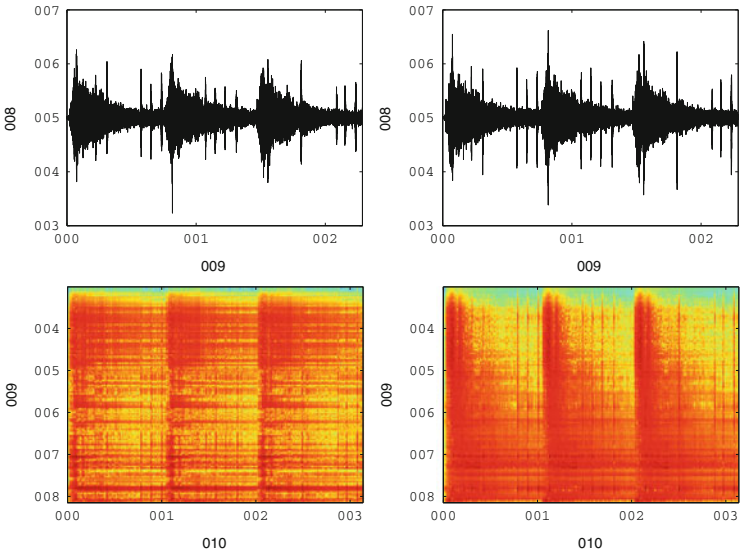


Fig. 14.8 Reconstruction of f as sum of the decomposed f_i by using NNPCA & NNMF (*left column*) and by using PCA & ICA (*right column*)

Likewise, for the reconstruction of the reduced signal, the combination of PCA and ICA provides an almost complete reproduction of the original signal f (see Fig. 14.8). Merely at time steps where a high amplitude of the cymbal exactly

matches the peaks of the castanets, a correct separation is not quite achieved. As for the NNMF, the spectrogram in Fig. 14.8 shows that information is being lost.

We finally remark that for signal separation *without* dimensionality reduction, NNMF is competitive to ICA (see e.g. [27]). This indicates that our use of NNPCA in combination with NNMF could be improved. Further improvements could be achieved by the use of more sophisticated (nonlinear) dimensionality reduction methods. On the other hand, this would lead to a much more complicated construction of the inverse transform, as required for the back-projection of the data. We defer these points to future research. Nevertheless, although PCA is only a *linear* projection method, our numerical results of this section, especially those obtained by the combination of PCA and ICA, are already quite promising.

References

1. Casey, M., Westner, A.: Separation of mixed audio sources by independent subspace analysis. In: Proceedings of the International Computer Music Conference, San Francisco, pp. 154–161 (2000)
2. Cohen, A., Daubechies, I., Feauveau, J.C.: Biorthogonal bases of compactly supported wavelets. Commun. Pure Appl. Math. **45**(5), 485–560 (1992)
3. Coifman, R., Donoho, D.: Translation-invariant de-noising. In: Wavelets and Statistics, pp. 125–150. Springer, New York (1995)
4. Comon, P.: Independent component analysis, a new concept? Signal Process. **36**(3), 287–314 (1994)
5. FitzGerald, D., Coyle, E., Lawlor, B.: Sub-band independent subspace analysis for drum transcription. In: Proceedings of the 5th International Conference on Digital Audio Effects (DAFX'02), Hamburg, pp. 65–69 (2002)
6. FitzGerald, D., Cranitch, M., Coyle, E.: Non-negative tensor factorisation for sound source separation. In: Proceedings of Irish Signals and Systems Conference, Dublin, pp. 8–12 (2005)
7. Guillemard, M., Iske, A.: Curvature analysis of frequency modulated manifolds in dimensionality reduction. Calcolo **48**(1), 107–125 (2011)
8. Guillemard, M.: Some Geometrical and Topological Aspects of Dimensionality Reduction in Signal Analysis. PhD thesis, University of Hamburg, 2011, ftp://ftp.math.tu-berlin.de/pub/numerik/guillem/prj2/mgyDiss.pdf.
9. Hall, B.C.: Lie Groups, Lie Algebras, and Representations. An Elementary Introduction. Springer, New York (2003)
10. Heinen, D., Plonka, G.: Wavelet shrinkage on paths for denoising of scattered data. Result. Math. **62**(3–4), 337–354 (2012)
11. Iserles, A., Munthe-Kaas, H., Nørsett, S., Zanna, A.: Lie-group methods. Acta Numer. **9**, 215–365 (2000)
12. Krause-Solberg, S., Iske, A.: Non-negative dimensionality reduction in signal separation. University of Hamburg (2014, preprint)
13. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems, vol. 13, pp. 556–562. MIT, Cambridge, Massachusetts (U.S.A.) (2000)
14. Lee, J.A., Verleysen, M.: Nonlinear Dimensionality Reduction. Springer, New York (2007)
15. Mallat, S.: A Wavelet Tour of Signal Processing. Academic, San Diego (1999)
16. Plonka, G.: The easy path wavelet transform: a new adaptive wavelet transform for sparse representation of two-dimensional data. Multiscale Model. Simul. **7**(3), 1474–1496 (2009)

17. Plonka, G., Iske, A., Tenorth, S.: Optimal representation of piecewise Hölder smooth bivariate functions by the easy path wavelet transform. J. Approx. Theory **176**, 42–67 (2013)
18. Plonka, G., Roşca, D.: Easy path wavelet transform on triangulations of the sphere. Math. Geosci. **42**(7), 839–855 (2010)
19. Plonka, G., Tenorth, S., Iske, A.: Optimally sparse image representation by the easy path wavelet transform. Int. J. Wavelets Multiresolut. Inf. Process. **10**(1), 1250,007, 20 (2012)
20. Plonka, G., Tenorth, S., Roşca, D.: A hybrid method for image approximation using the easy path wavelet transform. IEEE Trans. Image Process. **20**(2), 372–381 (2011)
21. Plumbley, M.D.: Geometrical methods for non-negative ICA: manifolds, Lie groups and toral subalgebras. Neurocomputing **67**, 161–197 (2005)
22. Ram, I., Elad, M., Cohen, I.: Generalized tree-based wavelet transform. IEEE Trans. Signal Process. **59**(9), 4199–4209 (2011)
23. Smaragdis, P., Brown, J.: Non-negative matrix factorization for polyphonic music transcription. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, pp. 177–180 (2003)
24. Tenorth, S.: Adaptive waveletmethoden zur approximation von bildern. Phd thesis, University of Göttingen (2011). http://hdl.handle.net/11858/00-1735-0000-0006-B3E7-A
25. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of 6th International Conference on Computer Vision (ICCV '98), Bombay, pp. 839–846. IEEE Computer Society (1998)
26. Uhle, C., Dittmar, C., Sporer, T.: Extraction of drum tracks from polyphonic music using independent subspace analysis. In: Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, pp. 843–848 (2003)
27. Virtanen, T.: Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. Trans. Audio Speech Lang. Proc. **15**(3), 1066–1074 (2007)
28. Welk, M., Weickert, J., Steidl, G.: A four-pixel scheme for singular differential equations. In: Scale-Space and PDE Methods in Computer Vision, pp. 610–621. Springer, Berlin (2005)

Chapter 15
Error Bound for Hybrid Models of Two-Scaled Stochastic Reaction Systems

Tobias Jahnke and Vikram Sunkara

Abstract Biochemical reaction systems are often modeled by a Markov jump process in order to account for the discreteness of the populations and the stochastic nature of their evolution. The associated time-dependent probability distribution is the solution of the Chemical Master Equation (CME), but solving the CME numerically is a considerable challenge due to the high dimension of the state space. In many applications, however, species with rather small population numbers interact with abundant species, and only the former group exhibits stochastic behavior. This has motivated the derivation of hybrid models where a low-dimensional CME is coupled to a set of ordinary differential equations representing the abundant species. Using such a hybrid model decreases the number of unknowns significantly but – in addition to the numerical error – causes a modeling error. We investigate the accuracy of the MRCE (= model reduction based on conditional expectations) approach with respect to a particular scaling of the reaction system and prove that the error is proportional to the scaling parameter.

15.1 Introduction

Biological systems such as gene-regulatory networks and cell metabolic processes consist of multiple species which are undergoing transformations via a set of reaction channels. If all populations are sufficiently large, then the evolution of the concentrations over time can be modeled by the classical reaction-rate equation, i.e. a system of ordinary differential equations; cf. [23]. In many applications, however, some of the species occur in low amounts, and it was observed that small stochastic fluctuations in their populations can cascade large effects to the other species. Important examples are gene-regulatory networks where the evolution of the entire system depends crucially on the stochastic behavior of a rather small number of transcription factors. In order to capture these effects, such systems must

T. Jahnke (✉) • V. Sunkara
Karlsruhe Institute of Technology, Kaiserstr. 89-93, 76133 Karlsruhe, Germany
e-mail: tobias.jahnke@kit.edu; vikram.sunkara@kit.edu

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_15

be described by a Markov jump processes, which respects the inherent discrete nature of the system and its stochastic interactions.

The associated time-dependent probability distribution is the solution of the *Chemical Master Equation (CME)*, but solving the CME is a considerable challenge, as the size of the state space scales exponentially in the number of species (*curse of dimensionality*). For this reason, Monte Carlo approaches based on the stochastic simulation algorithm from [6] or related methods are often used. In an alternative line of research, numerical techniques have been applied to the CME in order to reduce the number of degrees of freedom, e.g. optimal state space truncation [1,26,27], spectral approximation [4], adaptive wavelet compression [16,20], sparse grids [11], or tensor product approximation [2, 10, 18, 21, 22] among others. But in spite of the progress achieved with these approaches, many biological systems are still out of reach of direct numerical approximation.

The size of the problem can be significantly reduced if only species with low populations are described by a probability distribution, whereas the abundant species are represented by (conditional) moments. This approach is motivated by the famous result in [23] which states, roughly speaking, that stochastic fluctuations in large populations are insignificant. In the last years, this has inspired the development of *hybrid models* where a low-dimensional CME is coupled to ordinary differential equations similar to the classical reaction-rate equation; cf. [5, 9, 11–13, 17, 25, 28].

In this article, we analyze the accuracy of a hybrid model called MRCE (*model reduction based on conditional expectations*). This approach has been proposed in [9, 17, 25], and it was demonstrated numerically that MRCE captures the critical bimodal solution profiles which appear in certain applications. In [9,17,28], numerical techniques for MRCE were introduced, and an error bound for the modeling error was proven in [28]. In the present article, we make the additional assumption that the reaction system involves two scales, i.e. that the ratio between the small and large populations is proportional to a scaling parameter $0 < \varepsilon \ll 1$. For such two-scaled systems, we prove that the modeling error of the MRCE approximation is proportional to ε. The proof blends ideas and techniques from [19] and [28].

15.2 The Chemical Master Equation of Two-scale Reaction Systems

We consider a partitioned reaction system with two groups of species denoted by $\mathsf{S}_1, \ldots, \mathsf{S}_d$ and $\mathsf{S}_{d+1}, \ldots, \mathsf{S}_{d+D}$, respectively, with $d, D \in \mathbb{N}$. Let $X(t) \in \mathbb{N}_0^d$ be the vector whose entries $X_1(t), \ldots X_d(t)$ indicate how many copies of each of the species $\mathsf{S}_1, \ldots, \mathsf{S}_d$ exist at time $t \in [0, t_{\text{end}}]$, and let $Y(t) = (Y_1(t), \ldots, Y_D(t))$ contain the copy numbers of $\mathsf{S}_{d+1}, \ldots, \mathsf{S}_{d+D}$. The species interact via $r \in \mathbb{N}$ reaction channels, $\mathsf{R}_1, \ldots, \mathsf{R}_r$, each of which is represented by a scheme

$$\mathsf{R}_j : \sum_{k=1}^{d} a_{jk}\mathsf{S}_k + \sum_{k=1}^{D} b_{jk}\mathsf{S}_{d+k} \xrightarrow{c_j} \sum_{k=1}^{d} \hat{a}_{jk}\mathsf{S}_k + \sum_{k=1}^{D} \hat{b}_{jk}\mathsf{S}_{d+k}, \tag{15.1}$$

with $a_{jk}, \hat{a}_{jk}, b_{jk}, \hat{b}_{jk} \in \mathbb{N}_0$ and $c_j > 0$. If the j-th reaction channel fires, then the population numbers jump from the current state $(X(t), Y(t)) = (n, m) \in \mathbb{N}_0^d \times \mathbb{N}_0^D$ to the new state $(n, m) + (\nu_j, \mu_j)$, where $(\nu_j, \mu_j) \in \mathbb{Z}^{d+D}$ is the stoichiometric vector associated to R_j, i.e.

$$\nu_j = \left(\hat{a}_{j1} - a_{j1}, \ldots, \hat{a}_{jd} - a_{jd}\right)^T \in \mathbb{Z}^d$$

$$\mu_j = \left(\hat{b}_{j1} - b_{j1}, \ldots, \hat{b}_{jd} - b_{jd}\right)^T \in \mathbb{Z}^D.$$

In stochastic reaction kinetics, the function $t \mapsto (X(t), Y(t))$ is a realization of a Markov jump process; cf. [6, 14]. According to [6] the transition rates of this process depend on the propensity functions of the reaction channels. We assume that the propensity function of R_j has the form $\alpha_j(n)\beta_j(m)$ with

$$\alpha_j(n) = c_j \prod_{k=1}^{d} \binom{n_k}{a_{jk}}, \qquad \beta_j(m) = \varepsilon^{\gamma(j)-1}\varepsilon^{|b_j|} \prod_{k=1}^{D} \binom{m_k}{b_{jk}}, \tag{15.2}$$

where $|b_j| = \sum_{i=1}^{d} b_{ji}$, and where $0 < \varepsilon \ll 1$ is a scaling parameter discussed below. The value of γ depends on whether or not the population numbers of the first group of species change when R_j fires. To be more precise, we partition the index set $\{1, \ldots, r\}$ into

$$J_0 = \left\{ j \in \{1, \ldots, r\} : \nu_j = (0, \ldots, 0)^T \right\}, \qquad J_1 = \{1, \ldots, r\} \setminus J_0$$

and let γ be the indicator function

$$\gamma(j) = \begin{cases} 0 \text{ if } j \in J_0, \\ 1 \text{ if } j \in J_1. \end{cases} \tag{15.3}$$

The reason for this particular scaling is the following: if $(X(t), Y(t)) = (n, m) \in \mathbb{N}_0^d \times \mathbb{N}_0^D$ with $n \in O(1)$ and $m \in O\left(\varepsilon^{-1}\right)$, then $\alpha_j(n)\beta_j(m) = O(c_j \varepsilon^{\gamma(j)-1})$ for all $j = 1, \ldots, r$. Hence, the population numbers of $\mathsf{S}_{d+1}, \ldots, \mathsf{S}_{d+D}$ may change with a rate of $O\left(\varepsilon^{-1}\right)$, whereas the populations of $\mathsf{S}_1, \ldots, \mathsf{S}_d$ only change with a rate of $O(1)$, provided that $c_j = O(1)$ for all j. For initial data $X(0) = O(1)$ and $Y(0) = O\left(\varepsilon^{-1}\right)$, one can thus expect that $\mathbb{E}(X(t)) = O(1)$ and $\mathbb{E}(Y(t)) = O\left(\varepsilon^{-1}\right)$ on bounded time intervals. Hence, ε is roughly speaking the ratio between the small and the large population numbers of the two groups $\mathsf{S}_1, \ldots, \mathsf{S}_d$ and $\mathsf{S}_{d+1}, \ldots, \mathsf{S}_{d+D}$, respectively. This scaling was extensively motivated and illustrated in [19], and

a very similar scaling was considered in [25]. For $d = 0$ and $\alpha_j(n) = c_j$, our scaling coincides with the thermodynamic limit which has been analyzed in [23]. For $\varepsilon = 1$, there is no qualitative difference between the two groups of species, such that our setting corresponds to the situation in [6] where no partition nor scaling is considered.

Let $p(t,n,m)$ be the probability that at time $t \geq 0$ the process is in state $(n,m) \in \mathbb{N}_0^d \times \mathbb{N}_0^D$, i.e. the probability that there are n_k copies of S_k for $k = 1,\ldots,d$, and m_l copies of S_{d+l} for $l = 1,\ldots,D$. It is well-known (see [6, 7]) that the probability distribution p evolves according to the Chemical Master Equation (CME)

$$\partial_t p(t,n,m) = \sum_{j=1}^{r} \Big(\alpha_j(n-\nu_j)\beta_j(m-\mu_j)p(t,n-\nu_j,m-\mu_j) - \alpha_j(n)\beta_j(m)p(t,n,m) \Big) \quad \forall (n,m) \in \mathbb{N}_0^d \times \mathbb{N}_0^D \tag{15.4}$$

$$p(0,n,m) = p_0(n,m) \tag{15.5}$$

with the convention that $p(t,n-\nu_j,m-\mu_j) = 0$ if $n-\nu_j \notin \mathbb{N}_0^d$ or $m-\mu_j \notin \mathbb{N}_0^D$. For the sake of a more compact notation we define the shift operators S_j^1 and S_j^2 by

$$S_j^1 u(n,m) = \begin{cases} u(n-\nu_j,m) & \text{if } n-\nu_j \in \mathbb{N}_0^d \\ 0 & \text{else} \end{cases}$$

$$S_j^2 u(n,m) = \begin{cases} u(n,m-\mu_j) & \text{if } m-\mu_j \in \mathbb{N}_0^D \\ 0 & \text{else} \end{cases}$$

for $u : \mathbb{N}_0^d \times \mathbb{N}_0^D \longrightarrow \mathbb{R}$. The two shift operators commute, i.e.

$$S_j^1 S_j^2 u(n,m) = S_j^2 S_j^1 u(n,m) = \begin{cases} u(n-\nu_j,m-\mu_j) & \text{if } n-\nu_j \in \mathbb{N}_0^d, m-\mu_j \in \mathbb{N}_0^D \\ 0 & \text{else.} \end{cases}$$

Products of functions are to be understood entry-wise, and applying a shift operator to a product $u(n,m)v(n,m)$ is to be understood in the sense that

$$\left(S_j^1 uv\right)(n,m) = \left(S_j^1(uv)\right)(n,m) = u(n-\nu_j,m)v(n-\nu_j,m) = \left(S_j^1 u\right)\left(S_j^1 v\right)(n,m).$$

With these operators, the CME (15.4) can be reformulated as

$$\partial_t p = \sum_{j=1}^{r} (S_j^1 S_j^2 - I)\left(\alpha_j \beta_j p\right). \tag{15.6}$$

The chemical master equation (15.6) is considered on the space

$$\ell^1 = \Big\{u : \mathbb{N}_0^d \times \mathbb{N}_0^D \longrightarrow \mathbb{R} \ : \ \sum_{n\in\mathbb{N}_0^d} \sum_{m\in\mathbb{N}_0^D} |u(n,m)| < \infty\Big\}.$$

of absolutely summable functions on $\mathbb{N}_0^d \times \mathbb{N}_0^D$. This is a straightforward extension of the standard ℓ^1-space. For vector-valued functions $u = (u_1, \ldots, u_N) : \mathbb{N}_0^d \times \mathbb{N}_0^D \longrightarrow \mathbb{R}^N$ with some $N > 1$, $u \in \ell^1$ means that $u_j \in \ell^1$ for all $j = 1, \ldots, N$. The space ℓ^1 is endowed with the norm

$$\|u\|_{\ell^1} = \sum_{n\in\mathbb{N}_0^d} \sum_{m\in\mathbb{N}_0^D} |u(n,m)|$$

where $|\cdot| = |\cdot|_1$ is the 1-norm on $\mathbb{R}^N$. We set $\mathscr{X}^0 = \ell^1$ and define the spaces $\mathscr{X}^i$ via the recursion

$$\mathscr{X}^{i+1} = \Big\{u \in \mathscr{X}^i \mid (n,m) \mapsto m_k u(n,m) \in \mathscr{X}^i \text{ for all } k \in \{1, \ldots, D\}\Big\}.$$

If $p(t,\cdot,\cdot) \in \ell^1$ is the solution of the CME (15.6), then $p_1(t,n) = \sum_m p(t,n,m)$ is the marginal distribution of $p(t,\cdot,\cdot)$, and if $p_1(t,n) \neq 0$, then

$$p_2(t, m \mid n) = \frac{p(t,n,m)}{p_1(t,n)} \tag{15.7}$$

is the conditional probability that at time t there are m_j particles of S_j given there are n_i particles of S_i ($i \in \{1, \ldots, d\}$, $j \in \{d+1, \ldots, d+D\}$). If $p(t,\cdot,\cdot) \in \mathscr{X}^2$, then the conditional central moments

$$\xi(t,n) = \sum_{m\in\mathbb{N}_0^D} m p_2(t, m \mid n) \tag{15.8}$$

$$C_\xi(t,n) = \sum_{m\in\mathbb{N}_0^D} \big(m - \xi(t,n)\big)\big(m - \xi(t,n)\big)^T p_2(t, m \mid n)$$

exist provided that $p_1(t,n) \neq 0$.

15.3 Model Reduction Based on Conditional Expectations

Solving the CME (15.4) or (15.6) numerically is a considerable challenge. First, the infinite state space $\mathbb{N}_0^d \times \mathbb{N}_0^D$ has to be truncated; this causes an error which has been analyzed in [26]. The truncated state space is finite, but still $(d+D)$-dimensional, and the total number of states is usually so large that standard numerical schemes cannot be applied.

On the other hand, the solution of the CME often provides more information than actually needed to understand the biological process. In many applications, one is mainly interested in the question how the stochastic behavior of $\mathsf{S}_1, \ldots, \mathsf{S}_d$ affects the dynamics of $\mathsf{S}_{d+1}, \ldots, \mathsf{S}_{d+D}$. If the population numbers of $\mathsf{S}_{d+1}, \ldots, \mathsf{S}_{d+D}$ are sufficiently large, then stochastic fluctuations within their populations can be neglected according to [23]. In this case, it is sufficient to compute the *marginal distribution* $p_1(t,n)$ of the species $\mathsf{S}_1, \ldots, \mathsf{S}_d$ along with conditional moments which measure the abundance of $\mathsf{S}_{d+1}, \ldots, \mathsf{S}_{d+D}$. This has motivated the construction of *hybrid models*: Instead of trying to solve the high-dimensional CME and then extracting the relevant information from the solution $p(t,n,m)$, one derives a reduced set of equations, namely a low-dimensional CME for the marginal distribution coupled with other ODEs; cf. [5, 9, 11–13, 17, 25, 28]. Hybrid models have the advantage that the huge number of unknowns is significantly reduced, which makes the problem computationally feasible. The price to pay is that hybrid models involve structurally more complicated differential equations than the (linear) CME, and that such a model reduction causes a modeling error in addition to the numerical error. The following hybrid model has been derived in [17]:

$$\partial_t w = \sum_{j \in J_1} (S_j^1 - I)\Big(\alpha_j \beta_j(\phi) w\Big) =: A(\phi) w \tag{15.9}$$

$$\begin{aligned} \partial_t (\phi w) &= \sum_{j \in J_1} (S_j^1 - I)\Big(\alpha_j \beta_j(\phi) \phi w\Big) + \sum_{j=1}^{r} \mu_j S_j^1 \Big(\alpha_j \beta_j(\phi) w\Big) \\ &=: F(\phi, w) + G(\phi, w) \end{aligned} \tag{15.10}$$

For fixed ϕ, (15.9) is again a CME, but on the lower-dimensional state space $\mathbb{N}_0^d$. The function $w(t,n)$ approximates the marginal distribution $p_1(t,n)$ of the full CME solution, whereas $\phi(t,n)$ approximates the conditional expectations $\xi(t,n)$ defined in (15.8). This is why this model was called *model reduction based on conditional expectations (MRCE)* in [17]. It was demonstrated by numerical examples that MRCE captures certain bimodal solution profiles correctly, in contrast to simpler hybrid models proposed in the literature. Since w and ϕ do not depend on m any more, the $(d + D)$-dimensional state space of the CME is replaced by a d-dimensional state space, which reduces the computational costs considerably. Similar approaches have been proposed in [8, 9, 13, 15, 24, 25] for the CME and related differential equations.

Approximating the conditional expectations $\xi(t,n)$ has the drawback that (15.7) and hence (15.8) cannot be properly defined if $p_1(t,n) = 0$. The same applies to the approximations $w(t,n) \approx p_1(t,n)$ and $\phi(t,n) \approx \xi(t,n)$. The hybrid model (15.9)–(15.10) is formulated in terms of w and ϕw, but in order to evaluate the term $\beta_j(\phi)$ on the right-hand side, we have to divide $\phi(t,n)w(t,n)$ by $w(t,n)$. This is only possible if $w(t,n) > 0$, and for $w(t,n) \approx 0$ such a division causes numerical

instability. Different strategies to cope with this problem have been proposed in [9, 13,17,25,28]. Since the main goal of the present article is an analysis of the accuracy of MRCE, we will avoid such technical problems by the following assumption:

Assumption 15.1. *We assume that the CME (15.6) with initial condition (15.5) has a unique classical solution $p(t,\cdot,\cdot) \in \ell^1$ with strictly positive marginal distribution $p_1(t,\cdot)$, i.e. $p_1(t,n) > 0$ for all $t \in [0,t_{end}]$ and all $n \in \mathbb{N}_0^d$. This implies that p_2, ξ and C_ξ are well-defined. Moreover, we assume that the hybrid model (15.9)–(15.10) with initial data*

$$w(0,n) = p_1(0,n) \qquad \text{and} \qquad \phi(0,n) = \xi(0,n) \tag{15.11}$$

has a unique solution, and that $w(t,\cdot) \in \ell^1$ is strictly positive for all $t \in [0,t_{end}]$.

This assumption seems to be a strong simplification because in typical applications it can be observed that for every threshold parameter $\delta \in (0,1)$, there are only finitely many states with $p(t,n,m) \geq \delta$. Roughly speaking, this means that $p(t,n,m) \approx 0$ for "most of" the states. However, if $p(t_\star,n_\star,m_\star) = 0$ for some $t_\star > 0$, then the state $(n_\star,m_\star) \in \mathbb{N}_0^d \times \mathbb{N}_0^D$ cannot be reached from neither of the states which had nonzero probability at time $t = 0$. As a consequence, one could simply exclude $(n_\star,m_\star)$ from the state space to avoid the problem, and in this sense, Assumption 15.1 is not a severe restriction. Since numerical methods for solving (15.9)–(15.10) are not discussed in this article, numerical instabilities are not an issue here.

15.4 Error Analysis for the Hybrid Model

Since $\beta_j(m)$ defined in (15.2) depends on the scaling parameter ε and since $\beta_j(m)$ appears both in the CME (15.6) and in the hybrid model (15.9)–(15.10), the functions $p(t,n,m)$, $p_1(t,n)$, $\xi(t,n)$, $w(t,n)$, and $\phi(t,n)$ all depend[1] on ε, too. In this section we prove that the modeling error of MRCE is bounded by $C\varepsilon$ (see Theorem 15.1 below). Throughout the article, C denotes a generic constant which may have different values at different occurrences. The proof combines the arguments from [17, 19] with the analysis from [28] where systems with no scaling have been investigated. Our error analysis is based on the following assumptions.

Assumption 15.2. *For every $j \in \{1,\ldots,r\}$ we assume that $|b_j| \leq 2$.*

This is a natural assumption, because the probability of a trimolecular reaction is negligible according to [7, page 418].

[1] We do not make this dependency explicit in the notation in order to keep the equations as simple as possible.

Assumption 15.3. *We assume that the solution of the CME (15.6) satisfies* $p(t,\cdot,\cdot) \in \mathscr{X}^3$ *for* $t \in [0, t_{end}]$ *and that*

$$(n,m) \mapsto \alpha_j(n)\, p(t,n,m) \in \mathscr{X}^3 \qquad \text{for all } j \in \{1,\dots,r\}.$$

Assumption 15.4. *We assume that*

$$\sup_{t\in[0,t_{end}]} \sup_{n\in\mathbb{N}_0^d} |\xi(t,n)| \le \frac{C}{\varepsilon}, \qquad \sup_{t\in[0,t_{end}]} \sup_{n\in\mathbb{N}_0^d} |C_\xi(t,n)| \le \frac{C}{\varepsilon},$$

$$\sup_{t\in[0,t_{end}]} \sup_{n\in\mathbb{N}_0^d} |\phi(t,n)| \le \frac{C}{\varepsilon}$$

with a constant which does not depend on ε*. Moreover, we assume that all third central moments of* $p_2(t,\cdot \mid n)$ *are bounded by* $C\varepsilon^{-2}$ *with a constant which does not depend on* $t \in [0, t_{end}]$*,* ε*, and* $n \in \mathbb{N}_0^d$*.*

Assumption 15.5. *Suppose that there is a constant* $C > 0$ *such that for all* $t \in [0, t_{end}]$ *and* $j \in \{1,\dots,r\}$ *the bound*

$$\max_{j=1,\dots,r} \left\| \alpha_j(\cdot) u(t,\cdot) \right\|_{\ell^1} \le C \left\| u(t,\cdot) \right\|_{\ell^1}$$

holds for each of the following functions:

$$u = p_1, \qquad u = \beta_j(\xi)\, p_1 - \beta_j(\phi) w, \qquad u = \beta_j(\xi)\xi p_1 - \beta_j(\phi)\phi w.$$

Note that Assumption 15.4 implies $u \in \ell^1$ in each case.

The following error bound for the modeling error of MRCE is the main result of this article.

Theorem 15.1. *Under Assumptions 15.1–15.5, there is a constant* $C_b > 0$ *such that the approximation error of MRCE is bounded by*

$$\sup_{t\in[0,t_{end}]} \| p_1(t,\cdot) - w(t,\cdot) \|_{\ell^1} \le C_b \varepsilon \tag{15.12}$$

$$\sup_{t\in[0,t_{end}]} \| \xi(t,\cdot)\, p_1(t,\cdot) - \phi(t,\cdot) w(t,\cdot) \|_{\ell^1} \le C_b. \tag{15.13}$$

If in addition

$$|b_j| \le 1 \quad \text{for all } j \in J_0, \qquad |b_j| = 0 \quad \text{for all } j \in J_1, \tag{15.14}$$

then MRCE is even exact, i.e. we can choose $C_b = 0$ *in* (15.12) *and* (15.13)*.*

According to (15.13) the error of the approximation $\xi p_1 \approx \phi w$ remains bounded, but does not decrease when $\varepsilon \to 0$. This is not obvious, because Assumption 15.4

implies that $\|\xi(t,\cdot)p_1(t,\cdot)\|_{\ell^1} = O\left(\varepsilon^{-1}\right)$ and $\|\phi(t,\cdot)w(t,\cdot)\|_{\ell^1} = O\left(\varepsilon^{-1}\right)$. Multiplying both sides of (15.13) by ε shows that the *relative error* converges linearly in ε.

Proof. It will be shown below in Lemmas 15.2 and 15.3 that

$$\begin{aligned} &\|p_1(t,\cdot) - w_1(t,\cdot)\|_{\ell^1} + \varepsilon \|\xi p_1(t,\cdot) - \phi w(t,\cdot)\|_{\ell^1} \\ &\leq \tilde{C}_b t\varepsilon + C\int_0^t \varepsilon \|(\xi p_1 - \phi w)(s,\cdot)\|_{\ell^1}\, ds + C\int_0^t \|p_1(s,\cdot) - w(s,\cdot)\|_{\ell^1}\, ds. \end{aligned} \tag{15.15}$$

for all $t \in [0, t_{\text{end}}]$ with constants $\tilde{C}_b$ and C which do not depend on t or ε. Hence, the Gronwall lemma yields

$$\|p_1(t,\cdot) - w_1(t,\cdot)\|_{\ell^1} + \varepsilon \|\xi p_1(t,\cdot) - \phi w(t,\cdot)\|_{\ell^1} \leq C_b\varepsilon \text{ with } C_b = \tilde{C}_b t_{\text{end}} e^{Ct_{\text{end}}}$$

which proves (15.12) and (15.13). Moreover, it will be shown that we can choose $\tilde{C}_b = 0$ in the special case (15.14). □

The remainder of this article is devoted to the proof of the Gronwall inequality (15.15). As a preparatory step, we prove the following lemma:

Lemma 15.1. *Let* $y : \mathbb{N}_0^d \longrightarrow \mathbb{R}^d$, $z : \mathbb{N}_0^d \longrightarrow \mathbb{R}^d$ *with*

$$\max_{n\in\mathbb{N}_0^d} |y(n)| \leq C/\varepsilon, \qquad \max_{n\in\mathbb{N}_0^d} |z(n)| \leq C/\varepsilon, \tag{15.16}$$

and let $u \in \ell^1$ *and* $v \in \ell^1$. *Then for every* $j \in \{1,\ldots,r\}$, *there is a constant* $C > 0$ *such that*

$$\left\|\beta_j(y)u - \beta_j(z)v\right\|_{\ell^1} \leq C\varepsilon^{\gamma(j)}\left(\|yu - zv\|_{\ell^1} + \varepsilon^{-1}\|u - v\|_{\ell^1}\right)$$

with $\gamma(j)$ *defined in (15.3). Note that the assumption (15.16) implies that* $yu - zv \in \ell^1$.

A similar lemma has been shown in [19, Lemma 4].

Proof. For $|\lambda_j| = 0$ the assertion is obvious, because in this case $\beta_j(y) = \varepsilon^{\gamma(j)-1}$ is constant. If $|\lambda_j| = 1$, then there is a $k \in \{1,\ldots,d\}$ such that $\beta_j(y) = \varepsilon^{\gamma(j)}y_k$, and the assertion follows. If $|\lambda_j| = 2$, then the propensity $\beta_j(y)$ takes the form

$$\beta_j(y) = \hat{c}_j\varepsilon^{\gamma(j)+1}y_k y_l \qquad \text{with } \hat{c}_j = \begin{cases} c_j & \text{if } k \neq l \\ \frac{1}{2}c_j & \text{if } k = l \end{cases}$$

for some $k, l \in \{1,\ldots,d\}$. Thus, we have to bound the difference

$$\beta_j(y)u - \beta_j(z)v = \hat{c}_j \varepsilon^{\gamma(j)+1} (y_k y_l u - z_k z_l v)$$
$$= \hat{c}_j \varepsilon^{\gamma(j)+1} \Big(y_k (y_l u - z_l v) + y_k z_l (v-u) + z_l (y_k u - z_k v) \Big).$$

Since (15.16) implies that $|\varepsilon y_k(n)| \leq C$ and $|\varepsilon z_l(n)| \leq C$, it follows that

$$\big\| \beta_j(y)u - \beta_j(z)v \big\|_{\ell^1} \leq C \varepsilon^{\gamma(j)} \|yu - zv\|_{\ell^1} + C \varepsilon^{\gamma(j)-1} \|u - v\|_{\ell^1}$$

which proves the assertion. □

Lemma 15.2. *Under the assumptions of Theorem 15.1 there are constants $\tilde{C}_b \geq 0$ and $C > 0$ such that*

$$\|p_1(t,\cdot) - w_1(t,\cdot)\|_{\ell^1} \leq \tilde{C}_b t \varepsilon + C \int_0^t \varepsilon \, \|(\xi p_1 - \phi w)(s,\cdot)\|_{\ell^1} \, ds$$
$$+ C \int_0^t \|p_1(s,\cdot) - w(s,\cdot)\|_{\ell^1} \, ds.$$

for all $t \in [0, t_{end}]$. The constants $\tilde{C}_b$ and C do not depend on t or ε, and we can choose $\tilde{C}_b = 0$ in the special case (15.14).

Proof. From the definition of the marginal distribution p_1 it follows that

$$\partial_t p_1 = \sum_{j \in J_0} \sum_{m \in \mathbb{N}_0^D} (S_j^1 S_j^2 - I) \alpha_j \beta_j p + \sum_{j \in J_1} \sum_{m \in \mathbb{N}_0^D} (S_j^1 S_j^2 - I) \alpha_j \beta_j p.$$

The first sum vanishes, because $S_j^1 = I$ for $j \in J_0$, and

$$\sum_{m \in \mathbb{N}_0^D} (S_j^2 - I) \alpha_j \beta_j p = 0 \tag{15.17}$$

by Lemma 2 in [19]. Since $S_j^1 S_j^2 - I = S_j^1 (S_j^2 - I) + (S_j^1 - I)$ and since (15.17) implies

$$\sum_{j \in J_1} \sum_{m \in \mathbb{N}_0^D} S_j^1 (S_j^2 - I) \alpha_j \beta_j p = 0,$$

we obtain

$$\partial_t p_1 = \sum_{j \in J_1} \sum_{m \in \mathbb{N}_0^D} (S_j^1 - I) \alpha_j \beta_j p = \sum_{j \in J_1} (S_j^1 - I) \alpha_j \sum_{m \in \mathbb{N}_0^D} \beta_j p_2 p_1. \tag{15.18}$$

Following the ideas of [3] we use the Taylor expansion

$$\beta_j(m) = \beta_j(\xi) + \nabla\beta_j(\xi)^T(m-\xi) + \frac{1}{2}(m-\xi)^T\left(\nabla^2\beta_j\right)(m-\xi) \quad (15.19)$$

where $\xi = \xi(t,n)$. Since β_j is at most quadratic by Assumption 15.2, all higher-order terms vanish. This yields[2]

$$\sum_{m\in\mathbb{N}_0^D} \beta_j(m)\,p_2(t,m|n) = \beta_j(\xi) + R_j(t,n) \quad (15.20)$$

$$\text{with} \quad R_j(t,n) = \operatorname{trace}\left(C_\xi(t,n)\nabla^2\beta_j\right)$$

because $\sum_{m\in\mathbb{N}_0^D} p_2(t,m|n) = 1$ and $\sum_{m\in\mathbb{N}_0^D}(m-\xi(t,n))p_2(t,m|n) = 0$. Substituting this into (15.18) gives

$$\partial_t p_1 = \sum_{j\in J_1}(S_j^1 - I)\alpha_j\beta_j(\xi)p_1 + \mathfrak{R} = A(\xi)p_1 + \mathfrak{R} \quad (15.21)$$

with a rest term $\mathfrak{R} = \mathfrak{R}(t,n)$ given by

$$\mathfrak{R} = \sum_{j\in J_1}(S_j^1 - I)\alpha_j R_j p_1.$$

Comparing (15.21) with (15.9) yields

$$\partial_t p_1 - \partial_t w = A(\xi)p_1 - A(\phi)w + \mathfrak{R}$$

and since $p_1(0,\cdot) = w(0,\cdot)$ according to (15.11), we obtain

$$\|p_1(t,\cdot) - w(t,\cdot)\|_{\ell^1} \le \int_0^t \left\|A\big(\xi(s,\cdot)\big)p_1(s,\cdot) - A\big(\phi(s,\cdot)\big)w(s,\cdot)\right\|_{\ell^1} ds \quad (15.22a)$$

$$+ \int_0^t \left\|\mathfrak{R}(s,\cdot)\right\|_{\ell^1} ds. \quad (15.22b)$$

Our next goal is to derive a bound for the second term (15.22b). According to Assumption 15.4 we have

$$\sup_{t\in[0,t_{\text{end}}]}\sup_{n\in\mathbb{N}_0^d}|C_\xi(t,n)| \le \frac{C}{\varepsilon},$$

[2]The remainder term R_j is not to be mixed up with the reaction channel R_j in (15.1).

whereas (15.2) yields

$$\nabla^2 \beta_j = \begin{cases} 0 & \text{if } |b_j| \leq 1 \\ \varepsilon^{\gamma(j)+1} & \text{if } |b_j| = 2. \end{cases}$$

By Assumption 15.2, no other cases have to be considered. Hence, it follows that

$$\sup_{s\in[0,t_{\text{end}}]} \sup_{n\in\mathbb{N}_0^d} |R_j(s,n)| = \sup_{s\in[0,t_{\text{end}}]} \sup_{n\in\mathbb{N}_0^d} \left|\text{trace}\left(C_\xi(s,n)\nabla^2\beta_j\right)\right| \leq C\varepsilon^{\gamma(j)}, \tag{15.23}$$

and Assumption 15.5 and the fact that $\gamma(j) = 1$ for all $j \in J_1$ yield the estimate

$$\int_0^t \left\|\mathfrak{R}(s,\cdot)\right\|_{\ell^1} ds \leq Ct\varepsilon \sup_{s\in[0,t]} \sup_{j\in J_1} \|\alpha_j(s,\cdot)p_1(s,\cdot)\|_{\ell^1} \leq \tilde{C}_b t\varepsilon. \tag{15.24}$$

If $|b_j| \in \{0,1\}$ for all $j = 1,\ldots,r$, then $\nabla^2\beta_j = 0$ and hence $\left\|\mathfrak{R}(s,\cdot)\right\|_{\ell^1} = 0$ such that one can choose $\tilde{C}_b = 0$ in the special case (15.14). The first error term (15.22a) can be bounded by

$$\begin{aligned} &\left\|A(\xi)p_1 - A(\phi)w\right\|_{\ell^1} \\ &= \Big\| \sum_{j\in J_1} (S_j^1 - I)\Big(\beta_j(\xi)\alpha_j p_1\Big) - \sum_{j\in J_1} (S_j^1 - I)\Big(\beta_j(\phi)\alpha_j w\Big) \Big\|_{\ell^1} \\ &\leq C \max_{j\in J_1} \left\|\beta_j(\xi)\alpha_j p_1 - \beta_j(\phi)\alpha_j w\right\|_{\ell^1} \\ &\leq C \max_{j\in J_1} \left\|\beta_j(\xi)p_1 - \beta_j(\phi)w\right\|_{\ell^1} \end{aligned} \tag{15.25}$$

due to Assumption 15.5. Applying Lemma 15.1 now yields

$$\left\|\beta_j(\xi)p_1 - \beta_j(\phi)w\right\|_{\ell^1} \leq C\varepsilon^{\gamma(j)}\left(\left\|\xi p_1 - \phi w\right\|_{\ell^1} + \varepsilon^{-1}\left\|p_1 - w\right\|_{\ell^1}\right),$$

and since the maximum in (15.25) is only taken over J_1, it follows that

$$\left\|A(\xi)p_1 - A(\phi)w\right\|_{\ell^1} \leq C\left(\varepsilon\left\|\xi p_1 - \phi w\right\|_{\ell^1} + \left\|p_1 - w\right\|_{\ell^1}\right). \tag{15.26}$$

Substituting (15.24) and (15.26) into (15.22a) and (15.22b) yields the assertion. □

Lemma 15.3. *Under the assumptions of Theorem 15.1 there are constants $\tilde{C}_b \geq 0$ and $C > 0$ such that*

$$\|(\xi p_1 - \phi w)(t,\cdot)\|_{\ell^1} \le \tilde{C}_b t + C \int_0^t \|(\xi p_1 - \phi w)(s,\cdot)\|_{\ell^1}\, ds + \frac{C}{\varepsilon} \int_0^t \|p_1(s,\cdot) - w(s,\cdot)\|_{\ell^1}\, ds.$$

for all $t \in [0, t_{end}]$. *The constants* $\tilde{C}_b$ *and* C *do not depend on* t *or* ε. *If* $|b_j| \in \{0,1\}$ *for all* $j = 1,\dots,r$, *then we can choose* $\tilde{C}_b = 0$.

Proof. With similar arguments as in the proof of Lemma 15.2, it can be shown that

$$\begin{aligned}\partial_t (\xi p_1)(t,n) = &\sum_{j=1}^{r} \mu_j \sum_{m\in\mathbb{N}_0^D} \left(S_j^1 \alpha_j \beta_j p\right)(t,n,m) \\ &+ \sum_{j\in J_1} \sum_{m\in\mathbb{N}_0^D} m \left((S_j^1 - 1)\alpha_j \beta_j p\right)(t,n,m)\end{aligned} \tag{15.27}$$

(see step 1 in the proof of Lemma 6 in [19] for details). For the first term on the right-hand side, (15.20) yields

$$\begin{aligned}\sum_{m\in\mathbb{N}_0^D} \left(S_j^1 \alpha_j \beta_j p\right)(t,n,m) &= \alpha_j(n-\nu_j) \left(\sum_{m\in\mathbb{N}_0^D} \beta_j(m) p_2(t,m \mid n-\nu_j) \right) p_1(t,n-\nu_j) \\ &= \alpha_j(n-\nu_j)\Big(\beta_j\big(\xi(t,n-\nu_j)\big) + R_j(t,n-\nu_j)\Big) p_1(t,n-\nu_j) \\ &= S_j^1\big(\alpha_j\big[\beta_j(\xi) + R_j\big] p_1\big)(t,n).\end{aligned}$$

Moreover, it follows from (15.19) that

$$\sum_{m\in\mathbb{N}_0^D} m\beta_j(m) p(t,n,m) = \Big(\beta_j(\xi)\xi + T_j(t,n)\Big) p_1(t,n)$$

with $\xi = \xi(t,n)$ and

$$\begin{aligned}T_j(t,n) =& C_\xi(t,n)\nabla\beta_j(\xi) + \frac{1}{2}\xi R_j(t,n) \\ &+ \frac{1}{2} \sum_{m\in\mathbb{N}_0^D} (m-\xi)(m-\xi)^T (\nabla^2\beta_j)(m-\xi) p_2(t,m \mid n).\end{aligned}$$

Substituting into (15.27) yields

$$\partial_t\,(\xi p_1)\,(t,n) = \sum_{j=1}^{r} \mu_j\, S_j^1\big(\alpha_j \beta_j\,(\xi)\, p_1\big)(t,n) + \sum_{j\in J_1} \Big((S_j^1 - 1)\alpha_j \beta_j\,(\xi)\xi p_1\Big)\,(t,n) + \mathfrak{R}(t,n)$$

with defect

$$\mathfrak{R}(t,n) = \sum_{j=1}^{r} \mu_j\, S_j^1\big(R_j \alpha_j\, p_1\big)(t,n) + \sum_{j\in J_1} \Big((S_j^1 - 1) T_j \alpha_j\, p_1\Big)\,(t,n).$$

Comparing this with (15.10) shows that

$$\partial_t(\xi p_1)(t,n) = F(\xi, p_1) + G(\xi, p_1) + \mathfrak{R}.$$

We will now prove that $\|\mathfrak{R}(t,\cdot)\|_{\ell^1} \leq \tilde{C}_b$ with a constant $\tilde{C}_b \geq 0$ which does not depend on ε nor on $t \in [0, t_{\text{end}}]$. In the special case (15.14), we have that $\nabla \beta_j\,(\xi) = 0$ for all $j \in J_1$ and $\nabla^2 \beta_j(\xi) = 0$ for all $j = 1, \ldots, r$. This implies $R_j = 0$ for all j, $T_j = 0$ for all $j \in J_1$ and hence $\big\|\mathfrak{R}(s,\cdot)\big\|_{\ell^1} = 0$ such that one can choose $\tilde{C}_b = 0$. If (15.14) is not true, then according to (15.23) we know that $|R_j\,(s,n)| \leq C$ for all $j = 1, \ldots, r$, and with straightforward calculations and Assumption 15.5 we obtain the bound

$$\Big\| \sum_{j=1}^{r} \mu_j\, S_j^1\big(R_j \alpha_j\, p_1\big)(t,\cdot)\Big\|_{\ell^1} \leq C \max_{j=1,\ldots,r} \|\alpha_j\,(\cdot)\, p_1(t,\cdot)\|_{\ell^1} \leq C.$$

Concerning the second term in $\mathfrak{R}$, Assumption 15.4 and (15.2) imply that

$$\sup_{t\in[0,t_{\text{end}}]} \sup_{n\in\mathbb{N}_0^d} |T_j\,(t,n)| \leq C\,\varepsilon^{\gamma(j)-1}.$$

The sum in the second term of $\mathfrak{R}$ is only taken over $j \in J_1$ such that $\gamma(j) = 1$. With Assumption 15.5, it thus follows that

$$\Big\| \sum_{j\in J_1} \sum_{m\in\mathbb{N}_0^D} \Big((S_j^1 - 1) T_j \alpha_j\, p_1\Big)\,(t,\cdot)\Big\|_{\ell^1} \leq C \max_{j=1,\ldots,r} \|\alpha_j\,(\cdot)\, p_1(t,\cdot)\|_{\ell^1} \leq C,$$

which proves $\|\mathfrak{R}(t,\cdot)\|_{\ell^1} \leq \tilde{C}_b$. Now the error $\xi p_1 - \phi w$ can be estimated by

$$\begin{aligned} \|(\xi p_1)(t,\cdot) - (\phi w)(t,\cdot)\|_{\ell^1} &\leq \int_0^t \|\partial_t(\xi p_1)(s,\cdot) - \partial_t(\phi w)(s,\cdot)\|_{\ell^1}\, ds \\ &\leq \int_0^t \|F(\xi p_1)(s,\cdot) - F(\phi w)(s,\cdot)\|_{\ell^1}\, ds \end{aligned} \tag{15.28}$$

$$+ \int_0^t \|G(\xi p_1)(s,\cdot) - G(\phi w)(s,\cdot)\|_{\ell^1}\, ds + \tilde{C}_b t.$$

It follows from Assumption 15.5 and Lemma 15.1 that

$$\begin{aligned}
&\|G(\xi, p_1)(s,\cdot) - G(\phi, w)(s,\cdot)\|_{\ell^1} \\
&\le C \max_{j=1,\dots,r} \left\|\alpha_j \left[\beta_j(\xi) p_1 - \beta_j(\phi) w\right](s,\cdot)\right\|_{\ell^1} \\
&\le C \max_{j=1,\dots,r} \left\|\left[\beta_j(\xi) p_1 - \beta_j(\phi) w\right](s,\cdot)\right\|_{\ell^1} \\
&\le C \|\xi p_1(s,\cdot) - \phi w(s,\cdot)\|_{\ell^1} + \frac{C}{\varepsilon} \|p_1(s,\cdot) - w(s,\cdot)\|_{\ell^1}. \qquad (15.29)
\end{aligned}$$

A corresponding bound has to be shown for $F(\xi p_1) - F(\phi w)$. Assumption 15.5 yields

$$\begin{aligned}
\|F(\xi p_1)(s,\cdot) - F(\phi w)(s,\cdot)\|_{\ell^1} &\le C \max_{j \in J_1} \left\|\alpha_j \beta_j(\xi) \xi p_1(s,\cdot) - \alpha_j \beta_j(\phi) \phi w(s,\cdot)\right\|_{\ell^1} \\
&\le C \max_{j \in J_1} \left\|\beta_j(\xi) \xi p_1(s,\cdot) - \beta_j(\phi) \phi w(s,\cdot)\right\|_{\ell^1}.
\end{aligned}$$

We decompose the error into three parts:

$$\begin{aligned}
&\|F(\xi p_1)(s,\cdot) - F(\phi w)(s,\cdot)\|_{\ell^1} \\
&\le C \max_{j \in J_1} \Big(\left\|\beta_j(\xi)[\xi p_1 - \phi w](s,\cdot)\right\|_{\ell^1} + \left\|\beta_j(\xi)\phi[w - p_1](s,\cdot)\right\|_{\ell^1} \\
&\qquad + \left\|\phi[\beta_j(\xi) p_1 - \beta_j(\phi) w](s,\cdot)\right\|_{\ell^1} \Big).
\end{aligned}$$

Since (15.2) and Assumption 15.4 imply that for every $j \in J_1$

$$\begin{aligned}
\sup_{t \in [0,t_{\text{end}}]} \sup_{n \in \mathbb{N}_0^d} \left|\beta_j\big(\xi(t,n)\big)\right| &\le C \varepsilon^{\gamma(j)-1} \varepsilon^{|b_j|} \sup_{t \in [0,t_{\text{end}}]} \sup_{n \in \mathbb{N}_0^d} \left|\big(\xi(t,n)\big)\right|^{|b_j|} \\
&\le C \varepsilon^{\gamma(j)-1} = C,
\end{aligned}$$

and since $\sup_{n \in \mathbb{N}_0^d} |\phi(s,n)| \le \frac{C}{\varepsilon}$ by Assumption 15.4, we obtain

$$\begin{aligned}
&\|F(\xi p_1)(s,\cdot) - F(\phi w)(s,\cdot)\|_{\ell^1} \\
&\le C \left\|[\xi p_1 - \phi w](s,\cdot)\right\|_{\ell^1} \\
&\qquad + \frac{C}{\varepsilon} \Big(\|w(s,\cdot) - p_1(s,\cdot)\|_{\ell^1} + \max_{j \in J_1} \left\|\left[\beta_j(\xi) p_1 - \beta_j(\phi) w\right](s,\cdot)\right\|_{\ell^1} \Big).
\end{aligned}$$

Applying Lemma 15.1 to the last term yields

$$\max_{j\in J_1}\left\|\left[\beta_j(\xi)p_1-\beta_j(\phi)w\right](s,\cdot)\right\|_{\ell^1}\leq C\varepsilon\|\left[\xi p_1-\phi w\right]\|_{\ell^1}+C\|p_1-w\|_{\ell^1}$$

because $\gamma(j)=1$ for $j\in J_1$. Hence, we have shown the estimate

$$\begin{aligned}\|F(\xi p_1)(s,\cdot)-F(\phi w)(s,\cdot)\|_{\ell^1}\leq& C\left\|\xi p_1(s,\cdot)-\phi w(s,\cdot)\right\|_{\ell^1}\\&+\frac{C}{\varepsilon}\left\|w(s,\cdot)-p_1(s,\cdot)\right\|_{\ell^1}\end{aligned}\tag{15.30}$$

Substituting (15.29) and (15.30) into (15.28) proves the assertion. □

References

1. Burrage, K., Hegland, M., MacNamara, S., Sidje, R.B.: A Krylov-based finite state projection algorithm for solving the chemical master equation arising in the discrete modelling of biological systems. In: Langville, A.N., Stewart, W.J. (eds.) Markov Anniversary Meeting: An International Conference to Celebrate the 150th Anniversary of the Birth of A.A. Markov, Charleston, pp. 21–38. Boson Books (2006)
2. Dolgov, S.V., Khoromskij, B.N.: Tensor-product approach to global time-space-parametric discretization of chemical master equation. Technical report, Max-Planck-Institut für Mathematik in den Naturwissenschaften (2012)
3. Engblom, S.: Computing the moments of high dimensional solutions of the master equation. Appl. Math. Comput. **180**(2), 498–515 (2006)
4. Engblom, S.: Spectral approximation of solutions to the chemical master equation. J. Comput. Appl. Math. **229**(1), 208–221 (2009)
5. Ferm, L., Lötstedt, P., Hellander, A.: A hierarchy of approximations of the master equation scaled by a size parameter. J. Sci. Comput. **34**, 127–151 (2008)
6. Gillespie, D.T.: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. **22**, 403–434 (1976)
7. Gillespie, D.T.: A rigorous derivation of the chemical master equation. Physica A **188**, 404–425 (1992)
8. Griebel, M., Jager, L.: The BGY3dM model for the approximation of solvent densities. J. Chem. Phys. **129**(17), 174,511–174,525 (2008)
9. Hasenauer, J., Wolf, V., Kazeroonian, A., Theis, F.: Method of conditional moments for the chemical master equation. J. Math. Biol. (2013, Publication online) (Technical report)
10. Hegland, M., Garcke, J.: On the numerical solution of the chemical master equation with sums of rank one tensors. In: McLean, W., Roberts, A.J. (eds.) Proceedings of the 15th Biennial Computational Techniques and Applications Conference (CTAC-2010), Sydney, pp. C628–C643 (2011)
11. Hegland, M., Hellander, A., Lötstedt, P.: Sparse grids and hybrid methods for the chemical master equation. BIT **48**, 265–284 (2008)
12. Hellander, A., Lötstedt, P.: Hybrid method for the chemical master equation. J. Comput. Phys. **227**, 100–122 (2007)
13. Henzinger, T., Mateescu, M., Mikeev, L., Wolf, V.: Hybrid numerical solution of the chemical master equation. In: Quaglia, P. (ed.) Proceedings of the 8th International Conference on Computational Methods in Systems Biology (CMSB'10), Trento, pp. 55–65. ACM (2010)

14. Higham, D.J.: Modeling and simulating chemical reactions. SIAM Rev. **50**(2), 347–368 (2008)
15. Iedema, P.D., Wulkow, M., Hoefsloot, H.C.J.: Modeling molecular weight and degree of branching distribution of low-density polyethylene. Macromolecules **33**, 7173–7184 (2000)
16. Jahnke, T.: An adaptive wavelet method for the chemical master equation. SIAM J. Sci. Comput. **31**(6), 4373–4394 (2010)
17. Jahnke, T.: On reduced models for the chemical master equation. SIAM Multiscale Model. Simul. **9**(4), 1646–1676 (2011)
18. Jahnke, T., Huisinga, W.: A dynamical low-rank approach to the chemical master equation. Bull. Math. Biol. **70**(8), 2283–2302 (2008)
19. Jahnke, T., Kreim, M.: Error bound for piecewise deterministic processes modeling stochastic reaction systems. SIAM Multiscale Model. Simul. **10**(4), 1119–1147 (2012)
20. Jahnke, T., Udrescu, T.: Solving chemical master equations by adaptive wavelet compression. J. Comput. Phys. **229**(16), 5724–5741 (2010)
21. Kazeev, V., Khammash, M., Nip, M., Schwab, C.: Direct solution of the chemical master equation using quantized tensor trains. Technical report, ETH Zurich (2013)
22. Kazeev, V., Schwab, C.: Tensor approximation of stationary distributions of chemical reaction networks. Technical report, ETH Zurich (2013)
23. Kurtz, T.G.: The relationship between stochastic and deterministic models of chemical reactions. J. Chem. Phys. **57**, 2976–2978 (1973)
24. Lötstedt, P., Ferm, L.: Dimensional reduction of the Fokker-Planck equation for stochastic chemical reactions. Multiscale Model. Simul. **5**, 593–614 (2006)
25. Menz, S., Latorre, J.C., Schütte, C., Huisinga, W.: Hybrid stochastic-deterministic solution of the chemical master equation. SIAM Multiscale Model. Simul. **10**(4), 1232–1262 (2012)
26. Munsky, B., Khammash, M.: The finite state projection algorithm for the solution of the chemical master equation. J. Chem. Phys. **124**(4), 044,104 (2006)
27. Sunkara, V.: Analysis and numerics of the chemical master equation. Ph.D. thesis, Australian National University (2013)
28. Sunkara, V.: Finite state projection method for hybrid models. Technical report, Karlsruhe Institute of Technology (2013)

Chapter 16
Valuation of Structured Financial Products by Adaptive Multiwavelet Methods in High Dimensions

Rüdiger Kiesel, Andreas Rupp, and Karsten Urban

Abstract We introduce a new numerical approach to value structured financial products. These financial products typically feature a large number of underlying assets and require the explicit modeling of the dependence structure of these assets. We follow the approach of Kraft and Steffensen (Rev Finance 11:209–252, 2006), who explicitly describe the possible value combinations of the assets via a Markov chain with a portfolio state space. As the number of states increases exponentially with the number of assets in the portfolio, this model so far has been – despite its theoretical appeal – not computationally tractable. The price of a structured financial product in this model is determined by a coupled system of parabolic PDEs, describing the value of the portfolio for each state of the Markov chain depending on the time and macroeconomic state variables. A typical portfolio of n assets leads to a system of $N = 2^n$ coupled parabolic partial differential equations. It is shown that this high number of PDEs can be solved by combining an adaptive multiwavelet method with the Hierarchical Tucker Format. We present numerical results for $n = 128$.

16.1 Introduction

The inadequate pricing of Asset-backed securities (ABS) and in particular Collateralized Debt Obligations (CDOs), on which we focus, is widely viewed as a main trigger of the financial crisis that started in 2007, [7, 17]. The lack of adequate mathematical models to capture the (dependency) risk structure, [23], of these assets is consistently identified as the main reason for the inaccurate pricing. Due to the

R. Kiesel (✉)
University of Duisburg-Essen, Universitätsstraße 12, 45141 Essen, Germany
e-mail: ruediger.kiesel@uni-due.de

A. Rupp • K. Urban
University of Ulm, Helmholtzstr. 20, 89069 Ulm, Germany
e-mail: andreas.rupp@uni-ulm.de; karsten.urban@uni-ulm.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_16

complexity of a CDO portfolio, which arises from the high number of possible default combinations, drastic simplifications of the underlying portfolio structure had to be made in order to compute a price, [8, 10, 43].

We consider the CDO model of [26] where the value of a CDO portfolio is determined by a system of coupled parabolic PDEs, each PDE describing the portfolio value for a specific default situation. These situations are characterized by a discrete Markov chain, where each state in the chain stands for a default state of the portfolio. Therefore, for a portfolio of n assets, there are $N = 2^n$ possible combinations of defaults and, therefore, 2^n states in the Markov chain. It will later turn out to be convenient to label the states in the index set $\mathcal{N} := \{0, \dots, N-1\}$. The value of the CDO portfolio in [26] is described by the function $\mathbf{u}(t,y) = (u^0(t,y), \dots, u^{N-1}(t,y))^T$ that satisfies the partial differential equation for all $t \in (0,T)$ ($T > 0$ being the maturity) and all $y \in \Omega \subset \mathbb{R}^M$. The y variables are used to incorporate M economic market factors which describe the state of the economy.

$$u_t^j(t,y) = -\frac{1}{2}\nabla \cdot (\mathbf{B}(t)\nabla u^j(t,y)) - \boldsymbol{\alpha}^T(t)\nabla u^j(t,y) + r(t,y)u^j(t,y) - \sum_{k \in \mathcal{N}\setminus\{j\}} q^{j,k}(t,y)(a^{j,k}(t,y) + u^k(t,y) - u^j(t,y)) - c^j(t,y), \tag{16.1a}$$

$$\mathbf{u}(t,y) = 0,\ t \in (0,T),\ y \in \partial\Omega, \quad \mathbf{u}(T,y) = (u_T^0(y), \dots, u_T^{N-1}(y))^T,\ y \in \Omega, \tag{16.1b}$$

for all $j \in \mathcal{N}$. The differential operator ∇ is to be understood w.r.t. y. Often the bounded domain Ω arises from localizing the problem from $\mathbb{R}^M$ to a bounded domain by truncation. This is a generalized Black-Scholes PDE with a linear coupling, homogeneous Dirichlet boundary conditions in y (possibly after localization) and terminal condition (16.1b). The remaining parameters can be interpreted as follows:

- The space variables $y \in \Omega \subset \mathbb{R}^M$ describe the current market situation by means of variables which describe the market influence on the CDO portfolio. This could be for example interest rates, foreign exchange rates, macroeconomic factors and other factors depending on the composition of the portfolio. These space variables are modeled via a market process $dY(t) = \boldsymbol{\alpha}(t)dt + \boldsymbol{\beta}(t)dW(t)$, where $W(t)$ is a M-dimensional standard Brownian motion, the drift $\boldsymbol{\alpha}(t)$ is a M-dimensional vector and the volatility $\boldsymbol{\beta}(t) \in \mathbb{R}^{M\times M}$. Then, we abbreviate $\mathbf{B}(t) := \boldsymbol{\beta}(t)\boldsymbol{\beta}(t)^T$. W.l.o.g., we may assume that $\Omega = [0,1]^M$.
- $\mathcal{N}$ is the state space of a Markov chain, where each state is a possible combination of defaults of the underlying portfolio.
- The function $r(t,y)$ describes the relevant market interest rate.
- The parameters $q^{j,k} \geq 0$ are the transition intensities, which is the instantaneous change in the transition probabilities, from state j into state k, where

$j, k \in \mathcal{N}$. Moreover, for any state j, all intensities sum up to zero, i.e., $q^{j,j} := -\sum_{k \in \mathcal{N} \setminus \{j\}} q^{j,k}$. The default probability is assumed to increase over time, [26].

- The payments $c^j(t, y)$, $j \in \mathcal{N}$, made by the CDO are continuous in time.
- The recovery payment, i.e., the distribution of the remaining funds of the defaulted firm, is denoted by $a^{j,k}(t, y)$. It depends on the transition from state j to state k, which means on the defaulted firm.
- Final payments at maturity can also be included. They also depend on the state and the current market situation and are denoted by $u_T^j(y)$.

All together, (16.1) is a system of $N = 2^n$ coupled time-dependent parabolic PDEs each in dimension M. The difficulty of this pricing approach is primarily the high number $N = 2^n$ of states in the Markov chain and, hence, the high number of coupled partial differential equations. In the following it will be shown, that under reasonable conditions, the high dimensionality resulting from the Markov chain can be separated as a time dependent factor from the actual solution of the partial differential equation. This allows us to represent the system of coupled partial differential equations in variational form as the variational formulation of a high dimensional parabolic partial differential equation. We propose to use orthogonal multiwavelet bases to develop an equivalent discrete but infinite-dimensional system. This particular choice allows us to write the system as a tensor product, which in turns leads to decoupling the Markov chain ingredients from the market parameters, i.e., the high dimensionality is separated from the integrals of the test and trial spaces. The hierarchical Tucker Format (HTF), is then applied to this tensor structure. To numerically approximate a solution for this system, multiwavelets ensure small condition numbers regardless of the dimension of the process. Moreover, this choice allows us to use asymptotically optimal adaptive schemes, e.g. [24].

In the context of wavelet approximations of solutions of partial differential equations, the term "high dimensional" commonly refers to the dimension M of the space variable, say $M \geq 5$. In our problem at hand, we also have a huge number $N = 2^n$ of coupled equations. As already mentioned, we will show that we can separate both ingredients, namely the Markov chain state space $\mathcal{N}$ and the macroeconomic model $\Omega \subseteq \mathbb{R}^M$. The latter one will be discretized by a tensor product multiwavelet basis. In general, the dimension of the basis grows exponentially with M. Thus, the number of macroeconomic variables that can be used is often strongly limited by the available memory. This can be seen in [13], where the number of degrees of freedom in 10 dimensions is not enough to reach the optimal convergence rate. In [34] it can also be seen, that the number of degrees of freedom which can be used in 5 dimensions is strongly limited. By applying principal component analysis, [35], the authors are able to solve a problem in 30 dimensions essentially by a reduction to 5 dimensions. In [22], 8 dimensions are reached for a full rank Black Scholes model and 16 dimensions, when a stochastic volatility model is considered.

The remainder of this paper is organized as follows. In Sect. 16.2, we derive a variational formulation to (16.1) and prove its well-posedness. Section 16.3 is devoted to the description of well-known multiwavelet bases and the collection of the main properties that are needed here. The discretization in Sect. 16.4 is done in three steps. First, we use the multiwavelet basis in order to derive an equivalent discrete but infinite-dimensional system. We also show that this approach allows us to decouple the market variables from the Markov chain state space in terms of a tensor product. The next two steps involve the discretization in time and market variables. Due to the mentioned separation, we can handle large portfolios of companies by the so-called Hierarchical Tucker Format (HTF) which is briefly reviewed in Sect. 16.5 concentrating on those properties that are relevant here. Finally, in Sect. 16.6, we report on some numerical experiments for realistic market scenarios. We collect some auxiliary facts in Sect. 16.7.

16.2 Variational Formulation

We start by deriving a variational formulation of the original system (16.1). We begin with some remarks on *systems* of elliptic partial differential equations. Let $V \hookrightarrow H \hookrightarrow V'$ be a Gelfand triple and $\mathbf{V} := V^N$ be the tensor product space. For $\mathbf{u} = (u^0, \dots, u^{N-1})^T$, $\mathbf{v} = (v^0, \dots, v^{N-1})^T \in \mathbf{V}$, let $a^j : \mathbf{V} \times V \to \mathbb{R}$ be a bilinear form and $f^j : V \to \mathbb{R}$, $j = 0, \dots, N-1$, a linear form. Then,

$$\mathbf{u} \in \mathbf{V}: \quad a^j(\mathbf{u}, v) = f^j(v) \quad \forall\, v \in V,\ j \in \mathscr{N}, \tag{16.2}$$

is a coupled linear system of N equations. Defining $\mathbf{a} : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$, $\mathbf{f} : \mathbf{V} \to \mathbb{R}$ by $\mathbf{a}(\mathbf{u}, \mathbf{v}) := \sum_{j \in \mathscr{N}} a^j(\mathbf{u}, v^j)$, $\mathbf{f}(\mathbf{v}) := \sum_{j \in \mathscr{N}} f^j(v^j)$, $\mathbf{u}, \mathbf{v} \in \mathbf{V}$, we obtain

$$\mathbf{u} \in \mathbf{V}: \quad \mathbf{a}(\mathbf{u}, \mathbf{v}) = \mathbf{f}(\mathbf{v}) \quad \forall\, \mathbf{v} \in \mathbf{V}, \tag{16.3}$$

which is well-posed provided the Nečas conditions are valid, [32]. Note that (16.2) and (16.3) are equivalent using the test functions $v^j \boldsymbol{\delta}_j$, where $\boldsymbol{\delta}_j = (\delta_{j,j'})_{j' \in \mathscr{N}}^T$ ($\delta_{i,j}$ is the Kronecker delta) in (16.3) yields (16.2); the other direction is trivial.

Next, we need to separate the high dimensional Markov chain parts from the variational formulation. This means that the state dependent variables are compound functions of a state dependent part (which might also depend on the time t) and a mutual factor depending on the space variables y. Hence, we assume that there exist functions $\tilde{q}^{j,k}, \tilde{a}^{j,k}, \tilde{c}^j : [0,T] \to \mathbb{R}$, constants $\tilde{a}^j \in \mathbb{R}$ and functions $h_q, h_a, h_c, h_{a(T)} : \Omega \to \mathbb{R}$ such that

$$q^{j,k}(t,y) = \tilde{q}^{j,k}(t)\, h_q(y), \qquad a^{j,k}(t,y) = \tilde{a}^{j,k}(t)\, h_a(y), \tag{16.4a}$$

$$c^j(t,y) = \tilde{c}^j(t)\, h_c(y), \qquad a^j(y) = \tilde{a}^j\, h_{a(T)}(y), \tag{16.4b}$$

for all $j,k \in \mathcal{N}$, $t \in [0,T]$ and $y \in \Omega$. This is a reasonable assumption from the financial point of view since it states that the dependency on the market process is the same for all points in time and for all states in the Markov chain. The fact, that changes of the state of the Markov chain cannot alter the dependency of the market process Y means that default of single firms in the CDO portfolio will not change the market situation. Finally, we remark that there are methods available in order to obtain an approximate representation of the form (16.4) even in cases where the functions do not directly allow such a separation of variables, see e.g. [5, 33].

We are now going to derive a variational formulation. We need one more abbreviation: If $v : (0,T) \times \Omega \to \mathbb{R}$ is a function in time and space, we will always abbreviate $v(t) : \Omega \to \mathbb{R}$, where $v(t)(y) := v(t,y)$, $y \in \Omega$.

Definition 16.1. Given assumption (16.4), a function $\mathbf{u} \in \mathbf{X} := L_2(0,T; H_0^1(\Omega)^N) \cap H^1(0,T; H^{-1}(\Omega)^N)$ is called *weak solution* of (16.1) if

$$\begin{cases} (\mathbf{u}_t(t), \mathbf{v})_{0;\Omega} + \mathbf{a}(\mathbf{u}(t), \mathbf{v}) = (\mathbf{f}(t), \mathbf{v})_{0;\Omega} \quad \text{for all } \mathbf{v} \in H_0^1(\Omega)^N, t \in (0,T) \\ \mathbf{u}(T,y) = \mathbf{u}_T(y) := (u_T^0(y), \ldots, u_T^{N-1}(y))^T \end{cases} \tag{16.5}$$

where $(\mathbf{w}, \mathbf{v})_{0;\Omega} = \sum_{j \in \mathcal{N}} (w^j, v^j)_{0;\Omega}$, $\mathbf{a}(\mathbf{w}, \mathbf{v}) := \sum_{j \in \mathcal{N}} a^j(\mathbf{w}, v^j)$ with $a^j(\mathbf{w}(t), v) := \frac{1}{2}(\nabla w^j(t), \mathbf{B}(t)\nabla v)_{0;\Omega} - (\boldsymbol{\alpha}(t)^T \nabla w^j(t) + \boldsymbol{\gamma}^j(t)^T \mathbf{w}(t), v)_{0;\Omega}$, the reaction coefficient

$$\gamma_k^j(t,y) := (\boldsymbol{\gamma}^j(t,y))_k := \begin{cases} -\tilde{q}^{j,k}(t)\, h_q(y) & \text{if } k \neq j, \\ r(t) - \sum_{k' \in \mathcal{N} \setminus \{j\}} \tilde{q}^{j,k'}(t)\, h_q(y) & \text{if } k = j. \end{cases} \quad (j,k \in \mathcal{N})$$

and the right-hand side $(\mathbf{f}(t), \mathbf{v})_{0;\Omega} := \sum_{j \in \mathcal{N}} (f^j(t), v^j)_{0;\Omega}$ with $f^j(t) := -\tilde{c}^j(t)\, h_c(y) - \sum_{k \in \mathcal{N} \setminus \{j\}} \tilde{q}^{j,k}(t)\, \tilde{a}^{j,k}(t)\, h_a(y)\, h_q(y)$.

Obviously, (16.5) is a system of instationary convection-diffusion-reaction equation and the linear coupling is in the zero-order (reactive) term.

Theorem 16.1. *Let (16.4) hold. If $\mathbf{u} \in C^1([0,T]; (C^2(\Omega))^N)$ is a classical solution of (16.1), then it is also a weak solution in the sense of Definition 16.1. On the other hand, if $\mathbf{u}$ is a weak solution and additionally $\mathbf{u} \in C^1([0,T]; (C^2(\Omega))^N)$, then $\mathbf{u}$ is also a classical solution of (16.1).*

Proof. We multiply (16.1) by $v^j \in H_0^1(\Omega)$ and obtain $(u_t^j(t), v^j)_{0;\Omega} = (r(t)u^j(t), v^j)_{0;\Omega} - (\boldsymbol{\alpha}(t)^T \nabla u^j(t), v^j)_{0;\Omega} - \frac{1}{2}(\nabla \cdot (\mathbf{B}(t)\nabla u^j(t)), v^j)_{0;\Omega} - \sum_{k \in \mathcal{N} \setminus \{j\}} \int_\Omega q^{j,k} \times (t,y)\,(u^k(t,y) - u^j(t,y))v^j(y)\,dy - \int_\Omega \{c^j(t,y) + \sum_{k \in \mathcal{N} \setminus \{j\}} q^{j,k}(t,y)\, a^{j,k}(t,y)\} \times v^j(y)\,dy$. Using assumption (16.4), the (negative of the) last term reads $\tilde{c}^j(t) \int_\Omega h_c(y)\, v^j(y)\,dy + \sum_{k \in \mathcal{N} \setminus \{j\}} \tilde{q}^{j,k}(t)\, \tilde{a}^{j,k}(t) \int_\Omega h_q(y)\, h_a(y)\, v^j(y)\,dy = (f^j(t), v^j)_{0;\Omega}$. Next, integration by parts gives for the last term

$$\begin{aligned} &\frac{1}{2}(\mathbf{B}(t)\nabla u^j(t)), \nabla v^j)_{0;\Omega} - (\boldsymbol{\alpha}(t)^T \nabla u^j(t), v^j)_{0;\Omega} \\ &+ \int_\Omega \Big\{ r(t)\, u^j(t,y) - \sum_{k \in \mathcal{N} \setminus \{j\}} \tilde{q}^{j,k}(t)\, h_q(y) \Big(u^k(t,y) - u^j(t,y) \Big) \Big\} v^j(y)\,dy, \end{aligned} \tag{16.6}$$

where the last term is equal to $(\boldsymbol{\gamma}^j(t)^T\mathbf{u}(t), v^j)_{0;\Omega}$. Summing over $j \in \mathcal{N}$ yields (16.5). The above derivation also proves the claim.

Theorem 16.2. *If* $\mathbf{B}(t)$ *has rank* M*, then* (16.5) *is well-posed.*

Proof. We need to show that the bilinear form $\mathbf{a}(\cdot,\cdot)$ satisfies the Gårding inequality and is continuous. Then, the claim follows from the Lax-Milgram theorem.

Remark 16.1. Note that (16.4) is not needed for the well-posedness of (16.5).

Finally, consider a space-time variational formulation of (16.5) by integrating over time. With $\mathbf{B}(\mathbf{u},\mathbf{v}) := \int_0^T [(\mathbf{u}_t(t),\mathbf{v}_1)_{0;\Omega} + \mathbf{a}(\mathbf{u}(t),\mathbf{v}_1)]\,dt + (\mathbf{u}(T),\mathbf{v}_2)_{0;\Omega}$ and $\mathbf{f}(\mathbf{v}) := \int_0^T (\mathbf{f}(t),\mathbf{v}_1(t))_{0;\Omega} + (\mathbf{u}_T,\mathbf{v}_2)_{0;\Omega}$ for $\mathbf{u} \in \mathbf{X}$ and $\mathbf{v} \in \mathbf{Y} := L_2(0,T;H_0^1(\Omega)^N) \times L_2(\Omega)^N$, $\mathbf{v} = (\mathbf{v}_1,\mathbf{v}_2) \in \mathbf{Y}$, the space-time variational formulation reads

$$\mathbf{u} \in \mathbf{X}: \qquad \mathbf{B}(\mathbf{u},\mathbf{v}) = \mathbf{f}(\mathbf{v}) \quad \forall\, \mathbf{v} \in \mathbf{Y}. \tag{16.7}$$

This latter problem is also well-posed following the arguments e.g. in [38, 46].

16.3 Multiwavelets

Since we want to use multiwavelets for the discretization of the macroeconomic variables, we briefly recall some facts of these function systems. A (standard, not multi-) wavelet system is a Riesz basis $\Psi := \{\psi_\lambda : \lambda \in \mathcal{J}\}$ of $L_2(\Omega)$, where $\lambda = (\ell,k)$, $|\lambda| := \ell \geq 0$ denotes the *level* (also steering the size of the support in the sense that $\operatorname{diam}(\operatorname{supp}\psi_\lambda) \sim 2^{-|\lambda|}$) and k indicates the type as well as the position of $\operatorname{supp}\psi_\lambda$, e.g. the center of the support. Wavelets are (among other parameters) characterized by a certain order d of *vanishing moments*, i.e., $\int_\Omega y^r\,\psi_\lambda(y)\,dy = 0$ for all $0 \leq |r| \leq d-1$ and all $\lambda \in \mathcal{J}, |\lambda| > 0$. This means that wavelets necessarily oscillate which also explains the name. Note that the vanishing moment property only holds for $|\lambda| > 0$. Those functions ψ_λ with $|\lambda| = 0$ are *not* wavelets but so-called *scaling functions* and those are generated by a *generator* $\varphi \in C_0(\Omega)$ in the sense that each ψ_λ, $|\lambda| = 0$, is a linear combination of (possibly to Ω restricted) shifts $\varphi(\cdot - k)$, $k \in \mathbb{Z}$. The wavelets ψ_λ, $|\lambda| > 0$, are linear combinations of dilated versions of scaling functions. *Multiwavelets* are built as linear combinations of shifts of *several* generators φ_i, $i = 1,\ldots,m$. The main advantage is that corresponding multiwavelets may be constructed that are (1) piecewise polynomial, (2) L_2-orthogonal and compactly supported with small support size. These three properties are quite useful for numercial methods since they allow an efficient evaluation of an approximation as well as well-conditioned and sparse system matrices.

We use B-spline multiple generators and wavelets as constructed in [16, 19]. These functions are also adapted to finite intervals and allow for homogeneous

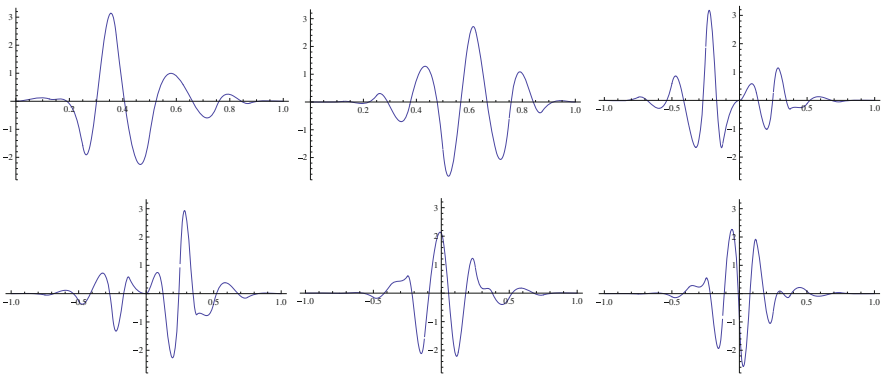


Fig. 16.1 Wavelets generated by the piecewise cubic MRA having one continuous derivative

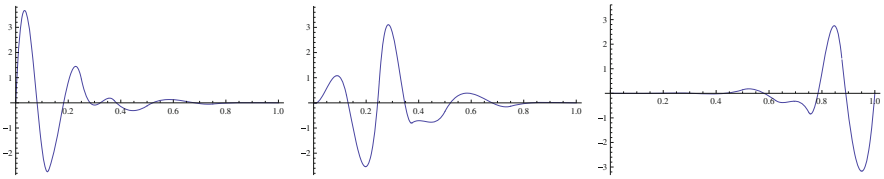


Fig. 16.2 Wavelets with homogeneous Dirichlet boundary conditions generated by a piecewise cubic MRA on $[0, 1]$ with one continuous derivative

Dirichlet boundary conditions, the latter construction was introduced in [36]. We faced some difficulties with the realization of the construction in [16] in particular for higher regularity. However, we finally came up with a realization using Mathematica® for almost arbitrary regularity. Some functions are shown in Figs. 16.1 and 16.2. Details can be found in [36]. Let us summarize some properties that we will need in the sequel.

Proposition 16.1 ([16]). *Let $\Psi = \{\psi_\lambda : \lambda \in \mathcal{J}\}$ be a system of multiwavelets on $\Omega = [0,1]$ from [16] normalized in $H^1(\Omega)$, i.e., $\|\psi_\lambda\|_{1;\Omega} \sim 1$. Then, (a) Ψ is L_2-orthogonal, i.e., $(\psi_\lambda, \psi_\mu)_{0;\Omega} = \delta_{\lambda,\mu}\|\psi_\lambda\|^2_{0;\Omega}$, $\lambda, \mu \in \mathcal{J}$; (b) $\psi_\lambda \in H^1_0(\Omega)$, $\lambda \in \mathcal{J}$; (c) The system Ψ is a Riesz basis for $H^1_0(\Omega)$ with L_2-orthogonal functions.*

Finally, denoting $\boldsymbol{\mathcal{J}} := \{(j,\lambda) : j \in \mathcal{N}; \lambda \in \mathcal{J}\} = \mathcal{N} \times \mathcal{J}$, $\boldsymbol{\lambda} := (j,\lambda) \in \boldsymbol{\mathcal{J}}$ and $\boldsymbol{\psi_\lambda} := \psi_\lambda \boldsymbol{\delta}_j$, $\boldsymbol{\lambda} = (j,\lambda) \in \boldsymbol{\mathcal{J}}$, the system $\boldsymbol{\Psi} := \{\boldsymbol{\psi_\lambda} : \boldsymbol{\lambda} \in \boldsymbol{\mathcal{J}}\}$ is a tensor product Riesz basis for $H^1_0(\Omega)^N$.

16.4 Discretization

16.4.1 An Equivalent ℓ_2-Problem

The first step towards an adaptive multiwavelet method is to rewrite the variational problem (16.5) and (16.7) in a discrete equivalent problem on the sequence space $\ell_2(\mathscr{J})$ for the multiwavelet expansion coefficients. It turns out that the assumption (16.4) is particularly useful here, since it allows us to separate state and space (and time), so that the discrete operators are of tensor product form. This also provides an efficient numerical realization, also for the space-time variational formulation [25] and, in particular, for larger M. Using $\boldsymbol{\Psi}$ as defined in Sect. 16.3, the solution $\mathbf{u}$ of (16.5) has a unique expansion of the form $\mathbf{u}(t,y) = \sum_{\boldsymbol{\lambda}\in\mathscr{J}} \mathbf{x}_{\boldsymbol{\lambda}}(t)\,\boldsymbol{\psi}_{\boldsymbol{\lambda}}(y), t \in (0,T)$, $y \in \Omega$, where $\mathbf{x}_{\boldsymbol{\lambda}}(t) = x^j_\lambda(t)$, $\boldsymbol{\lambda} = (j,\lambda)$, $\mathbf{x}^j(t) = (x^j_\lambda(t))_{\lambda\in\mathscr{J}} \in \ell_2(\mathscr{J})$. The above sum is to be understood componentwise, i.e., $u^j(t,y) = \sum_{\lambda\in\mathscr{J}} x^j_\lambda(t)\,\psi_\lambda(y)$ for $j \in \mathscr{N}$. Then, for $\boldsymbol{\lambda} = (j,\lambda) \in \mathscr{J}$, we get

$$a^j(\mathbf{u}(t),\psi_\lambda) = \sum_{\mu\in\mathscr{J}} x_\mu(t)\Big\{\frac{1}{2}(\nabla\psi_\mu, \mathbf{B}(t)\,\nabla\psi_\lambda)_{0;\Omega} - (\boldsymbol{\alpha}(t)^T\,\nabla\psi_\mu,\psi_\lambda)_{0;\Omega}\Big\} + \sum_{k\in\mathscr{N}} (\gamma^j_k(t)u^k(t),\psi_\lambda)_{0;\Omega}.$$

Defining $\mathbf{A}(t) := \big(\frac{1}{2}(\nabla\psi_\mu,\mathbf{B}(t)\,\nabla\psi_\lambda)_{0;\Omega} - (\boldsymbol{\alpha}(t)^T\,\nabla\psi_\mu,\psi_\lambda)_{0;\Omega}\big)_{\lambda,\mu\in\mathscr{J}}$, the first sum can be abbreviated as $\mathbf{A}(t)\mathbf{x}^j(t)$. Moreover, $\sum_{k\in\mathscr{N}}(\gamma^j_k(t)u^k(t),\psi_\lambda)_{0;\Omega} = \sum_{k\in\mathscr{N}}\sum_{\mu\in\mathscr{J}} x^k_\mu(t)(\gamma^j_k(t)\psi_\mu,\psi_\lambda)_{0;\Omega} = \Big[\sum_{k\in\mathscr{N}} \mathbf{C}^{j,k}(t)\mathbf{x}^k(t)\Big]_\lambda$, where we set $\mathbf{C}^{j,k}(t) := ((\gamma^j_k(t)\psi_\lambda,\psi_\mu)_{0;\Omega})_{\lambda,\mu\in\mathscr{J}} \in \mathbb{R}^{\mathscr{J}\times\mathscr{J}}$.[1] Next, we get $(u^j_t(t),\psi_\lambda)_{0;\Omega} = \sum_{\mu\in\mathscr{J}} \dot{x}_\mu(t)\,(\psi_\mu,\psi_\lambda)_{0;\Omega} = \dot{x}_\lambda(t)$, if the ψ_λ are L_2-orthonormalized. Thus, we obtain $\dot{\mathbf{x}}^j(t) + \mathbf{A}(t)\,\mathbf{x}^j(t) + \sum_{k\in\mathscr{N}} \mathbf{C}^{j,k}(t)\mathbf{x}^k(t)$ for $j \in \mathscr{N}$, or, written as a system

$$\dot{\mathbf{x}}(t) + (\mathscr{A}(t) + \mathscr{C}(t))\,\mathbf{x}(t) = \boldsymbol{f}(t), \tag{16.8}$$

where $\mathscr{A}(t), \mathscr{C}(t) \in \mathbb{R}^{\mathscr{J}\times\mathscr{J}}$ are given by $\mathscr{A}(t) = \operatorname{diag}(\mathbf{A}(t)\ldots,\mathbf{A}(t))$, $(\mathscr{C}(t))_{j,k} = \mathbf{C}^{j,k}(t)$ and $\boldsymbol{f}(t) = (\mathbf{f}(t),\boldsymbol{\psi}_{\boldsymbol{\mu}})_{0;\Omega})_{\boldsymbol{\mu}\in\mathscr{J}}$. Obviously, (16.8) is a coupled system of ODEs in the sequence space $\ell_2(\mathscr{J})$. We will now show that (16.8) allows us to use tensor product techniques for the numerical solution. For that, we need to review some facts on tensor products, which can be found in Sect. 16.7. We detail the coupling term

[1] With a slight abuse of notation, we set $\mathbb{R}^{\mathscr{I}} = \ell_2(\mathscr{I})$ for any countable (possibly infinite) set $\mathscr{I}$ as well as $\mathbb{R}^{\mathscr{I}\times\mathscr{J}}$ as the set of linear operators from $\ell_2(\mathscr{I})$ into $\ell_2(\mathscr{J})$.

$$[\mathbf{C}^{j,k}(t)]_{\lambda,\mu} = (\gamma_k^j(t)\psi_\lambda, \psi_\mu)_{0;\Omega} =: d^{j,k}(t)\mathbf{M}^q_{\lambda,\mu} + r(t)\,\delta_{j,k}\,\delta_{\lambda,\mu}$$
$$= \begin{cases} -\tilde{q}^{j,k}(t)\,(h_q\psi_\lambda, \psi_\mu)_{0;\Omega}, & \text{if } k \neq j, \\ r(t)\delta_{\lambda,\mu} - \sum_{k'\in\mathcal{N}\setminus\{j\}} \tilde{q}^{j,k'}(t)\,(h_q\psi_\lambda, \psi_\mu)_{0;\Omega}, & \text{if } k = j, \end{cases}$$

where $(\mathbf{M}^q)_{\lambda,\mu} := (h_q\psi_\lambda, \psi_\mu)_{0;\Omega}$ is a weighted mass matrix and

$$\mathbb{R} \ni d^{j,k}(t) := \begin{cases} -\tilde{q}^{j,k}(t), & \text{if } k \neq j, \\ -\sum_{m\in\mathcal{N}\setminus\{j\}} \tilde{q}^{j,m}(t), & \text{if } k = j. \end{cases} \tag{16.9}$$

We denote $\mathbf{D}(t) := (d^{j,k}(t))_{j,k\in\mathcal{N}}$.

Theorem 16.3. *Let Assumption (16.4) hold and assume that Ψ satisfies the properties in Proposition 16.1. Then, (16.5) is equivalent to*

$$(\mathbf{I}_\mathcal{N}\otimes\mathbf{I}_\mathscr{J})\dot{\mathbf{x}}(t)+[(\mathbf{I}_\mathcal{N}\otimes[\mathbf{A}(t)+r(t)\mathbf{I}_\mathscr{J}])+(\mathbf{D}(t)\otimes\mathbf{M}^q)]\,\mathbf{x}(t) = \mathbf{b}(t)\otimes\mathbf{g}^1-\tilde{\mathbf{c}}(t)\otimes\mathbf{g}^2, \tag{16.10}$$

where $\mathbf{I}_\mathscr{J}$ denotes the identity w.r.t. an index set $\mathscr{J}$, $\mathbf{b}(t) = (b^j(t))_{j\in\mathcal{N}}$, $b^j(t) := -\sum_{k\in\mathcal{N}\setminus\{j\}} \tilde{q}^{j,k}(t)\,\tilde{a}^{j,k}(t)$, $\tilde{\mathbf{c}}(t) := (\tilde{c}^j(t))_{j\in\mathcal{N}}$, $\mathbf{g}^1 = (g^1_\lambda)_{\lambda\in\mathscr{J}}$, $g^1_\lambda := (h_q\,h_a, \psi_\lambda)_{0;\Omega}$ and $\mathbf{g}^2 = (g^2_\lambda)_{\lambda\in\mathscr{J}}$ with $g^2_\lambda := (h_c, \psi_\lambda)_{0;\Omega}$.

Proof. Let $j \in \mathcal{N}$ and $\lambda \in \mathscr{J}$ so that $\boldsymbol{\lambda} = (j,\lambda) \in \boldsymbol{\mathscr{J}}$. Then,

$$(\boldsymbol{\mathscr{C}}(t)\,\mathbf{x}(t))_{\boldsymbol{\lambda}} = \sum_{k\in\mathcal{N}}\sum_{\mu\in\mathscr{J}} [d^{j,k}(t)\mathbf{M}^q_{\lambda,\mu} + r(t)\delta_{j,k}\delta_{\lambda,\mu}]x^k_\mu(t)$$
$$= \sum_{k\in\mathcal{N}}\sum_{\mu\in\mathscr{J}} [\mathbf{M}^q_{\lambda,\mu}x^k_\mu(t)d^{j,k}(t) + r(t)\delta_{\lambda,\mu}x^k_\mu(t)\delta_{j,k}]$$
$$= \Big([(\mathbf{D}(t)\otimes\mathbf{M}^q) + r(t)(\mathbf{I}_\mathcal{N}\otimes\mathbf{I}_\mathscr{J})]\mathbf{x}(t)\Big)_{\boldsymbol{\lambda}},$$

where we have used Lemma 16.7 in the last step. Note that $\mathbf{D}(t) \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$ $\mathbf{M}^q \in \mathbb{R}^{\mathscr{J}\times\mathscr{J}}$, thus $(\mathbf{D}(t)\otimes\mathbf{M}^q) \in \mathbb{R}^{(\mathcal{N}\times\mathscr{J})\times(\mathcal{N}\times\mathscr{J})} = \mathbb{R}^{\boldsymbol{\mathscr{J}}\times\boldsymbol{\mathscr{J}}}$. Finally,

$$(\mathbf{b}(t)\otimes\mathbf{g}^1 - \tilde{\mathbf{c}}(t)\otimes\mathbf{g}^2)_{(j,\lambda)} = -\sum_{k\in\mathcal{N}\setminus\{j\}} \tilde{q}^{j,k}(t)\,\tilde{a}^{j,k}(t)\,(h_q\,h_a, \psi_\lambda)_{0;\Omega}$$
$$-\tilde{c}^j(t)\,(h_c, \psi_\lambda)_{0;\Omega}$$

which equals $(f^j(t), \psi_\lambda)_{0;\Omega}$, so that the claim is proven.

16.4.2 Temporal Discretization

So far (16.10) is equivalent to the original PDE. As a first step towards a (finite) discretization, we consider the time interval $[0, T]$, fix some $K \in \mathbb{N}$ and set $\Delta t := \frac{1}{K}$ as well as $t^k := k\,\Delta t$, $k = 0, \ldots, K$. Obviously, (16.10) takes the form $\mathscr{M}\dot{\mathbf{x}}(t) + \mathscr{A}(t)\,\mathbf{x}(t) = \mathbf{f}(t)$, $\mathbf{x}(T) = \mathbf{x}_T$, to which we apply the standard θ-scheme ($\theta \in [0, 1]$) to derive an approximation $\mathbf{x}^k \approx \mathbf{x}(t^k)$ by solving $\mathbf{x}^K = \mathbf{x}_T$ and $\frac{1}{\Delta t}\mathscr{M}(\mathbf{x}^{k+1} - \mathbf{x}^k) + \theta\mathscr{A}(t^{k+1})\mathbf{x}^{k+1} + (1-\theta)\mathscr{A}(t^k)\mathbf{x}^k = \theta\mathbf{f}(t^{k+1}) + (1-\theta)\mathbf{f}(t^k)$, for $k = K-1, \ldots, 0$. Obviously, the second equation amounts to solving the following linear system in each time step

$$(\mathscr{M} + \Delta t(\theta - 1)\mathscr{A}(t^k))\mathbf{x}^k = (\mathscr{M} + \Delta t\,\theta\mathscr{A}(t^{k+1}))\mathbf{x}^{k+1} + \theta\mathbf{f}(t^{k+1}) + (1-\theta)\mathbf{f}(t^k). \tag{16.11}$$

16.4.3 Wavelet Galerkin Methods

The last step towards a fully discrete system in space and time is the discretization with respect to the economic variable $y \in \Omega$. After having transformed (16.5) into the discrete but infinite-dimensional system (16.10), this can easily be done by selecting a finite index set $\Lambda \subset \mathscr{J}$. Hence, we obtain $\mathscr{M}_\Lambda := \mathbf{I}_\mathscr{N} \otimes \mathbf{I}_\Lambda$, $\mathscr{A}_\Lambda(t) := \mathbf{I}_\mathscr{N} \otimes [\mathbf{A}_\Lambda(t) + r(t)\mathbf{I}_\Lambda] + \mathbf{D}(t) \otimes \mathbf{M}^q_\Lambda$, which is then inserted into (16.11) for $\mathscr{M}$ and $\mathscr{A}(t)$, respectively, in order to get a finite system. We denote by $\mathbf{A}_\Lambda(t) := \mathbf{A}(t)|_\Lambda = (a_{\lambda,\mu}(t))_{\lambda,\mu\in\Lambda}$ the restriction of the original bi-infinite operator $\mathbf{A}(t)$ to a finite index set $\Lambda \subset \mathscr{J}$, $|\Lambda| < \infty$ (and similarly $\mathbf{I}_\Lambda$, $\mathbf{M}^q_\Lambda$). The choice of Λ is done in an adaptive manner, i.e., we get a sequence $\Lambda^{(0)} \to \Lambda^{(1)} \to \Lambda^{(2)} \to \cdots$ by one of the known adaptive wavelet schemes that have been proven to be asymptotically optimal, [11, 12, 24, 37, 45].

16.5 The Hierarchical Tucker Format (HTF)

In this section, we briefly recall the main properties of the *Hierarchical Tucker Format (HTF)* and describe key features of our implementation. We concentrate on those issues needed for the pricing problem under consideration and refer e.g. to [20, 21, 27, 36] for more details. We call $\mathbf{w} \in \mathbb{R}^\mathscr{K}$, $\mathscr{K} = \times_{j\in\mathscr{N}}\mathscr{K}_j$ with entries $\mathbf{w}_\mathbf{i} \in \mathbb{R}$, $\mathbf{i} = (i_0, \ldots, i_{N-1})^T = (i_j)_{j\in\mathscr{N}}$, $i_j \in \mathscr{K}_j$, a *tensor* of *order* N.[2] Note that we will consider the cases $\mathscr{K}_j = \mathscr{I}_j$ (a *vector-tensor*) as well

[2]The indexation here is adapted to our problem at hand and thus differs from the standard literature on the Hierarchical Tucker Format.

as $\mathscr{K}_j = \mathscr{J}_j \times \mathscr{I}_j$ (a *matrix-tensor*), where $\mathscr{I}_j$ and $\mathscr{J}_j$ are suitable (possibly adaptively chosen) index sets. Storing and numerically manipulating tensors exactly is extremely expensive since the amount of storage and work grows exponentially with the order. Hence, one wishes to approximate a tensor $\mathbf{w}$ (or $\mathrm{vec}(\mathbf{w})$, which denotes the vector storage of $\mathbf{w}$ using reverse lexicographical order w.r.t. the indices) by some efficient format. One example is the *Tucker Format*, [44], where one aims at determining an approximation

$$\mathrm{vec}(\mathbf{w}) \approx \mathbf{V}\,\mathrm{vec}(\mathbf{c}) = (V_{N-1} \otimes \cdots \otimes V_0)\,\mathrm{vec}(\mathbf{c}), \quad V_j \in \mathbb{R}^{\mathscr{I}_j \times \mathscr{J}_j},\ j \in \mathscr{N}, \tag{16.12}$$

with the so called *core tensor* $\mathbf{c} \in \mathbb{R}^{\boldsymbol{\mathscr{J}}}$, $\boldsymbol{\mathscr{J}} = \times_{j\in\mathscr{N}} \mathscr{J}_j$. It is known that the *High-Order Singular Value Decomposition (HOSVD)* (see (16.13) below) yields a 'nearly' optimal solution to the approximation problem (16.12) which is also easy to realize numerically, [36]. However, the storage amount for the core tensor $\mathbf{c}$ still grows exponentially with N. This is the reason to consider alternative formats such as the HTF which provides an efficient multilevel format for the core tensor $\mathbf{c}$. In order to be able to describe the HTF, it is useful to introduce the concept of matricization as well as to describe the HOSVD in some more detail. The direction indices $0,\ldots,N-1$ of a tensor $\mathbf{w} \in \mathbb{R}^{\boldsymbol{\mathscr{I}}}$ are also called *modes*. Consider a splitting of the set of all modes $\{0,\ldots,N-1\} = \mathscr{N}$ into disjoint sets, i.e., $\mathscr{N} = t \cup s$, $t = \{t_1,\ldots,t_k\}$, $s = \{s_1,\ldots,s_{N-k}\} = t^{\complement}$, then the *matricization* $\mathbf{w}^{(t)} \in \mathbb{R}^{\mathscr{I}_t \times \mathscr{I}_t^{\complement}}$ of the tensor $\mathbf{w}$ w.r.t. the modes t is defined as follows

$$\mathscr{I}_t := \times_{i=1}^{k} \mathscr{I}_{t_i}, \quad \mathscr{I}_t^{\complement} := \times_{i=1}^{N-k} \mathscr{I}_{s_i}, \quad (\mathbf{w}^{(t)})_{(i_{t_1},\ldots,i_{t_k}),(i_{s_1},\ldots,i_{s_{N-k}})} := \mathbf{w}_{\mathbf{i}}.$$

Note that $\mathrm{vec}(\mathbf{w}) = \mathbf{w}^{(\mathscr{N})}$. A special case is the *μ-matricization* for $\mu \in \mathscr{N}$, where $t = \{\mu\}$ and $\mathscr{I}_\mu^{\complement} = \mathscr{I}_0 \times \cdots \times \mathscr{I}_{\mu-1} \times \mathscr{I}_{\mu+1} \times \cdots \times \mathscr{I}_{N-1}$. We set $r_\mu := \mathrm{rank}(\mathbf{w}^{(\mu)})$ and call $\mathbf{r} = (r_0,\ldots,r_{N-1})^T$ the *rank* of $\mathbf{w}$.

One idea to obtain an approximation $\tilde{\mathbf{w}}$ of $\mathbf{w}$ requiring less storage is a low-rank approximation, i.e., to determine a tensor $\tilde{\mathbf{w}}$ of rank $\tilde{\mathbf{r}}$ with $\tilde{r}_\mu \le r_\mu \le \#\mathscr{I}_\mu$. This can be achieved by a truncated SVD of each $\mathbf{w}^{(\mu)}$ in the sense that $\mathbf{w}^{(\mu)} \approx U_\mu \Sigma_\mu V_\mu^T$, i.e., $U_\mu \in \mathbb{R}^{\mathscr{I}_\mu \times \tilde{r}_\mu}$ contains the most significant $\tilde{r}_\mu$ left singular vectors of $\mathbf{w}^{(\mu)}$. Then

$$\mathrm{vec}(\mathbf{w}) \approx \mathrm{vec}(\tilde{\mathbf{w}}) := (U_{N-1} \otimes \cdots \otimes U_0)\,\mathrm{vec}(\mathbf{c}) \tag{16.13}$$

with the core tensor $\mathrm{vec}(\mathbf{c}) := (U_{N-1}^T \otimes \cdots \otimes U_0^T)\,\mathrm{vec}(\mathbf{w}) \in \mathbb{R}^{\tilde{r}_0 \times \tilde{r}_{N-1}}$ and the *mode frames* U_μ, $\mu \in \mathscr{N}$. The approximation in (16.13) is precisely the HOSVD and this choice of $\mathbf{c}$ can easily be shown to minimize $\|\mathbf{w} - \tilde{\mathbf{w}}\|_2$. Moreover, it holds that $\|\mathbf{w} - \tilde{\mathbf{w}}\|_2 \le \sqrt{N}\,\inf\{\|\mathbf{w} - \mathbf{v}\|_2 : \mathbf{v} \in \mathbb{R}^{\mathscr{I}},\ \mathrm{rank}(\mathbf{v}^{(\mu)}) \le \tilde{r}_\mu, \mu \in \mathscr{N}\}$. The main idea of the HTF is to construct a hierarchy of matricizations and to make use of the arising multilevel structure. It has been shown in [20, Lemma 17] that for

disjoint $t = t_\ell \cup t_r$, $t \subseteq \mathcal{N}$, we have $\operatorname{span}(\mathbf{w}^{(t)}) \subseteq \operatorname{span}(\mathbf{w}^{(t_r)} \otimes \mathbf{w}^{(t_\ell)})$.[3] This result has the following consequence: If we consider $\mathbf{w}^{(t)}$, $\mathbf{w}^{(t_\ell)}$ and $\mathbf{w}^{(t_r)}$ as defined above and denote by U_t, U_{t_ℓ} and U_{t_r} any (column-wise) bases for the corresponding column spaces, then the result ensures the existence of a so called *transfer matrix* $B_t \in \mathbb{R}^{r_{t_\ell} r_{t_r} \times r_t}$ such that $U_t = (U_{t_r} \otimes U_{t_\ell})\, B_t$, where r_t, r_{t_ℓ} and r_{t_r} denote the ranks of the corresponding matricizations. Obviously t_r and t_ℓ provide a subdivision of a mode t. If one recursively applies such a subdivision to $\operatorname{vec}(\mathbf{w}) = U_{\mathcal{N}}$, one obtains a multilevel-type hierarchy. One continues until t_ℓ, t_r become singletons. The corresponding general decomposition is formulated in terms of a so called dimension tree.

Definition 16.2 ([27, Def. 2.3]). A binary tree $\mathcal{T}_{\mathcal{N}}$ with each node represented by a subset of $\mathcal{N}$ is called a *dimension tree* if the root is $\mathcal{N}$, each leaf node is a singleton, and each parent node is the disjoint union of its two children. We denote by $\mathcal{L}(\mathcal{T}_{\mathcal{N}})$ the set of all leave nodes and by $\mathcal{I}(\mathcal{T}_{\mathcal{N}}) := \mathcal{T}_N \setminus \mathcal{L}(\mathcal{T}_N)$ the set of inner nodes.

Now, we define those tensors that can exactly be represented in HTF, called Hierarchical Tucker Tensor (HTT).

Definition 16.3. Let $\mathcal{T}_{\mathcal{N}}$ be a dimension tree and let $\mathbf{r} := (r_t)_{t \in \mathcal{T}_{\mathcal{N}}} \in \mathbb{N}^{\mathcal{T}_{\mathcal{N}}}$, $r_t \in \mathbb{N}$ be a family of non-negative integers. A tensor $\mathbf{w} \in \mathbb{R}^{\mathcal{I}}$, $\mathcal{I} = \times_{j \in \mathcal{N}} \mathcal{I}_j$, is called *Hierarchical Tucker Tensor (HTT)* of rank $\mathbf{r}$ if there exist families (i) $\mathbf{U} := (U_t)_{t \in \mathcal{L}(\mathcal{T}_{\mathcal{N}})}$ of matrices $U_t \in \mathbb{R}^{\mathcal{I}_t \times r_t}$, $\operatorname{rank}(U_t) \le r_t$ (a nested frame tree), (ii) $\mathbf{B} := (B_t)_{t \in \mathcal{I}(\mathcal{T}_{\mathcal{N}})}$ of matrices (the transfer tensors), such that $\operatorname{vec}(\mathbf{w}) = U_{\mathcal{N}}$ and for each inner node $t \in \mathcal{I}(\mathcal{T}_{\mathcal{N}})$ with children t_ℓ, t_r it holds that $U_t = (U_{t_r} \otimes U_{t_\ell}) B_t$ with $B_t \in \mathbb{R}^{r_{t_\ell} r_{t_r} \times r_t}$.

In order to keep track of the dependencies, we will write $\mathbf{w} = (\mathcal{T}^{\mathbf{w}}_{\mathcal{N}}, \mathbf{r}^{\mathbf{w}}, \mathbf{U}^{\mathbf{w}}, \mathbf{B}^{\mathbf{w}})$.

Remark 16.2. (a) By Definition 16.3, we get a Tucker representation $\operatorname{vec}(\mathbf{w}) = (U_{\{N-1\}} \otimes \cdots \otimes U_{\{0\}}) \operatorname{vec}(\mathbf{c})$ with $\operatorname{vec}(\mathbf{c})$ formed as a multilevel product of the transfer tensors.

(b) For the numerical realization, it is useful to consider different representations of tensors in terms of different matricizations. This can be seen as different views on the tensor with the same data: (i) The 'standard' view, i.e., $B_t \in \mathbb{R}^{k_{t_\ell} k_{t_r} \times k_t}$; (ii) The 'tensor view', $\mathscr{B}_t \in \mathbb{R}^{k_t \times k_{t_\ell} \times k_{t_r}}$, where $B_t = \mathscr{B}_t^{(\{2,3\})}$; (c) The storage of an HTT thus requires N matrices $U_\mu \in \mathbb{R}^{\mathcal{I}_\mu \times k_\mu}$, $\mu \in \mathcal{N}$ and $(N-1) = \#\mathcal{I}(\mathcal{T}_{\mathcal{N}})$ transfer tensors B_t, $t \in \mathcal{I}(\mathcal{T}_{\mathcal{N}})$; (d) From (i) it follows that $\operatorname{rank}(\mathbf{w}^{(\mu)}) \le r_\mu$; (e) An HTT for a vector $\mathbf{w} \in \mathbb{R}^{\mathcal{I}}$ is called *HT-vector* and for a matrix $\mathbf{a} \in \mathbb{R}^{\mathcal{I} \times \mathcal{I}}$ *HT-matrix*.

Computing with hierarchical tensors. We are now going to describe some of the algebraic operations for HTT's that are required for the numerical realization of an adaptive wavelet method. Some issues are similar to existing software [27], some

[3] By $\operatorname{span}(A)$ we denote the linear span of the column vectors of a matrix A.

are specific due to our wavelet discretization. Our numerical realization is described in detail in [36], where also the source code is available.

Lemma 16.1 ([20]). *Let* $\mathbf{v} = (\mathscr{T}_\mathscr{N}, \mathbf{r}^\mathbf{v}, \mathbf{U}^\mathbf{v}, \mathbf{B}^\mathbf{v})$ *and* $\mathbf{w} = (\mathscr{T}_\mathscr{N}, \mathbf{r}^\mathbf{w}, \mathbf{U}^\mathbf{w}, \mathbf{B}^\mathbf{w})$ *be HTT's of order N w.r.t. the same* $\mathscr{T}_\mathscr{N}$. *Then,* $\mathbf{v}+\mathbf{w} = (\mathscr{T}_\mathscr{N}, \mathbf{r}^\mathbf{v} + \mathbf{r}^\mathbf{w}, \mathbf{U}^{\mathbf{v}+\mathbf{w}}, \mathbf{B}^{\mathbf{v}+\mathbf{w}})$, *where* $U_t^{\mathbf{v}+\mathbf{w}} = (U_t^\mathbf{v} U_t^\mathbf{w}) \in \mathbb{R}^{\mathscr{I}_t^\mathbf{v} \cup \mathscr{I}_t^\mathbf{w} \times (r_t^\mathbf{v} + r_t^\mathbf{w})}$ *and* $\mathscr{B}_t^{\mathbf{v}+\mathbf{w}} \in \mathbb{R}^{(r_t^\mathbf{v}+r_t^\mathbf{w})\times(r_{t_\ell}^\mathbf{v}+r_{t_\ell}^\mathbf{w})\times(r_{t_r}^\mathbf{v}+r_{t_r}^\mathbf{w})}$ *for* $t = t_\ell \cup t_r$ *is given by*

$$(\mathscr{B}_t^{\mathbf{v}+\mathbf{w}})_{i,j,k} = \begin{cases} (\mathscr{B}_t^\mathbf{v})_{i,j,k},\ 1 \le i \le r_t^\mathbf{v}, 1 \le j \le r_{t_\ell}^\mathbf{v}, 1 \le k \le r_{t_r}^\mathbf{v}, \\ (\mathscr{B}_t^\mathbf{w})_{i,j,k},\ r_t^\mathbf{v} < i \le r_t^\mathbf{v}+r_t^\mathbf{w}, r_{t_\ell}^\mathbf{v} < j \le r_{t_\ell}^\mathbf{v}+r_{t_\ell}^\mathbf{w}, r_{t_r}^\mathbf{v} < k \le r_{t_r}^\mathbf{v}+r_{t_r}^\mathbf{w}, \\ 0, \qquad \text{otherwise}, \end{cases}$$

for $t \in \mathscr{I}(\mathscr{T}_\mathscr{N}) \setminus \mathscr{N}$ *and at the root node* $t = \mathscr{N}$

$$(\mathscr{B}_\mathscr{N}^{\mathbf{v}+\mathbf{w}})_{1,j,k} = \begin{cases} (\mathscr{B}_\mathscr{N}^\mathbf{v})_{1,j,k},\ 1 \le j \le r_{t_\ell}^\mathbf{v}, 1 \le k \le r_{t_r}^\mathbf{v}, \\ (\mathscr{B}_\mathscr{N}^\mathbf{w})_{1,j,k},\ r_{t_\ell}^\mathbf{v} < j \le r_{t_\ell}^\mathbf{v} + r_{t_\ell}^\mathbf{w}, r_{t_r}^\mathbf{v} \le k < r_{t_r}^\mathbf{v} + r_{t_r}^\mathbf{w}, \\ 0, \qquad \text{otherwise}. \end{cases}$$

It is particularly worth mentioning that the HTF of the sum of two HTF only requires a reorganization of the data and no additional computational work. On the other hand, however, we see that the rank of the sum is the sum of the ranks. We will come back to that point later. Let us now consider the matrix-vector multiplication.

Lemma 16.2 ([27]). *Let* $\mathbf{A} = (\mathscr{T}_\mathscr{N}, \mathbf{r}^\mathbf{A}, \mathbf{U}^\mathbf{A}, \mathbf{B}^\mathbf{A}) \in \mathbb{R}^{\mathscr{I}\times\mathscr{I}}$ *be a matrix HTT and* $\mathbf{w} = (\mathscr{T}_\mathscr{N}, \mathbf{r}^\mathbf{w}, \mathbf{U}^\mathbf{w}, \mathbf{B}^\mathbf{w}) \in \mathbb{R}^\mathscr{I}$ *be a vector HTT w.r.t. the same dimension tree* $\mathscr{T}_\mathscr{N}$. *Then, the matrix-vector product reads* $\mathbf{Aw} = (\mathscr{T}_\mathscr{N}, \mathbf{r}^\mathbf{Aw}, \mathbf{U}^\mathbf{Aw}, \mathbf{B}^\mathbf{Aw})$, *where*

- $r_t^\mathbf{Aw} = r_t^\mathbf{A} r_t^\mathbf{w}$, $t \in \mathscr{T}_\mathscr{N}$; $B_t^\mathbf{Aw} = B_t^\mathbf{A} \otimes B_t^\mathbf{w}$, $t \in \mathscr{I}(\mathscr{T}_N)$;
- $V_t^{(i)} \in \mathbb{R}^{\mathscr{I}_t \times \mathscr{I}_t}$ *is chosen such that* $(U_t^\mathbf{A})_i = vec(V_t^{(i)})$ *for* $t \in \mathscr{L}(\mathscr{T}_N)$ *(reinterpretation of the columns of the leaf bases as matrices);*
- $U_t^\mathbf{Aw} = [V_t^{(1)} U_t^\mathbf{w}, \ldots, V_t^{(r_t^\mathbf{A})} U_t^\mathbf{w}] \in \mathbb{R}^{\mathscr{I}_t \times r_t^\mathbf{A} r_t^\mathbf{w}}$.

Again, the computational work is almost negligible. Note again, that the rank grows and is the product of the two original ranks. It should be noted that the HT-matrix $\mathbf{A}$ has to be represented in the *same* hierarchical order as the HT-vector $\mathbf{w}$ (i.e., w.r.t. the same dimension tree). This might require a conversion of a given matrix into the HTF given by $\mathbf{w}$, see Fig. 16.3. If we can efficiently realize a conversion from $\mathbb{R}^{\boldsymbol{\mathscr{I}}\times\boldsymbol{\mathscr{I}}}$ to $\mathbb{R}^{\boldsymbol{\mathscr{I}}\cdot\boldsymbol{\mathscr{I}}}$, where $\boldsymbol{\mathscr{I}}\cdot\boldsymbol{\mathscr{I}} := \times_{j\in\mathscr{N}}(\mathscr{I}_j \times \mathscr{I}_j)$, then we can use a standard matrix-vector-multiplication for each $j \in \mathscr{N}$. Such a transformation can be done by a reverse lexicographical ordering of the indices, i.e., applying the vec-routine to a matrix-tensor.

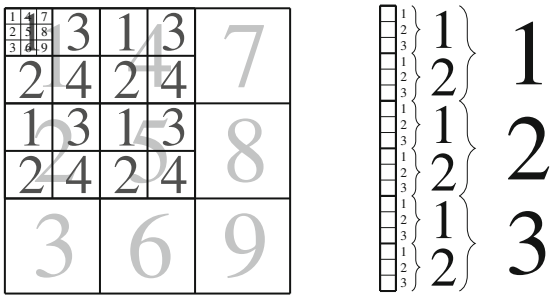


Fig. 16.3 Conversion of a matrix. The *left picture* shows the hierarchical order of the matrix blocks which have to be converted into vectors, the *right picture* the hierarchical order of a vector that is multiplied with the matrix

Lemma 16.3. *Let* $\mathbf{A} = (\mathscr{T}_{\mathscr{N}}, \mathbf{r}^{\mathbf{A}}, \mathbf{U}^{\mathbf{A}}, \mathbf{B}^{\mathbf{A}}) \in \mathbb{R}^{\mathscr{I} \times \mathscr{I}}$ *be a square HTT-matrix, then* $\mathbf{D}_{\mathbf{A}} = diag(\mathbf{A}) \in \mathbb{R}^{\mathscr{I}}$ *is given as* $\mathbf{D}_{\mathbf{A}} = (\mathscr{T}_{\mathscr{N}}, \mathbf{r}^{\mathbf{A}}, \mathbf{U}^{\mathbf{D}_{\mathbf{A}}}, \mathbf{B}^{\mathbf{A}})$*, where* $U_t^{\mathbf{D}_{\mathbf{A}}} \in \mathbb{R}^{\mathscr{I}_t \times r_t^{\mathbf{A}}}$ *is given as* $U_t^{\mathbf{A}} \in \mathbb{R}^{(\mathscr{I}_t \times \mathscr{I}_t) \times r_t^{\mathbf{A}}}$*,* $t \in \mathscr{T}_{\mathscr{N}}$*, by* $(U_t^{\mathbf{D}_{\mathbf{A}}})_{i,k} := (U_t^{\mathbf{A}})_{(i,i),k}$*,* $k = 1, \ldots, r_t^{\mathbf{A}}$*,* $i \in \mathscr{I}_t$*.*

Lemma 16.4. *Let* $\mathscr{A} = \sum_{k=1}^{m} \times_{\mu \in \mathscr{N}} A_{k,\mu} \in \mathbb{R}^{\mathscr{K}}$ *be a Kronecker tensor, then* $\mathscr{A} = (\mathscr{T}_{\mathscr{N}}, \mathbf{r}^{\mathscr{A}}, \mathbf{U}^{\mathscr{A}}, \mathbf{B}^{\mathscr{A}})$*, where (i)* $U_\mu^{\mathscr{A}} := A_{k,\mu}$*,* $k = 1, \ldots, m$*,* $\mu \in \mathscr{L}(\mathscr{T}_{\mathscr{N}})$*,* $r_\mu^{\mathscr{A}} := k$*; (ii) for* $t \in \mathscr{I}(\mathscr{T}_{\mathscr{N}}) \setminus \{\mathscr{N}\}$*:* $B_t^{\mathscr{A}} \in \mathbb{R}^{m \times m \times m}$*,* $r_t^{\mathscr{A}} := m$*,* $(B_t^{\mathscr{A}})_{i,j,\ell} := \delta_{i=j=\ell}$*; (iii)* $(B_{\mathscr{N}}^{\mathscr{A}})_{i,j,\ell} = \delta_{j,\ell}$*,* $B_{\mathscr{N}}^{\mathscr{A}} \in \mathbb{R}^{1 \times k \times k}$*,* $r_{\mathscr{N}}^{\mathscr{A}} := 1$*.*

Remark 16.3. Since the Kronecker sum in Definition 16.7 is a special case of a Kronecker tensor, Lemma 16.4 also provides an HTF for Kronecker sums.

Truncation of Tensors. We have seen that vector-vector addition and matrix-vector multiplication can be efficiently done for tensors in HTF. However, by Lemmata 16.1 and 16.2 the hierarchical rank grows with each addition or multiplication, so that only a certain number of such operations can be done in a numerical (iterative) scheme until the resulting HTTs get too large to be handled efficiently. Thus, a truncation is required. The basic idea is to apply a singular value decomposition (SVD) on the matricizations $\mathbf{w}^{(t)}$ of the tensor $\mathbf{w}$ and restrict these to the dominant singular values. This can be realized without setting up $\mathbf{w}^{(t)}$ explicitly. Since, by construction, the columns of the mode frames U_t contain a basis for the column span of $\mathbf{w}^{(t)}$ there is a matrix $V_t \in \mathbb{R}^{\mathscr{I}^{(t)} \times r_t}$ such that $\mathbf{w}^{(t)} = U_t V_t^T$. Only the left singular vectors of the SVD of $\mathbf{w}^{(t)}$ are thus needed. Thus, the symmetric singular value decomposition of $\mathbf{w}^{(t)}(\mathbf{w}^{(t)})^T = U_t V_t^T V_t U_t^T =: U_t G_t U_t^T$ yields the same result. The matrices $G_t := V_t^T V_t \in \mathbb{R}^{r_t \times r_t}$ are called *reduced Gramian*. They are always of small size and can be computed recursively within the tree structure. In [20, Lemma 4.6] it is shown, that the reduced Gramians correspond to the accumulated transfer tensors for orthogonal HTTs. This statement also holds for general HTTs, see also [27].

The truncation of an HTT can then be computed by the computation of the QR decomposition $U_t = Q_t R_t$ for $t \in \mathscr{L}(T_d)$ or $(\hat{S}_{t_l}^T R_{t_l} \otimes \hat{S}_{t_r}^T R_{t_r}) B_t = Q_t R_t$ for $t \in \mathscr{I}(T_d)$. Subsequently, the symmetric eigenvalue decomposition of $R_t G_t R_t^T = S_t \Sigma^2 S_t^T$ is computed and the truncated matrix is then given by $U_t = Q_t \widehat{S}_t$,

$t \in \mathscr{L}(T_d)$, or $B_t = Q_t \widehat{S}_t$, $t \in \mathscr{I}(T_d)$, where $\hat{S}_t$ is a restriction of S_t to the first r_t columns. Finally, we recall a well-known estimate for the truncation error.

Lemma 16.5. *Let* $\mathbf{w} = (\mathscr{T}_\mathscr{N}, \mathbf{r}, \mathbf{U}, \mathbf{B})$ *be an HTT and let* $\tilde{\mathbf{w}} = (\mathscr{T}_\mathscr{N}, \tilde{\mathbf{r}}, \tilde{\mathbf{U}}, \tilde{\mathbf{B}})$ *be the truncation of* $\mathbf{w}$ *such that* $\operatorname{rank}(\tilde{\mathbf{w}}^{(t)}) = \tilde{r}_t \leq r_t$. *Then, for* $\mathscr{I}_t = \{1, \ldots, n_t\}$, *we have* $\|\mathbf{w} - \tilde{\mathbf{w}}\|_2 \leq \left(\sum_{i=r_t+1}^{n_t} \sigma_i^2\right)^{1/2} \leq \sqrt{2d-3}\, \inf_{\mathbf{v} \in \mathscr{H}(\tilde{\mathbf{r}})} \|\mathbf{w} - \mathbf{v}\|_2$, *where* $\mathscr{H}(\mathbf{r}) := \{\mathbf{v} = (\mathscr{T}_\mathscr{N}, \mathbf{r}, \mathbf{U}, \mathbf{B}) : \operatorname{rank}(\mathbf{v}^{(t)}) \leq r_t, t \in \mathscr{T}_\mathscr{N}\}$ *and* σ_i *are the sigular values of* $\mathbf{w}^{(t)}$ *such that* $\sigma_i \geq \sigma_j$ *if* $i < j$, $i = 1, \ldots, n_t$.

Together with the vector-vector and matrix-vector addition as well as the truncation, the linear solver can be realized, [3], in particular in an adaptive setting, [2]. The main difference is that after each addition or multiplication a truncation has to be made in order to keep the hierarchical rank small, which is not always easy to realize, [27]. For the HTF a Matlab implementation is available, [27]. We have developed an HT library in C++ in [36] based on BLAS, [9,14,15,28] and LAPACK [1] routines, which are efficiently accessed via the FLENS interface, [29, 30]. The reason for our implementation is that FLENS is the basis of the LAWA library (Library for Adaptive Wavelet Applications), [40]. The coupling of the proposed adaptive wavelet scheme with the HT structure could thus be efficiently realized. All subsequent numerical experiments have been performed with this software.

16.6 Numerical Experiments

We report on numerical experiments, first (in order to describe some fundamental mechanisms) for a simple CDO and then in a more realistic framework.

16.6.1 Encoding of Defaults

We are given n assets and hence the state dimension is $N = 2^n$. Let $j \in \mathscr{N}$ and let $\mathtt{j} \in \mathtt{N} := \{0,1\}^n$ be its binary representation, i.e., a binary vector $(\mathtt{j}_1, \ldots, \mathtt{j}_n)^T$ of length n, where $\mathtt{j}_i = 1$ means that asset number i is defaulted. For $j, k \in \mathscr{N}$ with binary representation $\mathtt{j}, \mathtt{k} \in \mathtt{N}$ and $\mathtt{j}|\mathtt{k}$ denoting the bitwise XOR, the number of ones in $\mathtt{j}|\mathtt{k}$ corresponds to the number of assets that change their state. This easy encoding is the reason why we used the labeling $\mathscr{N} = \{0, \ldots, N-1\}$.

Once defaulted, always defaulted. For our numerical experiments, we assume for simplicity that an asset that is defaulted, stays defaulted for all future times, it cannot be reactivated (the theory and our implementation, however, is not restricted to this case). This means that $q^{j,k}(t,y) = 0$ if there exists an index $1 \leq i \leq n$ such that $\mathtt{j}_i = 1, \mathtt{k}_i = 0$. Both in the usual and in the binary ordering this last statement means $q^{j,k}(t,y) = 0$ if $j > k$, which in turns means that the

$\mathbf{Q}(t,y) := (q^{j,k}(t,y))_{j,k\in\mathcal{N}}$ is an upper triangular matrix. Moreover, recall that $q^{j,j}(t,y) = -\sum_{k\in\mathcal{N},k>j} q^{j,k}(t,y)$, so that $\mathbf{Q}$ can be stored as a strict lower triangular matrix, i.e., $\mathbf{Q} = (q^{j,k})_{k>j} \in \mathbb{R}^{\mathcal{N}\times\mathcal{N}}$.

Independent defaults. We assume that we have independent defaults. If defaults are independent, the transition of asset i from one state to another is independent of the state for all other assets as long as their states remain unchanged. Before we are going to formalize this, the following example may be helpful for the understanding.

Example 16.1. If a portfolio has 3 assets, asset 2 defaults when changing from $0 = (000)_2$ to $2 = (010)_2$, from $1 = (001)_2$ to $3 = (011)_2$, from $4 = (100)_2$ to $6 = (110)_2$ and from $5 = (101)_2$ to $7 = (111)_2$. These are all transitions where only asset 2 defaults. In case of independent defaults it follows that $q^{0,2} = q^{1,3} = q^{4,6} = q^{5,7}$. Note that $0|2 = 1|3 = 4|6 = 5|7 = (010)_2$.

Let $j,k \in \mathcal{N}$, then $\mathtt{j}|\mathtt{k}$ indicates a state change of asset i if the i-th component of $\mathtt{j}|\mathtt{k}$ is one. Hence, if $\mathtt{j}_1|\mathtt{k}_1 = \mathtt{j}_2|\mathtt{k}_2$, then the same assets change their state. Since the change $1 \to 0$ is not allowed, we obtain that $\mathtt{j}_1|\mathtt{k}_1 = \mathtt{j}_2|\mathtt{k}_2 \to q^{j_1,k_1} = q^{j_2,k_2}$.

Only one default at a time. If only one asset can default at a time, the transition intensity $q^{j,k}$ is zero if $\mathtt{j}|\mathtt{k}$ has more than one "1". On the other hand, if $\mathtt{j}|\mathtt{k}$ has only one "1", then $\mathtt{j}|\mathtt{k}$ must be a power of 2. Since $q^{j,k}(t,y) = 0$ for $j > k$, it suffices to consider the case $k > j$ ($q^{j,j}$ is determined by the condition on the sum over all intensities). For $k > j$ being $\mathtt{j}|\mathtt{k}$ a power of 2 means that $\log_2(k-j) \in \mathbb{N}$. In this case, we have $\mathtt{j}|\mathtt{k} = 0|(\mathtt{k}|\mathtt{j})$, so that for $j,k \in \mathcal{N}$ we have $q^{j,k}(t,y) = q^{0,k|j}(t,y)$ if $k > j$ and $\log_2(k-j) \in \mathbb{N}$ and 0 otherwise.

16.6.2 A Model Problem

The idea of our first numerical example is to showcase the numerical manageability, where the focus is on the combination of the multiwavelet components and the high dimensional Markov chain components. To this end, we start with a simplified CDO:

- The macroeconomic process Y is one dimensional with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ that are constant in time. This implies that $\mathbf{B}(t) \equiv \mathbf{B}$.
- The interest rate $r(t) \equiv r$ is constant in time and does not depend on Y.
- The state dependent parameters $q^{i,j}(t,y)$ and $c^j(t,y)$ are constant in time and do not depend on Y, i.e., $h_q(y) \equiv 1$ and $h_c(y) \equiv 1$, hence $q^{j,k}(t,y) \equiv \tilde{q}^{j,k}$ and $c^j(t,y) \equiv \tilde{c}^j$. This means that $\mathbf{Q}(t,y) \equiv \mathbf{Q} = \tilde{\mathbf{Q}}$, where $\tilde{\mathbf{Q}} = (\tilde{q}^{j,k})_{k>j;j,k\in\mathcal{N}}$.
- There is no recovery and no final payments, i.e., $a^{j,k}(t,y) \equiv 0$ for all $j,k \in \mathcal{N}$ and $a^j(y) = 0$ for $j \in \mathcal{N}$.
- There is only one tranche covering the entire CDO portfolio.

Thus all involved matrices are time-independent. In particular, we have $\mathbf{C}^{j,k}(t) \equiv \gamma_k^j ((\psi_\lambda, \psi_\mu)_{0;\Omega})_{\lambda,\mu\in\mathscr{J}} = \gamma_k^j \mathbf{I}_{\mathscr{J}}$. Moreover $(\mathbf{M}^q)_{\lambda,\mu} = (h_q\psi_\lambda, \psi_\mu)_{0;\Omega} = \delta_{\lambda,\mu}$ and $\mathbf{D}(t) \equiv \mathbf{D} = (d^{j,k})_{j,k\in\mathscr{N}}$. Next, we have by $a^{j,k}(t,y) \equiv 0$ that $b^j(t) \equiv 0$, $\tilde{c}^j(t) \equiv \tilde{c}^j$, $g_\lambda^1 = g_\lambda^2 = (1,\psi_\lambda)_{0;\Omega}$ ($= 0$ for $|\lambda| > 0$) so that (16.10) simplifies to

$$(\mathbf{I}_{\mathscr{N}} \otimes \mathbf{I}_{\mathscr{J}})\dot{\mathbf{x}}(t) + [(\mathbf{I}_{\mathscr{N}} \otimes [\mathbf{A} + r\mathbf{I}_{\mathscr{J}}]) + (\mathbf{D} \otimes \mathbf{I}_{\mathscr{J}})]\,\mathbf{x}(t) = (-\tilde{\mathbf{c}}) \otimes ((1,\psi_\lambda)_{0;\Omega})_{\lambda\in\mathscr{J}}. \tag{16.14}$$

For later reference, recall that (16.9) in this case implies $d^{k,k} = -\sum_{m\in\mathscr{N}\setminus\{k\}, m>k} \tilde{q}^{k,m}$. In turns, this means that $\mathbf{D} = \tilde{\mathbf{Q}} + \mathrm{diag}(\tilde{\mathbf{Q}}\mathbb{1}_{\mathscr{N}})$, where $\mathbb{1}_{\mathscr{N}} := (1,\ldots,1)^T \in \mathbb{R}^{\mathscr{N}}$. Note that even though $\mathbf{D}$ is time-independent, the huge dimension requires storage as an HT-matrix (in particular, it is impossible to store $\mathbf{D}$ directly). We use a standard implicit θ-scheme for the time-discretization of this Sylvester-type equation, [42]. The Barlets-Steward algorithm [6], is a well-known method for solving such Sylvester equations. It is based on a Schur decomposition. However, we cannot use this method here, since, to the best of our knowledge, there is no algorithm for the QR decomposition of HT-matrices available. Alternatively, an iterative scheme (CG, GMRES or BiCGStab) may be used. We have used BiCGStab as $\mathbf{D}$ is (in general) not symmetric. While generally GMRES yields faster convergence in terms of iteration numbers, it requires more computational steps and therefore, in the context of HT-matrices, more truncations are needed, which is computationally expensive. For systems with small condition numbers, BiCGStab requires only a few iterations, [41]. Using any iterative solver requires matrix-vector multiplications, here of the type $(\mathbf{I}_{\mathscr{N}} \otimes \mathbf{A} + \mathbf{D} \otimes \mathbf{I}_{\mathscr{J}})\mathbf{x}$, where $\mathbf{x} = \mathbf{x}_1 \otimes \mathbf{x}_2$ is also a Kronecker product of the appropriate dimension. Then, we obtain $(\mathbf{I}_{\mathscr{N}} \otimes \mathbf{A} + \mathbf{D} \otimes \mathbf{I}_{\mathscr{J}})(\mathbf{x}_1 \otimes \mathbf{x}_2) = \mathbf{x}_1 \otimes \mathbf{A}\mathbf{x}_2 + \mathbf{D}\mathbf{x}_1 \otimes \mathbf{x}_2$, which can be represented as a Kronecker product of an HT-matrix and a matrix. For details of the implementation, we refer to [36].

Construction of the intensity matrix D. We describe the representation of $\mathbf{D}$ into HTF in case of independent defaults and only one default at a time. Recall that $\mathbf{D} = \tilde{\mathbf{Q}} + \mathrm{diag}(\tilde{\mathbf{Q}}\mathbb{1}_{\mathscr{N}})$ and $\tilde{\mathbf{Q}} = (\tilde{q}_{j,k})_{k>j} \in \mathbb{R}^{\mathscr{N}\times\mathscr{N}}$. Hence, we start by deriving a Kronecker sum representation for $\tilde{\mathbf{Q}}$.

Theorem 16.4. *In case of independent defaults and one default at a time, the matrix $\tilde{\mathbf{Q}} \in \mathbb{R}^{\mathscr{N}\times\mathscr{N}}$ can be written as a Kronecker sum (see Definition 16.7 below)*

$$\tilde{\mathbf{Q}} = \bigoplus_{k=1}^{n} \begin{pmatrix} 0 & q^{0,2^{n-k}} \\ 0 & 0 \end{pmatrix} = \sum_{k=1}^{n} \left\{ \bigotimes_{\ell=1}^{k-1} I_{2\times 2} \otimes \begin{pmatrix} 0 & q^{0,2^{n-k}} \\ 0 & 0 \end{pmatrix} \otimes \bigotimes_{\ell=k+1}^{n} I_{2\times 2} \right\}, \tag{16.15}$$

where $I_{2\times 2} \in \mathbb{R}^{2\times 2}$ denotes the identity matrix of corresponding size.

Proof. The k-th summand of the Kronecker sum on the right-hand side of (16.15) reads $\tilde{\mathbf{Q}}_k := \bigotimes_{\ell=1}^{k-1} I_{2\times 2} \otimes \begin{pmatrix} 0 & q^{0,2^{n-k}} \\ 0 & 0 \end{pmatrix} \otimes \bigotimes_{\ell=k+1}^{n} I_{2\times 2}$. It is readily seen that $\tilde{\mathbf{Q}}_k$

is a matrix having the entries $q^{0,n-k}$ at the positions $(2^{n-k+1}(\nu-1)+\mu,\ 2^{n-k+1}(\nu-1)+\mu+2^{n-k})$, for $\nu = 1,\ldots,2^{k-1}$, $\mu = 1,\ldots,2^{n-k}$, i.e., $2^{k-1}\cdot 2^{n-k} = 2^{n-1}$ entries. Note that 2^{n-1} is the number of all combinations with a state change of asset k. Since $2^{n-k+1}(\nu-1)+\mu \neq 2^{n-k}$ for all possible ν and μ, we obtain that $\mathtt{j}\,|\mathtt{k}$ has exactly one "1" at position $i = n-k$. This, in turns, means that $q^{j,k} = q^{0,n-k}$. This shows that $\tilde{\mathbf{Q}}_k$ contains all transition intensities corresponding to asset k at the right positions. It remains to show that the sum over all $\tilde{\mathbf{Q}}_k$ does not cause overlapping indices. Since $\bigoplus_{\ell=1}^{k-1} I_{2\times 2} \in \mathbb{R}^{2^{k-1}\times 2^{k-1}}$, each $\tilde{\mathbf{Q}}_k$ is a block matrix with blocks at different positions. Thus $\bigoplus_{k=1}^{n} \tilde{\mathbf{Q}}_k$ collects the transition intensities for all assets k.

Having the Kronecker sum (16.15) at hand, the next step is to derive the HTF of $\mathbf{D}$. As we have seen in Lemma 16.4 and Remark 16.3, the HTF of $\tilde{\mathbf{Q}}$ can easily be derived from the Kronecker sum representation. Next, note that $\mathbb{R}^{\mathcal{N}} \ni \mathbb{1}_{\mathcal{N}} = \bigotimes_{k=1}^{n}(1,1)^T$, so that the HTF of $\tilde{\mathbf{Q}}\,\mathbb{1}_{\mathcal{N}}$ can be obtained via Lemma 16.3 and the HTF of $\mathbf{D}$ by Lemma 16.1. Finally, it is easily seen that $\tilde{\mathbf{c}} = \bigoplus_{k=1}^{n}(\tilde{c}_k, 0)^T$ so that the HTF of the right-hand side is easily be derived.

Results. We have used fictional data as marked data is generally not publicly available for CDOs. The fictional CDO portfolio consists of $n = 128$ assets, which is a reasonable size. The macroeconomic process is assumed to be one dimensional. This leads to a system of $N = 2^{128}$ coupled partial differential equations. The maturity is assumed to be $T = 1$ and the time interval is discretized into 20 time steps. For the Galerkin approximation, piecewise cubic L_2-orthonormal Multiwavelets with Dirichlet boundary conditions. We have fixed the lowest and highest level to $j_0 = 2$ and $J = 4$. This turned out to be sufficient in this example due to the smoothness of the right-hand side. For the θ-scheme, the parameter $\theta = 0.5$ has been chosen (Crank-Nicholson). To solve the linear system, BiCGStab has been used. The stopping criterion of the linear solver has been set as a relative error of the L_2-norm of the residual to 10^{-13} and HTTs are truncated to a rank of at most 5. It turned out that a fixed upper bound for the rank is sufficient for this example, for an adaptive strategy, see e.g. [2].

In Fig. 16.4, left, the portfolio value depending on the time t and the macroeconomic state y of the CDO in the first stage, where no firm has defaulted, is shown. Whenever a firm defaults, the Markov chain changes its state and the portfolio value

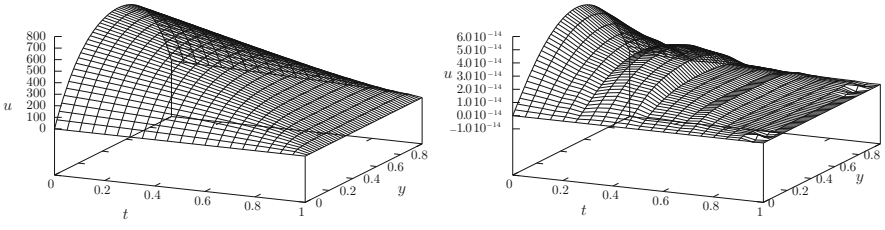


Fig. 16.4 CDO portfolio value in a portfolio of 128 assets. First state (no defaults, *left*) and state (all firms have defaulted, *right*). Note the scaling of the vertical axis in the *right figure*

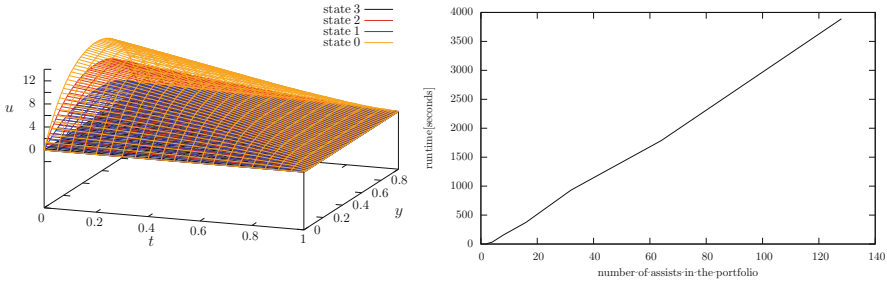


Fig. 16.5 CDO portfolio value of all states in a portfolio of 2 assets (*left*) and runtime of the pricing of a CDO portfolio of $N \in \{2, 4, 8, 16, 32, 64, 128\}$ assets

jumps to a lower value due to the immediate loss of all future continuous payments. To illustrate the jumps of the portfolio value, when a default has occurred, the values of all states of a portfolio of two assets have been combined in Fig. 16.5, left. The last stage is used for error analysis, since here the portfolio value has to be zero as all firms have defaulted. Figure 16.4 shows on the right the portfolio value in the last state of this simulation. This can be interpreted as the relative error arising from the Galerkin approximation with the L_2-orthonormal multiwavelets and the truncation of the HTTs after each addition or multiplication. It can be observed, that the relative error of this computation is smaller than 10^{-13}, which corresponds to the stopping criterion of the linear solver. In each time step, the BiCGStab algorithm took on average 54 iterations.

The computations were performed on a Dell XPS with Intel T9300 Dual Core CPU and 3 GB storage and took about 1 h, the runtime for different portfolio sizes is summarized in Fig. 16.5, right. We observe a linear scaling with respect to n, i.e., only a logarithmic scaling compared to the number of equations $N = 2^n$.

16.6.3 A Realistic Scenario

The example presented in the previous section corresponds to a simplified and idealized situation. As we will show now, the extension to a realistic scenario (given sufficient data), is not too difficult. In fact:

- Observe, that time-dependent parameters do not affect the runtime of the pricing as the condition number of the matrix is not affected, assuming the parameters are sufficiently smooth in the time t.
- The matrices $\mathbf{A}(t)$ and $\mathbf{D}(t)$ need to be setup for any time instant t^k. For $\mathbf{D}(t)$, this only amounts to constructing an HTF exactly as described in the previous section. For the matrix $\mathbf{A}(t)$, we have to compute integrals of the type $(\nabla\psi_\mu, \mathbf{B}(t)\,\nabla\psi_\lambda)_{0;\Omega}$ and $(\boldsymbol{\alpha}(t)^T\,\nabla\psi_\mu, \psi_\lambda)_{0;\Omega}$ for any t^k followed by a transformation to HTF. All these operations can be performed efficiently.

- Space dependent parameters, however, will affect the condition of the linear system and possibly will require the use of a preconditioner. The construction of such a preconditioner in HTF will thus be discussed in the following subsection.

Furthermore, independent defaults are a very restrictive assumption. Other HTFs of dependency structures have to be developed for each given dependency structure. If no explicit HTF can be set up, the transition matrix can as well be approximated using the so-called *Black Box Algorithm*, [4, 18, 36].

Preconditioning. One of the features of wavelet Galerkin schemes is the availability of asymptotically optimal preconditioners, e.g. [45, Ch. 6]. This means that $\mathbf{A}_\Lambda(t)$ from Sect. 16.4 has a uniformly bounded condition number, i.e., $\kappa_2(\mathbf{A}_\Lambda(t)) = \mathscr{O}(1)$ as $|\Lambda| \to \infty$. This, however, does not immediately imply that $\mathscr{A}_\Lambda(t)$ is well-conditioned. We use a simple Jacobi-type preconditioner, i.e.,

$$\mathscr{D}_\Lambda(t) := \operatorname{diag}(\mathscr{A}_\Lambda(t))^{-\frac{1}{2}} = \left[\left(\mathbf{I}_\mathscr{N} \otimes [\mathbf{A}_\Lambda(t) + r(t)\mathbf{I}_\Lambda] + \mathbf{D}(t) \otimes \mathbf{M}_\Lambda^q\right)^{-\frac{1}{2}}_{\lambda,\lambda}\right]_{\lambda\in\Lambda} \tag{16.16}$$

One reason for this choice is the fact that the HTF of such a preconditioner can efficiently be derived. Moreover, the numerical performance has been quite satisfactory, at least in our experiments. The computation of the HTF of a diagonal matrix is provided by Lemma 16.3. The sum in (16.16) can be transformed into HTF by Lemma 16.1 so that we are left with determining the HTF of the four matrices $\mathbf{I}_\mathscr{N}$, $\mathbf{D}_\mathscr{N}$ and $\mathbf{A}_\Lambda$, $\mathbf{M}_\Lambda$. The first two ones are trivial or have been derived above and for the second two ones we are using again the Black Box Algorithm. Finally, also the power $-1/2$ of a tensor can be transformed into HTF by the Black Box Algorithm.

Computing CDO tranches. So far, only single tranche portfolios were considered. However, in practice, a CDO is usually sold in tranches such that the first defaults only affect a certain tranche. Therefore, by construction, this is the riskiest tranche, the so called *equity tranche*. As soon as this first tranche has defaulted completely, subsequent defaults begin to affect a second trance. Therefore, this second tranche, called *mezzanine tranche*, is less risky than the first one. And finally, if this second tranche also has defaulted, the last tranche, the so called *senior tranche* is affected. Sometimes, these three tranches can be further split into sub-tranches. To compute the price of a CDO tranche, the cash flows of (16.1a) have to be adapted to the cash flows which affect the tranche under consideration. The construction of these adapted parameters is described in the following.

Given a portfolio of n assets with nominal values $\pi_1, \ldots, \pi_n$, the S tranches are defined by their upper boundary b_s, $s = 1, \ldots, S$, given in percentages $b_0 = 0 < b_1 < \cdots < b_{S-1} < b_S = 1$ of the total portfolio nominal $\Pi := \sum_{i=1}^n \pi_i$. Let L^j be the accumulated loss in state $j \in \mathscr{N}$, i.e., $L^j = \sum_{i\in\mathscr{D}(j)} \pi_i$, where $\mathscr{D}(j) := \left\{i \in \{1, \ldots, n\} \mid \exists x_k \in \{0,1\}, k = 1, \ldots, n,\ x_i = 1 : j = \sum_{k=1}^n x_k 2^{k-1}\right\}$ denotes the set of all defaulted firms in state j. Then, the cash flows are distributed as follows:

- The amount of the state dependent continuous payments c^j which is assigned to the s-th tranche, $s = 1, \ldots, S$, in state $j \in \mathcal{N}$ can be computed as the percentage of the nominal of the tranche divided by the accumulated nominals of the assets not in default:

$$\tilde{c}_s^j(t) = \begin{cases} \tilde{c}^j(t)\frac{\Pi(b_s - b_{s-1})}{\Pi - L^j} & \text{if } L^j < \Pi b_{s-1} \\ \tilde{c}^j(t)\frac{\Pi(1-b_s) - L^j}{\Pi - L^j} & \text{if } \Pi b_{s-1} \le L^j < \Pi b_s, \\ 0 & \text{otherwise.} \end{cases} \tag{16.17}$$

- The final payments u_T^j are distributed to the tranches in the same way as the continuous payments $\tilde{c}$, i.e., (16.17) holds with $\tilde{c}_s^j(t)$ replaced by $u_{T,s}^j$, the final payment of tranche number s.
- The recovery payments are paid out as a single payment to the tranche in which the default occurred. If several tranches are affected, the recovery is paid out proportional to their nominals. This means

$$\tilde{a}_s^{j,k}(t) = \begin{cases} \tilde{a}^{j,k}(t) & \text{if } \Pi b_{s-1} < L^j < L^k \le \Pi b_s, \\ \tilde{a}^{j,k}(t)\frac{L^k - \Pi b_{s-1}}{L^k - L^j} & \text{if } L^j \le \Pi b_{s-1} < L^k \le \Pi b_s, \\ \tilde{a}^{j,k}(t)\frac{\Pi b_s - \Pi b_{s-1}}{L^k - L^j} & \text{if } L^j \le \Pi b_{s-1} < \Pi b_s < L^k, \\ \tilde{a}^{j,k}(t)\frac{\Pi b_s - L^j}{L^k - L^j} & \text{if } \Pi b_{s-1} < L^j \le \Pi b_s < L^k. \end{cases}$$

  In this setting only recoveries are considered, when a default occurs. However, the model also enables the realization of temporarily defaults of firms. This means, there can also be payments if $L^k < L^j$. These cases are omitted in the following as they can be handled exactly as the cases where $L^j < L^k$.

The difficulty of computing the payoffs of the s-th tranche is that certain specific states within the huge amount of states of the Markov chain have to be found. Therefore, we define vectors $\mathbb{1}_s^<$, $\mathbb{1}_s^=$ and $\mathbb{1}_s^>$, by $(\mathbb{1}_s^<)_j := \chi_{\{L^j < \Pi b_{s-1}\}}$, $(\mathbb{1}_s^=)_j := \chi_{\{\Pi b_{s-1} \le L^j < \Pi b_s\}}$, $(\mathbb{1}_s^>)_j := \chi_{\{L^j \ge \Pi b_s\}}$ for $j \in \mathcal{N}$. In order to explain the realization of some required operations, we introduce the following short-hand notation for HTTs. Let $\mathbf{w} \in \mathbb{R}^{\mathcal{K}}$ be a tensor with HTF $\mathbf{w} = (\mathcal{T}_{\mathcal{N}}, \mathbf{r}, \mathbf{U}, \mathbf{B})$, then we abbreviate by $\mathscr{H}(\mathbf{w})$ its HTF without specifying the quantities involved in the HTF. For $\mathbf{A} \in \mathbb{R}^{\mathcal{K}\times\mathcal{K}}$ and $\mathbf{b} \in \mathbb{R}^{\mathcal{K}}$, we indicate by $\mathscr{H}(\mathbf{A})\,\mathscr{H}(\mathbf{u}) = \mathscr{H}(\mathbf{b})$ that $\mathbf{u} \in \mathbb{R}^{\mathcal{K}}$ is determined as the solution of the linear system $\mathbf{Au} = \mathbf{b}$ but only with numerical routines using the HTF-variants. Finally, we abbreviate $\mathscr{D}(\mathbf{w}) := \mathscr{H}(\mathrm{diag}(\mathbf{w})) \in \mathbb{R}^{\mathcal{K}}$ for $\mathbf{w} \in \mathbb{R}^{\mathcal{K}}$, where $(\mathrm{diag}(\mathbf{w}))_{\mathbf{i},\mathbf{j}} := \delta_{\mathbf{i},\mathbf{j}}\,\mathbf{w}_{\mathbf{i}}$, $\mathbf{i} \in \mathcal{K}$. Now, denote $\mathbf{L} := (L^j)_{j\in\mathcal{N}}$. Then, we need to compute the reciprocal value of each component of the vector $\mathbf{R} := \Pi - \mathbf{L} \in \mathbb{R}^{\mathcal{N}}$ in HTF denoted by $\mathbf{R}^{(-1)}$. This can be achieved by solving the linear system $\mathscr{D}(\mathscr{H}(\mathbf{R}))x = \mathscr{H}(\mathbb{1})$ for x. Then $\mathscr{H}(\mathbf{R}^{(-1)}) := x$ and $\mathscr{H}(\tilde{c}_s(t)) = \mathscr{D}(\mathscr{H}(\tilde{c}(t))\mathscr{D}(\mathscr{H}(\mathbf{R}^{-1}))\mathscr{H}(\mathbb{1}_\mu^<)\Pi(b_s - b_{s-1}) + \mathscr{D}(\mathscr{H}(\tilde{c}(t))\mathscr{D}(\mathscr{H}(\mathbf{R}^{-1}))\mathscr{D}(\mathscr{H}(\mathbb{1}_\mu^=))(\Pi(1 - b_s)\mathscr{H}(\mathbb{1}) - \mathscr{H}(\mathbf{L}))$. We

obtain a similar formula for the final payments $(u^j_{T,s})_{j\in\mathcal{N}}$. The HTF of the matrix $\tilde{\mathbf{A}}_s(t) := (\tilde{a}^{j,k}_s(t))_{j,k\in\mathcal{N}}$ can be set up similarly. Therefore, the matrix $\mathbf{S} := (L^k - L^j)_{j,k\in\mathcal{N}}$ is required. The HTF of this matrix can be obtained by $\mathscr{H}(\mathbf{S}) = \mathscr{H}(\mathbb{1}\cdot\mathbb{1}^T)\mathscr{D}(\mathscr{H}(\mathbf{L})) - \mathscr{D}(\mathscr{H}(\mathbf{L}))\mathscr{H}(\mathbb{1}\cdot\mathbb{1}^T)$. As before, the HTF of the matrix $\mathbf{S}^{(-1)}$ containing the reciprocals of the entries of $\mathbf{S}$ can be found by solving $\mathscr{D}(\mathbf{S})x = \mathscr{H}(\mathbb{1})$ for x and setting $\mathscr{H}(\mathbf{S}^{(-1)}) := x$. Defining $\mathscr{A}(\mathbf{v},\mathbf{w}) := \mathscr{D}(\mathscr{H}(\mathbf{V}))\mathscr{H}((\tilde{a}^{i,j}(t))_{i,j\in\mathcal{N}})\mathscr{D}(\mathscr{H}(\mathbf{w}))$, we obtain

$$\begin{aligned}\mathscr{H}(\tilde{\mathbf{A}}_s(t)) &= \mathscr{A}(\mathbb{1}^{=}_s,\mathbb{1}^{=}_s)\\ &+ \mathscr{D}(\mathscr{A}(\mathbb{1}^{<}_s,\mathbb{1}^{=}_s))\mathscr{D}(\mathscr{H}(\mathbf{S}^{(-1)}))(\mathscr{H}(\mathbb{1}\cdot\mathbb{1}^T)(\mathscr{D}(\mathscr{H}(\mathbf{L})) - \mathscr{D}(\mathscr{H}(\mathbb{1}))\Pi b_{s-1}))\\ &+ \mathscr{D}(\mathscr{A}(\mathbb{1}^{<}_s,\mathbb{1}^{>}_s))\mathscr{D}(\mathscr{H}(\mathbf{S}^{(-1)}))(\mathscr{H}(\mathbb{1}\cdot\mathbb{1}^T)(\mathscr{D}(\mathscr{H}(\mathbb{1}))\Pi(b_s - b_{s-1}))\\ &+ \mathscr{D}(\mathscr{A}(\mathbb{1}^{=}_s,\mathbf{1}^{>}_s))\mathscr{D}(\mathscr{H}(\mathbf{S}^{(-1)}))(\mathscr{H}(\mathbb{1}\cdot\mathbb{1}^T)(\mathscr{D}(\mathscr{H}(\mathbb{1}))\Pi b_s) - \mathscr{D}(\mathscr{H}(\mathbf{L}))).\end{aligned}$$

With these adapted payments, the value of a portfolio tranche can now be determined. A key point of this approach is the construction of the HTF of the vectors $\mathbb{1}^{<}_s$, $\mathbb{1}^{=}_s$ and $\mathbb{1}^{>}_s$. This can be computed by $\mathscr{H}(\mathbb{1}^{<}_s) := \max\{\Pi b_{s-1}\mathscr{H}(\mathbb{1}) - \mathscr{H}(\mathbf{L}), 0\}(\Pi b_{s-1}\mathscr{H}(\mathbb{1}) - \mathscr{H}(\mathbf{L}))^{-1}$, $\mathscr{H}(\mathbb{1}^{>}_s) := \max\{\mathscr{H}(\mathbf{L}) - \Pi b_s\mathscr{H}(\mathbb{1}), 0\}(\mathscr{H}(\mathbf{L}) - \Pi b_s\mathscr{H}(\mathbb{1}))^{-1}$ and $\mathscr{H}(\mathbb{1}^{=}_s) := \mathscr{H}(\mathbb{1}) - \mathscr{H}(\mathbb{1}^{<}_s) - \mathscr{H}(\mathbb{1}^{>}_s)$. The component-wise $\max\{\mathscr{H}(\cdot), 0\}$ of any HT-vector can be determined via $\max\{\mathscr{H}(\cdot), 0\} = \frac{1}{2}(\mathscr{H}(\cdot) + |\mathscr{H}(\cdot)|)$. The absolute value $|\mathscr{H}(\mathbf{w})|$ can be computed by the component-wise Newton iteration $\mathscr{H}(\mathbf{w}^{(n+1)}) = \mathscr{H}(\mathbf{w}^{(n)}) - D((\mathscr{H}(\mathbf{w}^{(n)}))^{-1})\mathscr{D}(\mathscr{H}(\mathbf{w}))\mathscr{H}(\mathbf{w})$. Note, that each iteration step requires the component-wise inversion of an HT-vector, i.e., the solution of a linear system. Let ν be such that $\mathbf{w}^{(\nu)}$ is of the desired accuracy, then $|\mathscr{H}(\mathbf{w})| \approx \mathscr{H}(\mathbf{w}^{(\nu)})$. Essentially, this corresponds to $|x| = \sqrt{x^2}$, the well-known Babylonian method, which converges quadratically for non-zero values. Moreover, for vectors $\mathbf{w}^{(0)}$ with positive entries, it always converges to the positive solution.

16.7 The Kronecker Product

The Kronecker product is a well-known technique when dealing with high dimensional problems, as it often provides the decomposition of a high dimensional problem into a product of problems of low dimension. The following facts can be found in [31, 39]. We note that all subsequent properties can also be extended to (infinite) countable index sets $\mathcal{J}$.

Definition 16.4. The *Kronecker product* $A\otimes B \in \mathbb{R}^{m_A m_B\times n_A n_B}$ of $A\in\mathbb{R}^{m_A\times n_A}$ and $B\in\mathbb{R}^{m_B\times n_B}$ is defined by $(A\otimes B)_{(\mu_1-1)m_B+\mu_2,(\nu_1-1)n_B+\nu_2} := A_{\mu_1,\nu_1}B_{\mu_2,\nu_2}$.

Lemma 16.6. *Let* $A\in\mathbb{R}^{m_A\times n_A}$, $B\in\mathbb{R}^{m_B\times n_B}$, $C\in\mathbb{R}^{m_C\times n_C}$, $D\in\mathbb{R}^{m_D\times n_D}$ *and* $\nu\in\mathbb{R}$. *Then: (1)* $(A\otimes B)^T = A^T\otimes B^T$, *(2)* $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$, *(3)*

$(A \otimes B)(C \otimes D) = AC \otimes BD$, *if* $n_A = m_C$ *and* $n_B = m_D$, *(4)* $A \otimes (B \otimes C) = (A \otimes B) \otimes C$, *(5)* $A \otimes (B + C) = A \otimes B + A \otimes B$ *and* $(B + C) \otimes A = B \otimes A + C \otimes A$, *(6)* $\nu(A \otimes B) = (\nu A) \otimes B = A \otimes (\nu B)$, *(7)* $\mathrm{tr}(A \otimes B) = \mathrm{tr}(A)\mathrm{tr}(B)$, *(8) if* $m_A = n_A$ *and* $m_B = n_B$, *then* $\det(A \otimes B) = (\det(A))^{n_B} (\det(B))^{n_A}$, *(9)* $\mathrm{rank}(A \otimes B) = \mathrm{rank}(A)\mathrm{rank}(B)$.

Definition 16.5. Let $A \in \mathbb{R}^{m_A \times n_A}$. Its *vectorization* is $\mathrm{vec}(A) := (A_{\cdot,1}^T, \ldots, A_{\cdot,n_A}^T)^T$, where $A_{\cdot,i}$, $i \in \{1, \ldots, n_A\}$, is the i-th column of the matrix A.

Lemma 16.7 ([31, (2)]). *Let* $A \in \mathbb{R}^{m_A \times n_A}$, $B \in \mathbb{R}^{m_B \times n_B}$, $C \in \mathbb{R}^{m_B \times m_A}$, $X \in \mathbb{R}^{n_B \times n_A}$, *then* $(A \otimes B)\mathrm{vec}(X) = \mathrm{vec}(C)$ *if and only if* $BXA^T = C$.

Definition 16.6. Let $A_{k,\mu} \in \mathbb{R}^{\mathscr{K}_\mu}$, $\mu \in \mathscr{N}$, $k = 1, \ldots, m$, $m \in \mathbb{N}$. Then, we define the *Kronecker tensor* as $\mathscr{A} := \sum_{k=1}^m \bigotimes_{\mu \in \mathscr{N}} A_{k,\mu} \in \mathbb{R}^{\mathscr{K}}$.

Definition 16.7. The *Kronecker sum* of $A_k \in \mathbb{R}^{n_k \times n_k}$, $k = 1, \ldots, m$, is defined as $\bigoplus_{k=1}^m A_k := \sum_{k=1}^m \left\{ \bigotimes_{\ell=1}^{k-1} I_{n_\ell \times n_\ell} \otimes A_k \otimes \bigotimes_{\ell=k+1}^m I_{n_\ell \times n_\ell} \right\} \in \mathbb{R}^{(n_1 \cdots n_m) \times (n_1 \cdots n_m)}$.

Obviously, the Kronecker sum is a special case of the Kronecker tensor, where $A_{k,\mu} = \delta_{k,\ell} A_k + (1 - \delta_{k,\ell}) I_{n_\ell \times n_\ell}$

References

1. Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., Sorensen, D.: LAPACK Users' Guide, 3rd edn. Society for Industrial and Applied Mathematics, Philadelphia (1999)
2. Bachmayr, M., Dahmen, W.: Adaptive near-optimal rank tensor approximation for high-dimensional operator equations. IGPM Report 363, RWTH Aachen (2013)
3. Ballani, J., Grasedyck, L.: A projection method to solve linear systems in tensor format. Numer. Linear Algebra Appl. **20**(1), 27–43 (2013)
4. Ballani, J., Grasedyck, L., Kluge, M.: Black box approximation of tensors in hierarchical Tucker format. Linear Algebra Appl. **438**(2), 639–657 (2013)
5. Barrault, M., Maday, Y., Nguyen, N., Patera, A.: An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. C.R. Math. **339**(9), 667–672 (2004)
6. Bartels, R.H., Stewart, G.W.: Solution of the matrix equation AX + XB = C [F4]. Commun. ACM **15**(9), 820–826 (1972)
7. Beaudry, P., Lahiri, A.: Risk allocation, debt fueled expansion and financial crisis. Working Paper 15110, National Bureau of Economic Research (2009)
8. Bielecki, T.R., Rutowski, M.: Credit Risk: Modeling, Valuation and Hedging. Springer, Berlin/Heidelberg/New York (2002)
9. Blackford, L.S., Demmel, J., Dongarra, J., Duff, I., Hammarling, S., Henry, G., Heroux, M., Kaufman, L., Lumsdaine, A., Petitet, A., Pozo, R., Remington, K., Whaley, R.C.: An updated set of basic linear algebra subprograms. ACM Trans. Math. Softw. **28**(2), 135–151 (2002)
10. Bluhm, C., Overbeck, L.: Structured Credit Portfolio Analysis, Baskets and CDOs. Chapman and Hall, Boca Raton/London/New York (2007)
11. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods for elliptic operator equations: convergence rates. Math. Comput. **70**(233), 27–75 (2001)
12. Cohen, A., Dahmen, W., DeVore, R.: Adaptive wavelet methods. II. Beyond the elliptic case. Found. Comput. Math. **2**(3), 203–245 (2002)

13. Dijkema, T.J., Schwab, C., Stevenson, R.: An adaptive wavelet method for solving high-dimensional elliptic PDEs. Constr. Approx. **30**, 423–455 (2009)
14. Dongarra, J.J., Du Croz, J., Hammarling, S., Duff, I.S.: A set of level 3 basic linear algebra subprograms. ACM Trans. Math. Softw. **16**(1), 1–17 (1990)
15. Dongarra, J.J., Du Croz, J., Hammarling, S., Hanson, R.J.: An extended set of Fortran basic linear algebra subprograms. ACM Trans. Math. Softw. **14**, 1–17 (1986)
16. Donovan, G., Geronimo, J., Hardin, D.: Orthogonal polynomials and the construction of piecewise polynomial smooth wavelets. SIAM J. Math. Anal. **30**(5), 1029–1056 (1999)
17. Dwyer, G.P.: Financial innovation and the financial crisis of 2007–2008. Papeles de Economia Espana **130** (2012)
18. Epsig, M., Grasedyck, L., Hackbusch, W.: Black box low tensor-rank approximation using fiber-crosses. Constr. Approx. **30**(3), 557–597 (2009)
19. Geronimo, J.S., Hardin, D.P., Massopust, P.R.: Fractal functions and wavelet expansions based on several scaling functions. J. Approx. Theory **78**(3), 373–401 (1994)
20. Grasedyck, L.: Hierarchical singular value decomposition of tensors. SIAM J. Matrix Anal. App. **31**(4), 2029–2054 (2010)
21. Hackbusch, W., Kühn, S.: A new scheme for the tensor representation. J. Fourier Anal. Appl. **15**(5), 706–722 (2009)
22. Hilber, N., Kehtari, S., Schwab, C., Winter, C.: Wavelet finite element method for option pricing in highdimensional diffusion market models. Research Report 01, ETH Zürich (2010)
23. Jarrow, R.A.: The role of ABS, CDS and CDOs in the credit crisis and the economy. Working Paper (2012)
24. Kestler, S.: On the adaptive tensor product wavelet Galerkin method with applications in finance. Ph.D. thesis, Ulm University (2013)
25. Kestler, S., Steih, K., Urban, K.: An efficient space-time adaptive wavelet Galerkin method for time-periodic parabolic partial differential equations. Report 06, Ulm University (2013)
26. Kraft, H., Steffensen, M.: Bankruptcy, counterparty risk, and contagion. Rev. Finance **11**, 209–252 (2006)
27. Kressner, D., Tobler, C.: Algorithm 941: htucker – a Matlab toolbox for tensors in hierarchical Tucker format. ACM Trans. Math. Software **40**(3), Art. 22, 22 pp. (2014)
28. Lawson, C.L., Hanson, R.J., Kincaid, D.R., Krogh, F.T.: Basic linear algebra subprograms for fortran usage. ACM Trans. Math. Softw. **5**(3), 308–323 (1979)
29. Lehn, M.: FLENS – a flexible library for efficient numerical solutions. Ph.D. thesis, Ulm University (2008)
30. Lehn, M., Stippler, A., Urban, K.: FLENS – a flexible library for efficient numerical solutions. In: Proceedings of Equadiff 11, pp. 467–473. Comenius University Press, Bratislava (SK) (2007)
31. van Loan, C.F.: The ubiquitous Kronecker product. J. Comput. Appl. Math. **123**(1–2), 85–100 (2000)
32. Nochetto, R., Siebert, K., Veeser, A.: Theory of adaptive finite element methods: an introduction. In: DeVore, R., Kunoth, A. (eds.) Multiscale, Nonlinear and Adaptive Approximation, pp. 409–542. Springer, Berlin (2009)
33. Nouy, A.: A priori model reduction through proper generalized decomposition for solving time-dependent partial differential equations. Comput. Methods Appl. Mech. Eng. **199**(23–24), 1603–1626 (2010)
34. Pflüger, D.: Spatially adaptive sparse grids for high-dimensional problems. Ph.D. thesis, TU München (2010)
35. Reisinger, C., Wittum, G.: Efficient hierarchical approximation of high-dimensional option pricing problems. SIAM J. Sci. Comput. **29**(1), 440–458 (2007)
36. Rupp, A.J.: High dimensional wavelet methods for structured financial products. Ph.D. thesis, Ulm University (2013)
37. Schwab, C., Stevenson, R.: Space-time adaptive wavelet methods for parabolic evolution problems. Math. Comput. **78**(267), 1293–1318 (2009)

38. Schwab, C., Stevenson, R.P.: Fast evaluation of nonlinear functionals of tensor product wavelet expansions. Numer. Math. **119**(4), 765–786 (2011)
39. Steeb, W.H., Shi, T.K.: Matrix Calculus and Kronecker Product with Applications and C++ Programs. World Scientific, Singapore/River Edge/London (1997)
40. Stippler, A.: LAWA – Library for Adaptive Wavelet Applications (2013). http://lawa.sourceforge.net/index.html
41. Swesty, F.D., Smolarski, D.C., Syalor, P.E.: A comparison of algorithms for the efficient solution of the linear systems arising from multigroup flux-limited diffusion problems. Astrophys. J. **153**, 369–387 (2004)
42. Sylvester, J.J.: Sur l'equations en matrices $px = xq$. C. R. Acad. Sci. Paris **99**, 67–71 (1884)
43. Tavakoli, J.M.: Collateralized Debt Obligations and Structured Finance. Wiley, Hoboken (2003)
44. Tucker, L.R.: Some mathematical notes on three-mode factor analysis. Psychometrika **31**(3), 279–311 (1966)
45. Urban, K.: Wavelet Methods for Elliptic Partial Differential Equations. Oxford University Press, Oxford (2009)
46. Urban, K., Patera, A.T.: A new error bound for reduced basis approximation of parabolic partial differential equations. C. R. Math. **350**(3–4), 203–207 (2012)

Chapter 17
Computational Methods for the Fourier Analysis of Sparse High-Dimensional Functions

Lutz Kämmerer, Stefan Kunis, Ines Melzer, Daniel Potts, and Toni Volkmer

Abstract A straightforward discretisation of high-dimensional problems often leads to a curse of dimensions and thus the use of sparsity has become a popular tool. Efficient algorithms like the fast Fourier transform (FFT) have to be customised to these thinner discretisations and we focus on two major topics regarding the Fourier analysis of high-dimensional functions: We present stable and effective algorithms for the fast evaluation and reconstruction of multivariate trigonometric polynomials with frequencies supported on an index set $\mathcal{I} \subset \mathbb{Z}^d$.

17.1 Introduction

Let $d \in \mathbb{N}$ be the spatial dimension and $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d \simeq [0,1)^d$ denote the torus. We consider multivariate trigonometric polynomials $f : \mathbb{T}^d \to \mathbb{C}$ with Fourier coefficients $\hat{f}_{\mathbf{k}} \in \mathbb{C}$ supported on the frequency index set $\mathcal{I} \subset \mathbb{Z}^d$ of finite cardinality. The evaluation of the trigonometric polynomial

$$f(\mathbf{x}) = \sum_{\mathbf{k} \in \mathcal{I}} \hat{f}_{\mathbf{k}} \, \mathrm{e}^{2\pi \mathrm{i} \mathbf{k} \cdot \mathbf{x}} \tag{17.1}$$

at a sampling set $\mathcal{X} \subset \mathbb{T}^d$ of finite cardinality can be written as the matrix-vector product

$$\mathbf{f} = \mathbf{A}\hat{\mathbf{f}}, \quad \mathbf{f} = (f(\mathbf{x}))_{\mathbf{x} \in \mathcal{X}} \in \mathbb{C}^{|\mathcal{X}|}, \ \hat{\mathbf{f}} = (\hat{f}_{\mathbf{k}})_{\mathbf{k} \in \mathcal{I}} \in \mathbb{C}^{|\mathcal{I}|}, \tag{17.2}$$

L. Kämmerer • D. Potts • T. Volkmer
Technical University of Chemnitz, Reichenhainer Str. 39, 09126 Chemnitz, Germany
e-mail: lutz.kaemmerer@mathematik.tu-chemnitz.de; daniel.potts@mathematik.tu-chemnitz.de; toni.volkmer@mathematik.tu-chemnitz.de

S. Kunis (✉) • I. Melzer
University of Osnabrück, Albrechtstrasse 28a, 49069 Osnabrück, Germany
e-mail: stefan.kunis@uos.de; ines.melzer@uos.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_17

with the Fourier matrix

$$\mathbf{A} = \mathbf{A}(\mathcal{X}, \mathcal{I}) = \left(\mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}}\right)_{\mathbf{x}\in\mathcal{X},\mathbf{k}\in\mathcal{I}} \in \mathbb{C}^{|\mathcal{X}|\times|\mathcal{I}|}.$$

We are interested in the following two problems:

1. Evaluation: given a support $\mathcal{I} \subset \mathbb{Z}^d$, Fourier coefficients $\hat{f}_{\mathbf{k}} \in \mathbb{C}$, $\mathbf{k} \in \mathcal{I}$, and sampling nodes $\mathcal{X} = \{\mathbf{x}_\ell \in \mathbb{T}^d : \ell = 0, \dots, L-1\}$, evaluate the trigonometric polynomial (17.1) efficiently, i.e., compute $\mathbf{f} = \mathbf{A}\hat{\mathbf{f}}$ by means of a fast algorithm.
2. Reconstruction: given a support of Fourier coefficients $\mathcal{I} \subset \mathbb{Z}^d$, construct a set of sampling nodes $\mathcal{X} \subset \mathbb{T}^d$ with small cardinality $L = |\mathcal{X}|$ which allows for the unique and stable reconstruction of all multivariate trigonometric polynomials (17.1) from their sampling values $f(\mathbf{x}_\ell)$. In particular, solve the system of linear equations $\mathbf{A}\hat{\mathbf{f}} \approx \mathbf{f}$.

As an extension to the reconstruction problem, we considered the efficient approximate reconstruction of a smooth function from subspaces of the Wiener algebra by a trigonometric polynomial (17.1), which guarantees a good approximation to the function, cf. [37, 38].

17.2 Evaluation of Multivariate Trigonometric Polynomials

One cornerstone in numerical Fourier analysis is the fast computation of certain trigonometric sums. A straightforward evaluation of the trigonometric polynomial (17.1) at all sampling nodes $\mathcal{X} \subset \mathbb{T}^d$, or equivalently the matrix vector multiplication (17.2), takes a quadratic number $\mathcal{O}(|\mathcal{X}| \cdot |\mathcal{I}|)$ of floating point operations. For equidistant cartesian grids, the well known fast Fourier transform (FFT) reduces this complexity to an almost linear scaling and this has proven an important reason for the success of numerical Fourier analysis in the last century. More recently, the concept of sparse discretisations has gained a lot of attention and we discuss three variants for the evaluation of sparse trigonometric sums subsequently.

17.2.1 Fast Fourier Transform

We consider multivariate trigonometric polynomials with frequencies supported on the full grid, i.e., with Fourier coefficients $\hat{f}_{\mathbf{k}}$ are defined on the full d-dimensional set $\mathcal{I} := \hat{G}_n^d = \mathbb{Z}^d \cap \times_{j=1}^d (-2^{n-1}, 2^{n-1}]$ of refinement $n \in \mathbb{N}$ and bandwidth $N = 2^n$ with the cardinality $|\mathcal{I}| = N^d$. The evaluation of the trigonometric polynomial

$$f(\mathbf{x}) = \sum_{\mathbf{k}\in\hat{G}_n^d} \hat{f}_{\mathbf{k}}\, \mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}} \tag{17.3}$$

at all sampling nodes of an equispaced grid $\mathbf{x} \in \mathcal{X} = (2^{-n}\hat{G}_n^d \bmod \mathbf{1})$, with the cardinality $|\mathcal{X}| = N^d$, requires only $\mathcal{O}(2^{nd}n) = \mathcal{O}(N^d \log N)$ floating point operations by the famous fast Fourier transform (FFT). A well understood generalisation considers an arbitrary sampling set $\mathcal{X} = \{\mathbf{x}_\ell \in \mathbb{T}^d : \ell = 0, \dots, L-1\}$ and leads to the so-called nonequispaced fast Fourier transform (NFFT) which takes $\mathcal{O}(2^{nd}n + |\log \varepsilon|^d L) = \mathcal{O}(N^d \log N + |\log \varepsilon|^d L)$ floating point operations for a target accuracy $\varepsilon > 0$, see e.g. [5, 16, 40, 52, 60] and the references therein. In both cases, already the huge cardinality of the support $\hat{G}_n^d$ of the Fourier coefficients $\hat{f}_\mathbf{k}$ causes immense computational costs for high dimensions d even for moderate refinement n. Hence, we restrict the index set $\mathcal{I}$ to smaller sets.

17.2.2 Hyperbolic Cross FFT

Functions of dominating mixed smoothness can be well approximated by multivariate trigonometric polynomials with frequencies supported on reduced frequency index sets, so called dyadic hyperbolic crosses

$$\mathcal{I} = H_n^d := \bigcup_{\substack{\mathbf{j}\in\mathbb{N}_0^d \\ \|\mathbf{j}\|_1 = n}} \left(\mathbb{Z}^d \cap \bigtimes_{l=1}^d (-2^{j_l-1}, 2^{j_l-1}] \right)$$

of dimension d and refinement n, cf. [58]. Compared to the trigonometric polynomial in (17.3), we strongly reduce the number of used Fourier coefficients $|H_n^d| = \mathcal{O}(2^n n^{d-1}) \ll 2^{nd}$. A natural spatial discretisation of trigonometric polynomials with frequencies supported on the dyadic hyperbolic cross H_n^d is given by the sparse grid

$$\mathcal{X} = S_n^d := \bigcup_{\substack{\mathbf{j}\in\mathbb{N}_0^d \\ \|\mathbf{j}\|_1 = n}} \bigtimes_{l=1}^d 2^{-j_l} (\mathbb{N}_0 \cap [0, 2^{j_l})).$$

The cardinalities of the sparse grid and the dyadic hyperbolic cross are $|S_n^d| = |H_n^d| = \mathcal{O}(2^n n^{d-1})$. Figure 17.1a(left) shows an example for a two-dimensional dyadic hyperbolic cross and Fig. 17.1a(right) depicts the corresponding sparse grid of identical cardinality. Based on [3, 27] there exists a fast algorithm for evaluating the trigonometric polynomial with frequencies supported on the hyperbolic cross H_n^d at all $\mathbf{x} \in S_n^d$ in $\mathcal{O}(2^n n^d)$ floating point operations, called hyperbolic cross fast Fourier transform (HCFFT). A generalisation to sparser index sets, i.e., to index sets for so called energy-norm based hyperbolic crosses, is presented in [22].

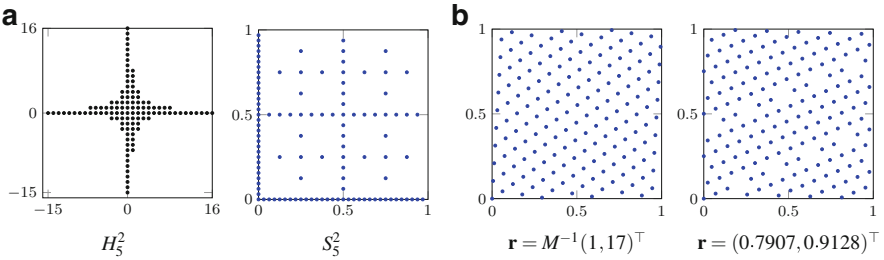


Fig. 17.1 (**a**) Dyadic hyperbolic cross H_5^2 (*left*) and sparse grid S_5^2 (*right*). (**b**) Rank-1 lattice (*left*) and generated set $\Lambda(\mathbf{r}, M)$ (*right*), $M = 163$

17.2.3 Lattice and Generated Set FFT

Using lattices as sampling set $\mathcal{X}$ is motivated from the numerical integration of functions of many variables by lattice rules, see [14, 48, 59] for an introduction. In contrast to general lattices which may be spanned by several vectors, we only consider so-called rank-1 lattices and a generalisation of this concept called generated sets [32]. For a given number $L \in \mathbb{N}$ of sampling nodes and a generating vector $\mathbf{r} \in \mathbb{R}^d$, we define the generated set

$$\mathcal{X} = \Lambda(\mathbf{r}, L) := \{\mathbf{x}_\ell = \ell\mathbf{r} \bmod \mathbf{1},\ \ell = 0, \dots, L-1\} \subset \mathbb{T}^d.$$

For $\ell = 0, \dots, L-1$, the evaluation of a d-variate trigonometric polynomial supported on an arbitrary frequency index set $\mathcal{I}$ simplifies dramatically since

$$f(\mathbf{x}_\ell) = \sum_{\mathbf{k}\in\mathcal{I}} \hat{f}_\mathbf{k}\, \mathrm{e}^{2\pi\mathrm{i}\mathbf{k}\cdot\mathbf{x}_\ell} = \sum_{\mathbf{k}\in\mathcal{I}} \hat{f}_\mathbf{k}\, \mathrm{e}^{2\pi\mathrm{i}\ell\mathbf{k}\cdot\mathbf{r}} = \sum_{y\in\mathcal{Y}} \hat{g}_y\, \mathrm{e}^{2\pi\mathrm{i}\ell y}, \tag{17.4}$$

with some set $\mathcal{Y} = \{\mathbf{k}\cdot\mathbf{r} \bmod 1 : \mathbf{k} \in \mathcal{I}\} \subset \mathbb{T}$ and the aliased coefficients

$$\hat{g}_y = \sum_{\mathbf{k}\cdot\mathbf{r}\equiv y \pmod 1} \hat{f}_\mathbf{k}. \tag{17.5}$$

Using a one-dimensional adjoint NFFT [40], this takes $\mathcal{O}(L\log L + (d + |\log\varepsilon|)|\mathcal{I}|)$ floating point operations for a target accuracy $\varepsilon > 0$. Moreover, given $L \in \mathbb{N}$ and a generating vector $\mathbf{r} = \mathbf{z}/L$, $\mathbf{z} \in \mathbb{Z}^d$, the sampling scheme $\Lambda(\mathbf{r}, L)$ is called rank-1 lattice and the computational costs of the evaluation reduce to $\mathcal{O}(L\log L + d|\mathcal{I}|)$ by applying a one dimensional FFT. We stress on the fact that in both cases, the computational costs only depend on the number L of samples subsequent to the aliasing step (17.5) which takes $d|\mathcal{I}|$ floating point operations. Figure 17.1b(left) and 17.1b(right) show an example for a two-dimensional rank-1 lattice and generated set, respectively.

17.2.4 Butterfly Sparse FFT

Another generalisation of the classical FFT to nonequispaced nodes has been suggested in [1,41,63]. While the above mentioned NFFT still relies on a equispaced FFT, the so-called butterfly scheme only relies on local low rank approximations of the complex exponentials – in particular this locality allows for its application to sparse data. The idea of local low rank approximations can be traced back at least to [4,21,26,64] for smooth kernel functions and to [13,46,49,62,65] for oscillatory kernels. In a linear algebra setting, it was pointed out in [17] that certain blocks of the Fourier matrix are approximately of low rank.

We consider real frequencies $\mathcal{I} \subset [0, 2^n)^d$ and nonequispaced evaluation nodes $\mathbf{x}_\ell \in \mathcal{X} \subset [0, 1)^d$ in

$$f(\mathbf{x}_\ell) = \sum_{\mathbf{k}\in\mathcal{I}} \hat{f}_\mathbf{k}\, \mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}_\ell}, \quad \ell = 0, \dots, L-1. \tag{17.6}$$

For ease of notation, we outline the main idea for the one-dimensional case. We decompose both domains dyadically starting with the whole interval $[0, 2^n)$ and $[0, 1)$ as root, respectively, see also Fig. 17.2(left, middle). Each pair of a frequency interval in the $(n - j)$-th level and a space interval in the j-th level now fulfils the admissibility condition $\mathrm{diam}(\mathcal{I}')\mathrm{diam}(\mathcal{X}') \leq 1$. These pairs are depicted in Fig. 17.2(right), where an edge in this butterfly graph is set if and only if the associated pairs of intervals are connected in both trees.

We note that the properly frequency shifted exponential function is a smooth function within the admissible region and can be well approximated by a trigonometric sum with equispaced frequencies interpolating in Chebyshev nodes, see [41, Thm. 2.6] for details.

The generalisation to spatial dimension $d \geq 2$ is straightforward by decomposing $\mathcal{I} \subset [0, 2^n)^d$ and $\mathcal{X} \subset [0, 1)^d$ dyadically in each coordinate, using a tensor product ansatz, and interpolate in a product grid. The butterfly scheme now traverses the butterfly graph top down. We start in the zeroth level, sum frequencies in the finest

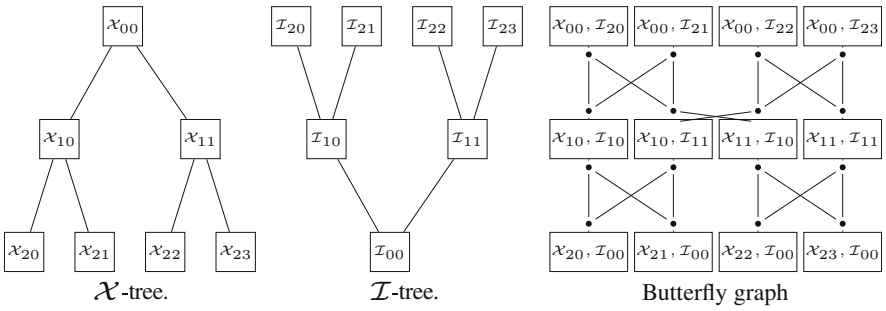


Fig. 17.2 Trees and butterfly graph for $N = 4$

decomposition, and approximate on the whole spatial domain. In each subsequent level, we sum up two predecessors including more frequencies and approximate on each smaller spatial box. The final approximation is a function piecewise defined on the finest spatial decomposition. The butterfly scheme guarantees the following target accuracy.

Theorem 17.1 ([41, Thm. 3.1]). *Let $d, n, p \in \mathbb{N}$, $p \geq 5$, $\mathcal{I} \subset [0, 2^n)^d$, $\mathcal{X} \subset [0,1)^d$, and the trigonometric sum f as in (17.6), then the butterfly approximation g obeys the error estimate*

$$\|f - g\|_\infty \leq \frac{(C_p + 1)(C_p^{d(n+1)} - 1)}{C_p - 1} c_p \|\hat{\mathbf{f}}\|_1.$$

The constants are explicitly given by

$$K_p := \left(\frac{2\pi^2}{\left(1 - \cos\frac{2\pi}{p-1}\right)(p-1)^2} \right)^{p-1}, \quad K_p \leq \frac{\pi^4}{16}, \quad \lim_{p\to\infty} K_p = 1,$$

$$C_p := \sqrt{K_p}\left(1 + \tfrac{2}{\pi}\log p\right), \qquad c_p := \frac{1}{\pi p}\left(\frac{\pi}{p-1}\right)^p.$$

In particular, the butterfly scheme achieves relative error at most ε if the local expansion degree fulfils $p \geq \max\{10, 2|\log\varepsilon|, 2d(n+1)\}$.

In case $1 \leq t < d$ and $|\mathcal{X}| = |\mathcal{I}| = 2^{nt}$ well distributed sets on smooth t-dimensional manifolds, the dyadic decompositions of the sets remain sparse. Consequently, the butterfly graph, which represents the admissible pairs where computations are performed, remains sparse as well and the computation of (17.6) takes $\mathcal{O}(2^{nt} n(n + |\log\varepsilon|)^{d+1})$ floating point operations only.

17.3 Reconstruction Using Multivariate Trigonometric Polynomials

Beyond the fast evaluation of Fourier expansions, the sampling problem is concerned with the recovery of the Fourier coefficients $\hat{f}_{\mathbf{k}} \in \mathbb{C}$, $\mathbf{k} \in \mathcal{I}$, from a sequence of function samples f_ℓ, $\ell = 0, \ldots, L-1$. This inverse transform constructs a trigonometric polynomial f, see (17.1), such that for given data points $(\mathbf{x}_\ell, f_\ell) \in \mathbb{T}^d \times \mathbb{C}$, $\ell = 0, \ldots, L-1$, the approximate identity

$$f(\mathbf{x}_\ell) \approx f_\ell$$

is fulfilled. Thus, we aim to solve the linear system of equations $\mathbf{A}\hat{\mathbf{f}} \approx \mathbf{f}$, i.e., we compute the vector of Fourier coefficients $\hat{\mathbf{f}} = (\hat{f}_{\mathbf{k}})_{\mathbf{k}\in\mathcal{I}} \in \mathbb{C}^{|\mathcal{I}|}$ from the given vector

of function samples $\mathbf{f} = (f_\ell)_{\ell=0,\dots,L-1} \in \mathbb{C}^L$. In contrast to the ordinary Fourier matrix, its generalized analogue $\mathbf{A}$ is in general neither unitary nor square. The meaningful variants of this reconstruction problem include

1. The weighted least squares approximation

$$\|\mathbf{f} - \mathbf{A}\hat{\mathbf{f}}\|_{\mathbf{W}}^2 = \sum_{\ell=0}^{L-1} w_\ell |f_\ell - f(\mathbf{x}_\ell)|^2 \xrightarrow{\hat{\mathbf{f}}} \min, \tag{17.7}$$

   for the over-determined case $|\mathcal{I}| < L = |\mathcal{X}|$, where the weights w_ℓ compensate for clusters in the sampling set,
2. The optimal interpolation problem

$$\|\hat{\mathbf{f}}\|_{\hat{\mathbf{W}}^{-1}}^2 = \sum_{\mathbf{k}\in\mathcal{I}} \frac{|\hat{f}_\mathbf{k}|^2}{\hat{w}_\mathbf{k}} \xrightarrow{\hat{\mathbf{f}}} \min \quad \text{subject to} \quad \mathbf{A}\hat{\mathbf{f}} = \mathbf{f}, \tag{17.8}$$

   for the under-determined case $|\mathcal{I}| > L = |\mathcal{X}|$, where the weights $\hat{w}_\mathbf{k}$ damp high-frequency components, and
3. The sparse recovery problem

$$\|\hat{\mathbf{f}}\|_0 = |\{\mathbf{k} \in \mathcal{I} : \hat{f}_\mathbf{k} \neq 0\}| \xrightarrow{\hat{\mathbf{f}}} \min \quad \text{subject to} \quad \mathbf{A}\hat{\mathbf{f}} = \mathbf{f}, \tag{17.9}$$

   for the under-determined case $|\mathcal{I}| > L = |\mathcal{X}|$.

The main tool in iterative methods to solve these three problems is the use of fast matrix-vector multiplications with the Fourier matrix $\mathbf{A}$ and its adjoint $\mathbf{A}^*$ as well as bounding involved condition numbers uniformly.

In the following subsections, we focus on the reconstruction of a multivariate trigonometric polynomial (17.1) from sampling values using different sampling schemes. Therefor, we consider different types of sampling sets $\mathcal{X}$ as introduced in Sect. 17.2. We discuss necessary and sufficient conditions on the frequency index set $\mathcal{I}$ and sampling set $\mathcal{X}$ such that the unique and stable reconstruction is guaranteed.

17.3.1 FFT and NFFT

Analog to Sect. 17.2.1, we consider multivariate trigonometric polynomials with frequencies supported on the full grid $\mathcal{I} = \hat{G}_n^d$. The reconstruction of the Fourier coefficients $\hat{f}_\mathbf{k}$, $\mathbf{k} \in \hat{G}_n^d$, from sampling values at an equispaced grid $\mathbf{x} \in \mathcal{X} = (2^{-n}\hat{G}_n^d \bmod \mathbf{1})$, see (17.3), can be realized by the inverse fast Fourier transform, since the Fourier matrix $\mathbf{F} := \mathbf{A}(2^{-n}\hat{G}_n^d, \hat{G}_n^d)$ has orthogonal columns, and takes $\mathcal{O}(N^d \log N)$ floating point operations. This is no longer true for the nonequispaced Fourier matrix given by

$$\mathbf{A} := \mathbf{A}(\mathcal{X}, \hat{G}_n^d) = \left(\mathrm{e}^{2\pi \mathrm{i}\mathbf{k}\cdot\mathbf{x}_\ell}\right)_{\ell=0,\dots,L-1,\mathbf{k}\in\hat{G}_n^d}.$$

Here, we use an iterative algorithm since the fast matrix times vector multiplication with the matrix $\mathbf{A}$ and $\mathbf{A}^*$ takes only $\mathcal{O}(2^{nd}n + |\log\varepsilon|^d L)$ floating point operations, see [40]. The conditioning of the reconstruction problems relies on the uniformity of $\mathcal{X}$, measured by the mesh norm and the separation distance

$$\delta := 2\max_{\mathbf{x}\in\mathbb{T}^d}\min_{j=0,\dots,L-1}\operatorname{dist}(\mathbf{x}_j,\mathbf{x}), \qquad q := \min_{j,l=0,\dots,L-1;j\neq l}\operatorname{dist}\left(\mathbf{x}_j,\mathbf{x}_l\right),$$

where $\operatorname{dist}(\mathbf{x},\mathbf{x}_0) := \min_{\mathbf{j}\in\mathbb{Z}^d}\|(\mathbf{x}+\mathbf{j})-\mathbf{x}_0\|_\infty$, respectively.

For the overdetermined case $N^d < L$, it has been proven in [24] that the reconstruction problem (17.7) has a unique solution if $N < (\frac{\pi}{\log 2}\, d\, \delta)^{-1}$. The solution is computed iteratively by means of the conjugate gradient method in [2, 18, 23], where the multilevel Toeplitz structure of $\mathbf{A}^*\mathbf{W}\mathbf{A}$ is used for fast matrix vector multiplications. Slightly more stable with respect to rounding errors is the CGNR method, cf. [6, pp. 288], which iterates the original residual $\mathbf{r}_l = \mathbf{y} - \mathbf{A}\hat{\mathbf{f}}_l$ instead of the residual $\mathbf{A}^*\mathbf{W}\mathbf{r}_l$ of the normal equations. Further analysis of the numerical stability of the least squares approximation (17.7) relies on so-called Marcinkiewicz-Zygmund inequalities which establish norm equivalences between a trigonometric polynomial and its samples, see e.g. [19, 39, 45, 61] and references therein for specific variants.

For the underdetermined case $N^d > L$, the optimal interpolation problem (17.8) has been shown to be stable in [42] if the sampling set is well separated with respect to the polynomial degree and the weights $\hat{w}_\mathbf{k}$ are constructed by means of a so-called smoothness-decay principle. In particular, we proved that the nonequispaced Fourier matrix $\mathbf{A}$ has full rank L for every polynomial degree $N > 2\,d\,q^{-1}$ and proposed to solve problem (17.8) by a version of the conjugate gradient method in combination with the NFFT to efficiently perform each iteration step.

17.3.2 Hyperbolic Cross FFT

For the HCFFT, see Sect. 17.2.2, there also exists a fast inverse algorithm. This inverse HCFFT is not an orthogonal transform and is realized by reverting all steps of the HCFFT, see [3, 27], which makes this spatial discretisation most attractive in terms of efficiency. Therefore, the inverse HCFFT requires also only $\mathcal{O}(2^n n^d)$ floating point operations. However, we proved in [35] that this transform is mildly ill conditioned, since the condition numbers of the Fourier matrices $\mathbf{A}(S_n^d, H_n^d)$ are bounded by

$$\begin{aligned} c_d 2^{\frac{n}{2}} n^{\frac{2d-3}{2}} \leq \operatorname{cond}_2\mathbf{A}(S_n^d, H_n^d) \leq C_d 2^{\frac{n}{2}} n^{2d-2}, \quad n\to\infty, \\ c_n d^{2n} \leq \operatorname{cond}_2\mathbf{A}(S_n^d, H_n^d) \leq C_n d^{2n}, \quad d\to\infty. \end{aligned}$$

In particular, we loose more than 5 decimal digits of accuracy already for $d = 10$ and $n = 5$ in the worst case.

17.3.3 Lattice and Generated Set FFT

As pointed out in Sect. 17.2.3, the evaluation of multivariate trigonometric polynomials with frequencies supported on an arbitrary index set $\mathcal{I}$, i.e., the mapping from the index set $\mathcal{I}$ in frequency domain to the rank-1 lattice in spatial domain reduces to a single one-dimensional FFT and thus can be computed very efficiently and stable. For the inverse transform, mapping the samples of a trigonometric polynomial to its Fourier coefficients on a specific frequency index set, we discuss the recently presented necessary and sufficient conditions on rank-1 lattices allowing a stable reconstruction of trigonometric polynomials with frequencies supported on hyperbolic crosses and the generalisation to arbitrary index sets in the frequency domain. Based on research results in the field of numerical integration [12], we suggest approaches for determining suitable rank-1 lattices using a component–by–component strategy, see [33, 34]. In conjunction with numerically found lattices, we showed that this new method outperforms the classical hyperbolic cross FFT for realistic problem sizes, cf. [36].

The use of generated sets, a generalisation of rank-1 lattices, as spatial discretisations offers an additional suitable possibility for sampling sparse trigonometric polynomials. The fast computation of trigonometric polynomials on generated sets can be realized using the NFFT. A simple sufficient condition on a generated set $\Lambda(\mathbf{r}, L)$ allows the fast, unique and stable reconstruction of the frequencies of a d-dimensional trigonometric polynomial from its samples along $\Lambda(\mathbf{r}, L)$. In contrast to searching for suitable rank-1 lattices, we can use continuous optimization methods in order to determine generated sets that are suitable for reconstruction, see [32].

Reconstruction using rank-1 lattices. In the following, a rank-lattice that allows for the unique reconstruction of all trigonometric polynomials with frequencies supported on the frequency index set $\mathcal{I}$ is called *reconstructing rank-lattice* for $\mathcal{I}$. In order to state constructive existence results for reconstructing rank-1 lattices, we define the difference set

$$\mathcal{D}(\mathcal{I}) := \{\mathbf{k} - \mathbf{l} : \mathbf{k}, \mathbf{l} \in \mathcal{I}\}$$

of the frequency index set $\mathcal{I}$. As a consequence of [34, Cor. 1] we formulate the following

Theorem 17.2. *Let $\mathcal{I} \subset \{\mathbf{k} \in \mathbb{Z}^d : \mathbf{k} - \mathbf{a} \in [0, |\mathcal{I}| - 1]^d\}$ for a fixed $\mathbf{a} \in \mathbb{Z}^d$ being a frequency index set of finite cardinality. Then there exists a reconstructing rank-1 lattice of prime cardinality L,*

$$|\mathcal{I}| \leq L \leq |\mathcal{D}(\mathcal{I})| \leq |\mathcal{I}|^2, \tag{17.10}$$

such that all multivariate trigonometric polynomials f with frequencies supported on $\mathcal{I}$ can be reconstructed from the sampling values $(f(\mathbf{x}))_{\mathbf{x}\in\Lambda(\mathbf{r},L)}$. Moreover, the corresponding generating vector $\mathbf{r} \in L^{-1}\mathbb{Z}^d$ can be determined using a component–by–component strategy and the reconstruction of the Fourier coefficients can be realized by a single one-dimensional FFT of length L, and thus takes $\mathcal{O}(L \log L + d\,|\mathcal{I}|)$ floating point operations.

Proof. The result follows from [34, Cor. 1], Bertrand's postulate, and Eqs. (17.4) and (17.5). □

We stress on the fact, that [34, Cor. 1] is a more general result on arbitrary frequency index sets $\mathcal{I}$. Some simple additional assumptions on L allow to replace the condition $\mathcal{I} \subset \{\mathbf{k} \in \mathbb{Z}^d : \mathbf{k} - \mathbf{a} \in [0, |\mathcal{I}| - 1]^d\}$ by $\mathcal{I} \subset \mathbb{Z}^d$, $|\mathcal{I}| < \infty$.

In fact, the cardinality of the difference set $\mathcal{D}(\mathcal{I})$ is the theoretical upper bound in (17.10) for the number of samples needed to reconstruct trigonometric polynomials with frequencies supported on the index set $\mathcal{I}$ using a rank-1 lattice. This cardinality depends mainly on the structure of $\mathcal{I}$.

Example 17.1. Let $\mathcal{I} = I^d_{p,N} := \{\mathbf{k} \in \mathbb{Z}^d : \|\mathbf{k}\|_p \leq N\}$, $N \in \mathbb{N}$, be the ℓ_p-ball, $0 < p \leq \infty$, of size N, see Fig. 17.3. The cardinality of $I^d_{p,N}$ is bounded by $c_{p,d} N^d \leq |I^d_{p,N}| \leq C_d N^d$ and $c_{p,d} N^d \leq \mathcal{D}(I^d_{p,N}) \leq C_d 2^d N^d$, $c_{p,d}, C_d \in \mathbb{R}$, $0 < c_{p,d} \leq C_d$. Consequently, we can find a reconstructing rank-1 lattice of size $L \leq \tilde{C}_{p,d} |I^d_{p,N}|$, $\tilde{C}_{p,d} > 0$, using a component–by–component strategy.

On the other hand, we obtain for the limit $p \to 0$ the frequency index set $\mathcal{I} := \{\mathbf{k} \in \mathbb{Z}^d : \|\mathbf{k}\|_1 = \|\mathbf{k}\|_\infty \leq N\}$, $N \in \mathbb{N}$, which is supported on the coordinate axis. We have $|\mathcal{I}| = 2dN + 1$ and $(2N+1)^2 \leq |\mathcal{D}(\mathcal{I})| \leq (2dN+1)^2$. Hence, we estimate $\tilde{c}_d |\mathcal{I}|^2 \leq |\mathcal{D}(\mathcal{I})|$, $\tilde{c}_d \in \mathbb{R}$, $0 < \tilde{c}_d$, and the theoretical upper bound on L is quadratic in $|\mathcal{I}|$ for fixed dimension d. In fact, reconstructing rank-1 lattices for these specific frequency index sets need at least a number of $L \in \Omega(N^2)$ nodes, cf. [36, Thm. 3.5]. □

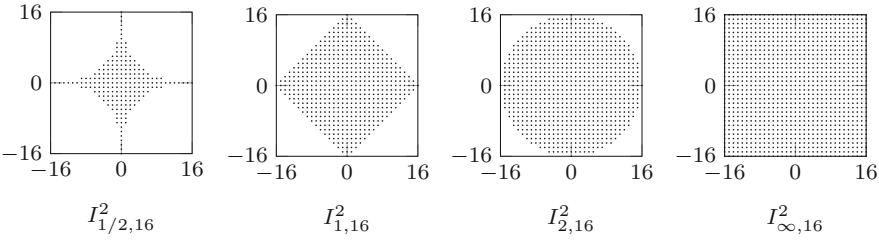


Fig. 17.3 Two-dimensional frequency index sets $I^2_{p,16}$ for $p \in \{\frac{1}{2}, 1, 2, \infty\}$

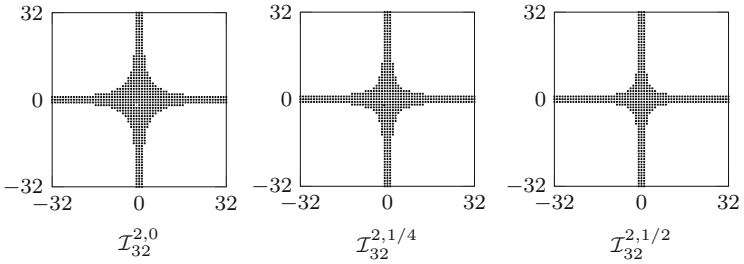


Fig. 17.4 Two-dimensional frequency index sets $\mathcal{I}_{32}^{2,T}$ for $T \in \{0, \frac{1}{4}, \frac{1}{2}\}$

Example 17.2. More useful frequency index sets in higher dimensions $d > 2$ are so-called (energy-norm based) hyperbolic crosses, cf. [3, 7, 8, 66]. In particular, we consider frequency index sets $\mathcal{I}$ of the form

$$\mathcal{I}_N^{d,T} := \left\{ \mathbf{k} \in \mathbb{Z}^d : \max(1, \|\mathbf{k}\|_1)^{\frac{T}{T-1}} \prod_{s=1}^{d} \max(1, |k_s|)^{\frac{1}{1-T}} \leq N \right\},$$

with parameter $T \in [0, 1)$ and $N \in \mathbb{N}$, see Fig. 17.4 for illustration. The frequency index set $\mathcal{I}_N^{d,0}$, i.e., $T = 0$, is in fact a symmetric hyperbolic cross and frequency index sets $\mathcal{I}_N^{d,T}$, $T \in (0, 1)$, are called energy-norm based hyperbolic crosses. The cardinality of $\mathcal{I}_N^{d,T}$ can be estimated, cf. [37, Lem. 2.6], by

$$c_{d,0} N \log^{d-1} N \leq |\mathcal{I}_N^{d,T}| \leq C_{d,0} N \log^{d-1} N, \quad \text{for } T = 0,$$
$$c_{d,T} N \leq |\mathcal{I}_N^{d,T}| \leq C_{d,T} N, \quad \text{for } T \in (0, 1),$$

where $c_{d,T}, C_{d,T} \in \mathbb{R}$, $0 < c_{d,T} \leq C_{d,T}$. Since the axis cross is a subset of the considered frequency index sets, i.e., $\{\mathbf{k} \in \mathbb{Z}^d : \|\mathbf{k}\|_1 = \|\mathbf{k}\|_\infty \leq N\} \subset \mathcal{I}_N^{d,T}$, $T \in [0, 1)$, we obtain $(2N + 1)^2 \leq |\mathcal{D}(\mathcal{I}_N^{d,T})|$. On the other hand, we obtain upper bounds of the cardinality of the difference set $\mathcal{D}(\mathcal{I}_N^{d,T})$

$$|\mathcal{D}(\mathcal{I}_N^{d,T})| \leq \tilde{C}_{d,0} N^2 \log^{d-2} N, \quad \text{for } T = 0, \text{ cf. [33, Thm. 4.8]},$$
$$|\mathcal{D}(\mathcal{I}_N^{d,T})| \leq |\mathcal{I}_N^{d,T}|^2 \leq C_{d,T}^2 N^2, \quad \text{for } T \in (0, 1).$$

Consequently, Theorem 17.2 offers a constructive strategy in order to find reconstructing rank-1 lattices for $\mathcal{I}_N^{d,T}$ of cardinality $L \leq |\mathcal{D}(\mathcal{I}_N^{d,T})|$. We would like to stress that, at least for $T \in (0, 1)$, we are able to construct rank-1 lattices of optimal order in N, cf. [33, Lem. 2.1, 2.3, and Cor. 2.4].

For instance, Fig. 17.1b(left) shows a reconstructing rank-1 lattice for the symmetric hyperbolic cross $\mathcal{I}_8^{2,0}$ and Fig. 17.1b(right) shows an example for a generated set, which allows the exact reconstruction of multivariate trigonometric polynomials

with frequencies supported on $\mathcal{I}_8^{2,0}$. The condition number of the Fourier matrix $\mathbf{A}(\mathcal{X},\mathcal{I})$ is always one when $\mathcal{X}$ is a reconstructing rank-1 lattice for $\mathcal{I}$, since the columns of the Fourier matrix $\mathbf{A}(\mathcal{X},\mathcal{I})$ are orthogonal. When the frequency index set $\mathcal{I} = \mathcal{I}_8^{2,0}$ and $\mathcal{X}$ is the specific generated set in Fig. 17.1b(right), then the condition number of the Fourier matrix $\mathbf{A}(\mathcal{X},\mathcal{I})$ is approximately 2.19. □

Reconstruction using generated sets. Up to now, we discussed reconstructing rank-1 lattices. We generalized this concept to so-called generated sets, cf. Sect. 17.2.3 and determined sufficient and necessary conditions on generated sets $\Lambda(\mathbf{r},L)$ guaranteeing a full rank and stable Fourier matrix $\mathbf{A}(\Lambda(\mathbf{r},L),\mathcal{I})$ in [32]. In general, the set $\mathcal{Y} = \{\mathbf{k}\cdot\mathbf{r} \bmod 1 : \mathbf{k}\in\mathcal{I}\} \subset \mathbb{T}$ is of our main interest, where $\mathbf{r}\in\mathbb{R}^d$ is the generating vector of the generated set $\Lambda(\mathbf{r},L)$. We determined the necessary condition $|\mathcal{Y}| = |\mathcal{I}|$ in order to obtain a Fourier matrix $\mathbf{A}(\Lambda(\mathbf{r},L),\mathcal{I})$ of full column rank.

Theorem 17.3. *Let $\mathcal{I}\subset\mathbb{Z}^d$ be an arbitrary d-dimensional index set of finite cardinality $|\mathcal{I}|$. Then, the exact reconstruction of a trigonometric polynomial with frequencies supported on $\mathcal{I}$ is possible from only $|\mathcal{I}|$ samples using a suitable generated set.*

Proof. Let $\mathbf{r}\in\mathbb{R}^d$ be a vector such that

$$\mathbf{k}\cdot\mathbf{r} \bmod 1 \neq \mathbf{k}'\cdot\mathbf{r} \bmod 1 \quad \text{for all } \mathbf{k},\mathbf{k}'\in\mathcal{I},\ \mathbf{k}\neq\mathbf{k}'. \tag{17.11}$$

For instance, Theorem 17.2 guarantees the existence of a reconstructing rank-1 lattice $\Lambda(\mathbf{r},L)$ for the index set $\mathcal{I}$, where $\mathbf{r}\in L^{-1}\mathbb{Z}^d$ fulfills property (17.11). The corresponding Fourier matrix $\mathbf{A} := (\mathrm{e}^{2\pi\mathrm{i}\mathbf{k}\cdot\mathbf{x}_\ell})_{\ell=0,\dots,L-1;\,\mathbf{k}\in\mathcal{I}} = (\mathrm{e}^{(2\pi\mathrm{i}\mathbf{k}\cdot\mathbf{r})^\ell})_{\ell=0,\dots,L-1;\,\mathbf{k}\in\mathcal{I}}$ is a transposed Vandermonde matrix of (full column) rank $|\mathcal{I}|$. If we use only the first $|\mathcal{I}|$ rows of the matrix $\mathbf{A}$ and denote this matrix by $\tilde{\mathbf{A}}$, the matrix $\tilde{\mathbf{A}} := (\mathrm{e}^{(2\pi\mathrm{i}\mathbf{k}\cdot\mathbf{r})^\ell})_{\ell=0,\dots,|\mathcal{I}|-1;\,\mathbf{k}\in\mathcal{I}} = (\mathrm{e}^{(2\pi\mathrm{i}y_j)^\ell})_{\ell=0,\dots,|\mathcal{I}|-1;\,j=0,\dots,|\mathcal{I}|-1}$ is a transposed Vandermonde matrix of size $|\mathcal{I}|\times|\mathcal{I}|$, where $y_j := \mathbf{k}_j\cdot\mathbf{r} \bmod 1$ and $\mathcal{I} = \{\mathbf{k}_0,\dots,\mathbf{k}_{|I|-1}\}$ in the specified order. Furthermore, the determinant of the transposed Vandermonde matrix $\tilde{\mathbf{A}}$, cf. [31, Sec. 6.1], is $\det\tilde{\mathbf{A}} = \prod_{1\le k<j\le|I|-1}(\mathrm{e}^{2\pi\mathrm{i}y_j} - \mathrm{e}^{2\pi\mathrm{i}y_k}) \neq 0$, since we have $\mathrm{e}^{2\pi\mathrm{i}\mathbf{k}\cdot\mathbf{r}} \neq \mathrm{e}^{2\pi\mathrm{i}\mathbf{k}'\cdot\mathbf{r}}$ for all $\mathbf{k},\mathbf{k}'\in\mathcal{I}$, $\mathbf{k}\neq\mathbf{k}'$, due to property (17.11). This means the transposed Vandermonde matrix $\tilde{\mathbf{A}}$ has full rank $|\mathcal{I}|$ and is invertible. □

Theorem 17.3 states that $L = |\mathcal{I}|$ many samples are sufficient to exactly reconstruct a trigonometric polynomial with frequencies supported on the index set $\mathcal{I}$. In general, we obtain a large condition number for the Fourier matrix $\tilde{\mathbf{A}} := (\mathrm{e}^{(2\pi\mathrm{i}\mathbf{k}\cdot\mathbf{r})^\ell})_{\ell=0,\dots,|\mathcal{I}|-1;\,\mathbf{k}\in\mathcal{I}}$. Using $L > |\mathcal{I}|$ samples, we also obtain matrices $\mathbf{A}(\Lambda(\mathbf{r},L),\mathcal{I})$ of full column rank, since the first $|\mathcal{I}|$ rows of the matrix $\mathbf{A}(\Lambda(\mathbf{r},L),\mathcal{I})$ are linear independent. In practice, growing oversampling, i.e., increasing $L > |\mathcal{I}|$, decreases at least an estimator of the condition number of $\mathbf{A}(\Lambda(\mathbf{r},L),\mathcal{I})$, as published in [32]. In this context, for each generating vector $\mathbf{r}\in\mathbb{R}^d$ bringing $|\mathcal{Y}| = |\mathcal{I}|$ and constant $C > 1$ we determined a generated set

of size L_C such that the Fourier matrix $\mathbf{A}(\Lambda(\mathbf{r}, L_C), \mathcal{I})$ has a condition number of at most C, cf. [32, Cor. 1]. We discuss a nonlinear optimization strategy in [32] in order to determine generated sets $\Lambda(\mathbf{r}, L)$ of relatively small cardinality bringing a Fourier matrix $\mathbf{A}(\Lambda(\mathbf{r}, L), \mathcal{I})$ with small condition number.

The reconstruction of multivariate trigonometric polynomials with frequencies supported on an fixed index set $\mathcal{I}$ from samples along a generated set can be realized solving the normal equation, which can be done in a fast way using the one-dimensional NFFT and a conjugate gradient (CG) method. One step of the CG method needs one NFFT of length L and one adjoint NFFT of length L. Consequently, one CG step has a complexity of $\mathcal{O}(L \log L + (d + |\log \varepsilon|)|I|)$, cf. Sect. 17.2.3. The convergence of the CG method depends on the condition number of the Fourier matrix $\mathbf{A}(\Lambda(\mathbf{r}, L), \mathcal{I})$. Hence, generated sets $\Lambda(\mathbf{r}, L)$ with small condition numbers of the Fourier matrices $\mathbf{A}(\Lambda(\mathbf{r}, L), \mathcal{I})$ guarantee a fast approximative computation of the reconstruction of trigonometric polynomials with frequencies supported on the index set $\mathcal{I}$.

17.3.4 Random Sampling and Sparse Recovery

Stable deterministic sampling schemes with a minimal number of nodes are constructed above. For arbitrary index sets of frequencies $\mathcal{I} \subset \mathbb{Z}^d$, we showed that orthogonality of the Fourier matrix necessarily implies $|\mathcal{X}| \geq |\mathcal{D}(\mathcal{I})|$ which scales (almost) quadratically in $|\mathcal{I}|$ for several interesting cases. In contrast, injectivity of the Fourier matrix can be guaranteed for a linear scaling and numerical results also support that a small oversampling factor suffices for stable reconstruction generically. Subsequently, we discuss known results for randomly chosen sampling nodes. Let $d \in \mathbb{N}$, arbitrary frequencies $\mathcal{I} \subset \mathbb{Z}^d$ be given, and sampling nodes $\mathcal{X}$ are drawn independently from the uniform distribution over the spatial domain $\mathbb{T}^d$, then [25] implies

$$\operatorname{cond}_2 \mathbf{A}(\mathcal{X}, \mathcal{I}) \leq \sqrt{\frac{1+\gamma}{1-\gamma}}, \ \gamma \in (0,1), \quad \text{if} \quad |\mathcal{X}| \geq \frac{C}{\gamma^2} |\mathcal{I}| \log \frac{|\mathcal{I}|}{\eta},$$

with probability $1 - \eta$, where $C > 0$ is some universal constant independent of the spatial dimension d. A partial derandomization can be obtained by randomly subsampling a fixed rank-1 lattice as constructed in Theorem 17.2.

Moreover, random sampling has been applied successfully in compressed sensing [9, 15, 20] to solve the sparse recovery problem (17.9), where both the support $\mathcal{I} \subset \mathcal{I}_0 \subset \mathbb{Z}^d$ as well as the Fourier coefficients $\hat{f}_{\mathbf{k}} \in \mathbb{C}$, $\mathbf{k} \in \mathcal{I}$, of the expansion (17.1) are sought. Provided a so-called restricted isometry condition is met, the sparse recovery problem can be solved efficiently, cf. [10, 43, 47, 55–57], and with probability at least $1 - \eta$ this is true if

$$|\mathcal{X}| \geq C\,|\mathcal{I}|\log^4|\mathcal{I}_0|\log\frac{1}{\eta}.$$

Well studied algorithmic approaches to actually solve the sparse recovery problem are then ℓ^1-minimisation [11], orthogonal matching pursuit [44], and their successors. Optimal variants of these algorithms have the same arithmetic complexity as one matrix vector multiplication with $\mathbf{A}(\mathcal{X},\mathcal{I}_0)$, which is however worse than the recent developments [28, 29].

Prony type methods. In contrast to compressed sensing approaches, Prony type methods aim to recover the finite and real support $\mathcal{I}$ within the bounded interval $[-\frac{N}{2},\frac{N}{2}]$ as well as the Fourier coefficients in the nonharmonic Fourier series

$$f(x)=\sum_{k\in\mathcal{I}}\hat{f}_k\mathrm{e}^{2\pi\mathrm{i}kx},$$

from equally spaced samples $f(\frac{\ell}{N})$, $\ell=0,\dots,L-1$, cf. [50,51,53]. If the number of samples fulfils a Nyquist type relation

$$|\mathcal{X}| \geq CNq_{\mathcal{I}}^{-1}$$

with respect to the nonharmonic bandwidth N and to the separation distance $q_{\mathcal{I}} := \min\{|k-k'| : k,k'\in\mathcal{I}, k\neq k'\}$, then a newly developed variant of the Prony method solves this reconstruction problem in a stable way, see e.g. [54]. The arithmetic complexity $\mathcal{O}(|\mathcal{I}|^3)$ has been improved for integer frequencies in [30] using ideas from [28, 29].

Acknowledgements We gratefully acknowledge support by the German Research Foundation (DFG) within the Priority Program 1324, project PO 711/10-2 and KU 2557/1-2. Moreover, Ines Melzer and Stefan Kunis gratefully acknowledge their support by the Helmholtz Association within the young investigator group VH-NG-526.

References

1. Aydıner, A.A., Chew, W.C., Song, J., Cui, T.J.: A sparse data fast Fourier transform (SDFFT). IEEE Trans. Antennas Propag. **51**(11), 3161–3170 (2003)
2. Bass, R.F., Gröchenig, K.: Random sampling of multivariate trigonometric polynomials. SIAM J. Math. Anal. **36**, 773–795 (2004)
3. Baszenski, G., Delvos, F.J.: A discrete Fourier transform scheme for Boolean sums of trigonometric operators. In: Chui, C.K., Schempp, W., Zeller, K. (eds.) Multivariate Approximation Theory IV. ISNM, vol. 90, pp. 15–24. Birkhäuser, Basel (1989)
4. Bebendorf, M.: Hierarchical Matrices. Lecture Notes in Computational Science and Engineering, vol. 63. Springer, Berlin (2008)
5. Beylkin, G.: On the fast Fourier transform of functions with singularities. Appl. Comput. Harmon. Anal. **2**, 363–381 (1995)

6. Björck, Å.: Numerical Methods for Least Squares Problems. SIAM, Philadelphia (1996)
7. Bungartz, H.J., Griebel, M.: A note on the complexity of solving Poisson's equation for spaces of bounded mixed derivatives. J. Complex. **15**, 167–199 (1999)
8. Bungartz, H.J., Griebel, M.: Sparse grids. Acta Numer. **13**, 147–269 (2004)
9. Candès, E.J.: Compressive sampling. In: International Congress of Mathematicians, vol. III, pp. 1433–1452. European Mathematical Society, Zürich (2006)
10. Candès, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**, 4203–4215 (2005)
11. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**, 33–61 (1998)
12. Cools, R., Kuo, F.Y., Nuyens, D.: Constructing lattice rules based on weighted degree of exactness and worst case error. Computing **87**, 63–89 (2010)
13. Demanet, L., Ferrara, M., Maxwell, N., Poulson, J., Ying, L.: A butterfly algorithm for synthetic aperture radar imaging. SIAM J. Imaging Sci. **5**, 203–243 (2012)
14. Dick, J., Kuo, F.Y., Sloan, I.H.: High-dimensional integration: the quasi-Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
15. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
16. Dutt, A., Rokhlin, V.: Fast Fourier transforms for nonequispaced data II. Appl. Comput. Harmon. Anal. **2**, 85–100 (1995)
17. Edelman, A., McCorquodale, P., Toledo, S.: The future fast Fourier transform? SIAM J. Sci. Comput. **20**, 1094–1114 (1999)
18. Feichtinger, H.G., Gröchenig, K., Strohmer, T.: Efficient numerical methods in non-uniform sampling theory. Numer. Math. **69**, 423–440 (1995)
19. Filbir, F., Themistoclakis, W.: Polynomial approximation on the sphere using scattered data. Math. Nachr. **281**, 650–668 (2008)
20. Foucart, S., Rauhut, H.: A mathematical introduction to compressive sensing. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, New York (2013)
21. Greengard, L., Rokhlin, V.: A fast algorithm for particle simulations. J. Comput. Phys. **73**, 325–348 (1987)
22. Griebel, M., Hamaekers, J.: Fast discrete Fourier transform on generalized sparse grids (2013). University of Bonn, INS Preprint No. 1305
23. Grishin, D., Strohmer, T.: Fast multi-dimensional scattered data approximation with Neumann boundary conditions. Linear Algebra Appl. **391**, 99–123 (2004)
24. Gröchenig, K.: Reconstruction algorithms in irregular sampling. Math. Comput. **59**, 181–194 (1992)
25. Gröchenig, K., Pötscher, B., Rauhut, H.: Learning trigonometric polynomials from random samples and exponential inequalities for eigenvalues of random matrices (2007, preprint). arXiv:math/0701781
26. Hackbusch, W.: Hierarchische Matrizen. Algorithmen und Analysis. Springer, Berlin/Heidelberg (2009)
27. Hallatschek, K.: Fouriertransformation auf dünnen Gittern mit hierarchischen Basen. Numer. Math. **63**, 83–97 (1992)
28. Hassanieh, H., Indyk, P., Katabi, D., Price, E.: Nearly optimal sparse Fourier transform. In: STOC, New York (2012)
29. Hassanieh, H., Indyk, P., Katabi, D., Price, E.: Simple and practical algorithm for sparse Fourier transform. In: SODA, Kyoto, pp. 1183–1194 (2012)
30. Heider, S., Kunis, S., Potts, D., Veit, M.: A sparse prony FFT. In: 10th International Conference on Sampling Theory and Applications, Bremen (2013)
31. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991)
32. Kämmerer, L.: Reconstructing multivariate trigonometric polynomials by sampling along generated sets. In: Dick, J., Kuo, F.Y., Peters, G.W., Sloan, I.H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2012, pp. 439–454. Springer, Berlin (2013)

33. Kämmerer, L.: Reconstructing hyperbolic cross trigonometric polynomials by sampling along rank-1 lattices. SIAM J. Numer. Anal. **51**, 2773–2796 (2013)
34. Kämmerer, L.: Reconstructing multivariate trigonometric polynomials from samples along rank-1 lattices. in: Approximation Theory XIV: San Antonio 2013, G.E. Fasshauer and L.L. Schumaker (eds.), Springer International Publishing, 255–271 (2014)
35. Kämmerer, L., Kunis, S.: On the stability of the hyperbolic cross discrete Fourier transform. Numer. Math. **117**, 581–600 (2011)
36. Kämmerer, L., Kunis, S., Potts, D.: Interpolation lattices for hyperbolic cross trigonometric polynomials. J. Complex. **28**, 76–92 (2012)
37. Kämmerer, L., Potts, D., Volkmer, T.: Approximation of multivariate functions by trigonometric polynomials based on rank-1 lattice sampling. Preprint 145, DFG Priority Program 1324 (2013)
38. Kämmerer, L., Potts, D., Volkmer, T.: Approximation of multivariate periodic functions by trigonometric polynomials based on sampling along rank-1 lattice with generating vector of Korobov form. Preprint 159, DFG Priority Program 1324 (2014)
39. Keiner, J., Kunis, S., Potts, D.: Fast summation of radial functions on the sphere. Computing **78**, 1–15 (2006)
40. Keiner, J., Kunis, S., Potts, D.: Using NFFT3 – a software library for various nonequispaced fast Fourier transforms. ACM Trans. Math. Softw. **36**, Article 19, 1–30 (2009)
41. Kunis, S., Melzer, I.: A stable and accurate butterfly sparse Fourier transform. SIAM J. Numer. Anal. **50**, 1777–1800 (2012)
42. Kunis, S., Potts, D.: Stability results for scattered data interpolation by trigonometric polynomials. SIAM J. Sci. Comput. **29**, 1403–1419 (2007)
43. Kunis, S., Rauhut, H.: Random sampling of sparse trigonometric polynomials II, orthogonal matching pursuit versus basis pursuit. Found. Comput. Math. **8**, 737–763 (2008)
44. Mallat, S., Zhang, Z.: Matching pursuit with time-frequency dictionaries. IEEE Trans. Signal Process. **41**, 3397–3415 (1993)
45. Mhaskar, H.N., Narcowich, F.J., Ward, J.D.: Spherical Marcinkiewicz-Zygmund inequalities and positive quadrature. Math. Comput. **70**, 1113–1130 (2001). Corrigendum on the positivity of the quadrature weights in **71**, 453–454 (2002)
46. Michielssen, E., Boag, A.: A multilevel matrix decomposition algorithm for analyzing scattering from large structures. IEEE Trans. Antennas Propag. **44**, 1086–1093 (1996)
47. Needell, D., Vershynin, R.: Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. Found. Comput. Math. **9**, 317–334 (2009)
48. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems Volume II: Standard Information for Functionals. EMS Tracts in Mathematics, vol. 12. European Mathematical Society, Zürich (2010)
49. O'Neil, M., Woolfe, F., Rokhlin, V.: An algorithm for the rapid evaluation of special function transforms. Appl. Comput. Harmon. Anal. **28**, 203–226 (2010)
50. Peter, T., Plonka, G.: A generalized prony method for reconstruction of sparse sums of eigenfunctions of linear operators. Inverse Probl. **29**, 025,001 (2013)
51. Peter, T., Potts, D., Tasche, M.: Nonlinear approximation by sums of exponentials and translates. SIAM J. Sci. Comput. **33**, 314–334 (2011)
52. Potts, D., Steidl, G., Tasche, M.: Fast Fourier transforms for nonequispaced data: a tutorial. In: Benedetto, J.J., Ferreira, P.J.S.G. (eds.) Modern Sampling Theory: Mathematics and Applications, pp. 247–270. Birkhäuser, Boston (2001)
53. Potts, D., Tasche, M.: Parameter estimation for exponential sums by approximate Prony method. Signal Process. **90**, 1631–1642 (2010)
54. Potts, D., Tasche, M.: Parameter estimation for nonincreasing exponential sums by Prony-like methods. Linear Algebra Appl. **439**, 1024–1039 (2013)
55. Rauhut, H.: Random sampling of sparse trigonometric polynomials. Appl. Comput. Harmon. Anal. **22**, 16–42 (2007)
56. Rauhut, H.: On the impossibility of uniform sparse reconstruction using greedy methods. Sampl. Theory Signal Image Process. **7**, 197–215 (2008)

57. Rauhut, H.: Stability results for random sampling of sparse trigonometric polynomials. IEEE Trans. Inf. Theory **54**, 5661–5670 (2008)
58. Sickel, W., Ullrich, T.: The Smolyak algorithm, sampling on sparse grids and function spaces of dominating mixed smoothness. East J. Approx. **13**, 387–425 (2007)
59. Sloan, I.H., Joe, S.: Lattice methods for multiple integration. Oxford Science Publications. The Clarendon Press/Oxford University Press, New York (1994)
60. Steidl, G.: A note on fast Fourier transforms for nonequispaced grids. Adv. Comput. Math. **9**, 337–353 (1998)
61. Temlyakov, V.N.: Approximation of functions with bounded mixed derivative. Trudy Mat. Inst. Steklov (Proc. Steklov Inst. Math. (1989)), vol. 178 (1986)
62. Tygert, M.: Fast algorithms for spherical harmonic expansions, III. J. Comput. Phys. **229**, 6181–6192 (2010)
63. Ying, L.: Sparse Fourier transform via butterfly algorithm. SIAM J. Sci. Comput. **31**, 1678–1694 (2009)
64. Ying, L., Biros, G., Zorin, D.: A kernel-independent adaptive fast multipole method in two and three dimensions. J. Comput. Phys. **196**, 591–626 (2004)
65. Ying, L., Fomel, S.: Fast computation of partial Fourier transforms. Multiscale Model. Simul. **8**, 110–124 (2009)
66. Yserentant, H.: Regularity and Approximability of Electronic Wave Functions. Lecture Notes in Mathematics. Springer, Berlin (2010)

Chapter 18
Sparsity and Compressed Sensing in Inverse Problems

Evelyn Herrholz, Dirk Lorenz, Gerd Teschke, and Dennis Trede

Abstract This chapter is concerned with two important topics in the context of sparse recovery in inverse and ill-posed problems. In first part we elaborate conditions for exact recovery. In particular, we describe how both ℓ^1-minimization and matching pursuit methods can be used to regularize ill-posed problems and moreover, state conditions which guarantee exact recovery of the support in the sparse case. The focus of the second part is on the incomplete data scenario. We discuss extensions of compressed sensing for specific infinite dimensional ill-posed measurement regimes. We are able to establish recovery error estimates when adequately relating the isometry constant of the sensing operator, the ill-posedness of the underlying model operator and the regularization parameter. Finally, we very briefly sketch how projected steepest descent iterations can be applied to retrieve the sparse solution.

18.1 Introduction

Many applications in science and engineering require the solution of an operator equation $Kx = y$. Often only noisy data y^δ with $\|y^\delta - y\| \leq \delta$ are available, and if the problem is ill-posed, regularization methods have to be applied. During the last three decades, the theory of regularization methods for treating linear problems in a Hilbert space framework has been well developed, see, e.g., [23, 29, 30, 39]. Influenced by the huge impact of sparse signal representations and the practical

E. Herrholz • G. Teschke (✉)
Neubrandenburg University of Applied Sciences, Brodaer Straße 2, 17033 Neubrandenburg, Germany
e-mail: evelyn@herrholz.de; teschke@hs-nb.de

D. Lorenz
Technical University of Braunschweig, Pockelsstr. 14, 38106 Braunschweig, Germany
e-mail: d.lorenz@tu-braunschweig.de

D. Trede
University of Bremen, Bibliothekstrae 1, Gebäude MZH, 28359 Bremen, Germany
e-mail: trede@math.uni-bremen.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_18

feasibility of advanced sparse recovery algorithms, the combination of sparse signal recovery and inverse problems emerged in the last decade as a new growing area. Currently, there exist a great variety of sparse recovery algorithms for inverse problems (linear as well as for nonlinear operator equations) within this context, see, e.g., [5–7, 14–16, 25, 26, 41, 44, 45]. These recovery algorithms are successful for many applications and have lead to breakthroughs in many fields. However, the feasibility is usually limited to problems for which the data are complete and where the problem is of moderate dimension. For really large-scale problems or problems with incomplete data, these algorithms are not well-suited and often far off exact recovery or fail completely.

Within this chapter we focus on two neighboring questions arising in sparse recovery of solutions of inverse problems. The first is concerned with *exact recovery conditions in the complete data scenario*, and the second is concerned with *sparse recovery in the compressively sensed data scenario*.

Exact recovery. The two most widely used recovery methods, namely ℓ^1-minimization and matching pursuit methods, can be related to two classical methods for regularization of ill-posed problems: ℓ^1-minimization is a special case of variational regularization in which the operator equation $Kx = y$ is replaced by a well-posed minimization problem with a sparsity constraint. Matching pursuit methods are related to iterative regularization methods in which one uses an iterative method to solve the operator equation and uses a stopping criterion to prevent noise amplification. We describe how both ℓ^1-minimization and matching pursuit methods can be used to regularize ill-posed problems and moreover, state conditions which guarantee exact recovery of the support in the sparse case.

Compressive sensing. For the incomplete data situation, the mathematical technology called compressive sensing, which turned out to be quite successful in sparse signal recovery, was established several years ago by D. Donoho, see [18]. A major breakthrough was achieved when it was proven that it is possible to reconstruct a signal from very few measurements under certain conditions on the signal and the measurement model, see [8–10, 18–20, 24, 42]. In [12] it was shown that if the sensing operator satisfies the restricted isometry property the solution can be reconstructed exactly by minimization of an ℓ_1 constrained problem, provided that the solution is sparse enough. Classical formulations of compressed sensing are finite dimensional. Quite recently, continuous formulations have appeared, see [1] (full continuous sensing model) and see, e.g., [22, 33, 38] (problem of analog-to-digital conversion). Within this chapter we summarize extensions of the infinite dimensional model in [22] to the case of compressively sampling ill-posed problems and provide iterative sparse recovery principles and corresponding error estimates, for detailed discussions see [31]. Further extensions towards generalized and compressive sampling also on the context of ill-posed problems can be found in [2, 3], and [4].

18.2 Exact Recovery for Ill-Posed Problems

In this section we describe how both ℓ^1-minimization and matching pursuit methods can be used to regularize ill-posed problems and moreover, state condition which guarantee exact recovery of the support in the sparse case.

18.2.1 Orthogonal Matching Pursuit

In a Banach space X, we assume that we have a given *dictionary* of unit-normed *atoms* $(e_i) = \mathscr{E}$. We assume that the solutions to an operator equation $Kx = y$ (with $K : X \to Y$ bounded, injective and linear and Y a Hilbert space) can be expressed sparsely in $\mathscr{E}$, i.e. that

$$x = \sum_{i \in \mathbb{Z}} \alpha_i e_i \quad \text{with} \quad \alpha_i \in \mathbb{R}, \quad \|\alpha\|_{\ell^0} =: N < \infty.$$

Now assume that instead of $y = Kx$ we are given a noisy measurement y^ε with $\|y - y^\varepsilon\| \leq \varepsilon$ and aim to recover a good approximation of x from the measurement y^ε.

In the following we denote with I the support of the coefficient vector α, i.e. $I = \{i \in \mathbb{Z} \,|\, \alpha_i \neq 0\}$. For any subset $J \subset \mathbb{Z}$ we denote $\mathscr{E}(J) := \{e_i \,|\, i \in J\}$.

The above setting is of practical relevance, e.g. in mass spectrometry [32] where the signal is modeled as a sum of Dirac peaks (so-called impulse trains) $x = \sum_{i \in \mathbb{Z}} \alpha_i \, \delta(\cdot - t_i)$. Another example can be found in digital droplet holography, cf. [43], where images arise as superposition of characteristic functions of balls with different centers t_i and radii r_j, $x = \sum_{i,j \in \mathbb{Z}} \alpha_{i,j} \, \chi_{B_{r_j}}(\cdot - t_i)$.

In this section we approach the problem "$Kx = y^\varepsilon$" by iteratively including more and more atoms in the representation of x in a "greedy" fashion—an algorithmic idea which is also known under the name *matching pursuit*. We define another normed dictionary

$$\mathscr{D} := \{d_i\}_{i \in \mathbb{Z}} := \left\{ \frac{Ke_i}{\|Ke_i\|} \right\}_{i \in \mathbb{Z}}.$$

Note that $\mathscr{D}$ is well defined by the injectivity of K. In any step of our iterative method we select that atom from the dictionary $\mathscr{D}$ which is correlates most with the residual (hence the name "greedy" method). To stabilize the solution of "$Kx = y^\varepsilon$" the iteration has to be stopped early enough. We only investigate the so-called *orthogonal matching pursuit* (OMP), first proposed in the signal processing context by Davis et al. in [36] and Pati et al. in [40] as an improvement upon the matching pursuit algorithm [37]. The algorithm is stated in Fig. 18.1.

1. Initialize $k = 0$, $I^0 = \emptyset$, $r^0 = y^\varepsilon$, $\widehat{x}^0 = 0$
2. While $\|r_k\| > \varepsilon$ do:
   a. Increase k and select an atom by $i_k \in \operatorname{argsup}\{|\langle r^{k-1}, d_i\rangle| \,|\, d_i \in \mathscr{D}\}$,
   b. Set $I^k = I^{k-1} \cup \{i_k\}$ and project x onto $\operatorname{span}\mathscr{E}(I^k)$, i.e. $\widehat{x}^k = \operatorname{argmin}\{\|v^\varepsilon - K\widehat{u}\|^2 \,|\, \widehat{u} \in \operatorname{span}\mathscr{E}(I^k)\}$,
   c. Set $r^k := y^\varepsilon - K\widehat{x}^k$.

Fig. 18.1 Orthogonal matching pursuit

Necessary and sufficient conditions for exact support recovery by OMP are given in [46]. Next, we list this result in the language of infinite-dimensional inverse problems. We define the linear continuous synthesis operator for the dictionary $\mathscr{D}$ via $D : \ell^1 \to Y$, $D\beta = \sum_{i\in\mathbb{Z}} \beta_i d_i = \sum_{i\in\mathbb{Z}} \beta_i \frac{Ke_i}{\|Ke_i\|}$. Furthermore, for $J \subset \mathbb{Z}$ we denote with $P_J : \ell^1 \to \ell^1$ the projection onto J and with $A^\dagger$ the pseudoinverse of an operator A. With this notation we state the following theorem.

Theorem 18.1 (Tropp [46]). *Let $\alpha \in \ell^0$ with $\operatorname{supp}\alpha = I$, $x = \sum_{i\in\mathbb{Z}} \alpha_i e_i$ be the source and $y = Kx$ the measured signal. If the operator $K : X \to Y$ and the dictionary $\mathscr{E} = \{e_i\}_{i\in\mathbb{Z}}$ fulfill the* Exact Recovery Condition (ERC)

$$\sup_{d\in\mathscr{D}(I^\complement)} \|(DP_I)^\dagger d\|_{\ell^1} < 1, \tag{18.1}$$

then OMP with its parameter ε set to 0 recovers α exactly.

The necessity of the condition (18.1) is shown in [46], by constructing a signal such that for ≥ 1 in (18.1), OMP fails to recover it.

A weaker sufficient condition is derives by Dossal and Mallat [21] and Gribonval and Nielsen [28] and this condition only depends on inner products of the dictionary atoms of $\mathscr{D}(I)$ and $\mathscr{D}(I^\complement)$ only and hence, is simpler to evaluate (although the condition is not necessary).

Proposition 18.1 (Dossal and Mallat [21], Gribonval and Nielsen [28]). *Let $\alpha \in \ell^0$ with $\operatorname{supp}\alpha = I$,. If the operator $K : X \to Y$ and the dictionary $\mathscr{E} = \{e_i\}_{i\in\mathbb{Z}}$ fulfill the* Neumann ERC

$$\sup_{i\in I} \sum_{j\in I,\ j\neq i} |\langle d_i, d_j\rangle| + \sup_{i\in I^\complement} \sum_{j\in I} |\langle d_i, d_j\rangle| < 1, \tag{18.2}$$

then OMP with its parameter ε set to 0 recovers α.

The transfer to noisy signals $y^\varepsilon = y + \eta = Kx + \eta$ with $\|v - v^\varepsilon\| = \|\eta\| \leq \varepsilon$ (where OMP has to stop as soon as $\varepsilon \geq \|r^k\|$) is contained in the following theorem from [17].

Theorem 18.2 (ERC in the Presence of Noise). *Let $\alpha \in \ell^0$ with $\operatorname{supp}\alpha = I$. Let $x = \sum_{i\in\mathbb{Z}} \alpha_i e_i$ be the source and $y^\varepsilon = Kx + \eta$ the noisy data with noise level*

$\|\eta\| \leq \varepsilon$ *and noise-to-signal-ratio*

$$r_{\varepsilon/\alpha} := \frac{\sup_{i\in\mathbb{Z}} |\langle \eta, d_i\rangle|}{\min_{i\in I} |\alpha_i| \|Ke_i\|}.$$

If the operator K *and the dictionary* $\mathscr{E}$ *fulfill the* Exact Recovery Condition in Presence of Noise (εERC)

$$\sup_{d\in\mathscr{D}(I^{\complement})} \|(DP_I)^{\dagger} d\|_{\ell^1} < 1 - 2\, r_{\varepsilon/\alpha} \frac{1}{1 - \sup_{i\in I} \sum_{j\in I,\ j\neq i} |\langle d_i, d_j\rangle|}, \tag{18.3}$$

and $\sup_{i\in I} \sum_{j\in I,\ j\neq i} |\langle d_i, d_j\rangle| < 1$, *then OMP recovers the support* I *of* α *exactly.*

To ensure the εERC (18.3) one has necessarily for the noise-to-signal-ratio $r_{\varepsilon/\alpha} < 1/2$. A rough upper bound for $\sup_{i\in\mathbb{Z}} |\langle \eta, d_i\rangle|$ is ε and hence, one may use $r_{\varepsilon/\alpha} \leq \varepsilon / (\min_{i\in I} |\alpha_i| \|Ke_i\|)$.

Similarly to the result of Dossal and Mallat, one can give a weaker sufficient recovery condition that depends on inner products of the dictionary atoms. It is proved analogously to Proposition 18.1 (see [17]).

Proposition 18.2 (Neumann ERC in the Presence of Noise). *If the operator* K *and the dictionary* $\mathscr{E}$ *fulfill the* Neumann εERC

$$\sup_{i\in I} \sum_{j\in I,\ j\neq i} |\langle d_i, d_j\rangle| + \sup_{i\in I^{\complement}} \sum_{j\in I} |\langle d_i, d_j\rangle| < 1 - 2\, r_{\varepsilon/\alpha}, \tag{18.4}$$

then OMP recovers the support I *of* α *exactly.*

Theorem 18.2 and Proposition 18.2 ensure that the correct support I is identified and the following proposition additionally shows that the reconstruction error is of the order of the noise level.

Proposition 18.3 (Error bounds for OMP in presence of noise). *If the* ε*ERC is fulfilled then there exists a constant* $C > 0$ *such that for the approximative solution* $\hat{\alpha}$ *determined by OMP it holds that* $\|\hat{\alpha} - \alpha\|_{\ell^1} \leq C\,\varepsilon$.

The proof can also be found in [17]

18.2.2 ℓ^1-Minimization

In ℓ^1-minimization one promotes sparsity of the approximate solution of $Kx = y^{\varepsilon}$ by a sparsity constraint. In this section we assume that x itself is the object which is sparse, i.e. $x \in \ell^2$ with $|\mathrm{supp} x| = N < \infty$. A typical sparsity constraint is given by the ℓ^1-norm and hence, we investigate the minimization problem

$$\min_x \left\{ T_\lambda(x) = \tfrac{1}{2}\|Kx - y^\varepsilon\|^2 + \lambda \|x\|_{\ell^1} \right\}.$$

This method is also called Basis Pursuit Denoising [13].

In [27, 34] it has been shown that ℓ^1 minimization is indeed a regularization method and also an error estimate have been derived. A central ingredient is the so-called Finite Basis Injectivity property (FBI-property) of the operator K introduced in [7]. An operator K has the FBI property if for all finite subsets $J \subset \mathbb{Z}$ the operator K restricted to $\operatorname{span}\{e_i \mid i \in J\}$ is injective, (in other words, for all $x, z \in \ell^2$ with $Kx = Kz$ and $x_k = z_k = 0$, for all $k \notin J$, it follows that $x = z$). Note that the FBI property can be seen as a variant of the restricted isometry property (introduced in the next section).

Theorem 18.3 (Error estimate). *Let K possess the FBI property, x be sparse with $\operatorname{supp} x = I$ be a minimum-$\|\cdot\|_{\ell^1}$ solution of $Kx = y$, and $\|y - y^\varepsilon\| \le \varepsilon$. Let the following source condition (SC) be fulfilled:*

$$\text{there exists } w \in Y \text{ such that } K^* w = \xi \in \operatorname{Sign}(x). \tag{18.5}$$

Moreover, let $\theta = \sup\{|\xi_k| \,|\, |\xi_k| < 1\}$ and $c > 0$ such that for all $z \in \ell^2$ with $\operatorname{supp}(z) \subset I$ it holds $\|Ku\| \ge c\|u\|$. Then for the minimizers $x^{\lambda,\varepsilon}$ of T_λ it holds

$$\|x^{\lambda,\varepsilon} - x\|_{\ell^1} \le \frac{\|K\| + 1}{1 - \theta}\frac{\varepsilon^2}{\lambda} + \Big(\frac{1}{c} + \|w\|\frac{\|K\| + 1}{1 - \theta}\Big)(\lambda + \varepsilon). \tag{18.6}$$

Especially, with $\lambda \asymp \varepsilon$ it holds

$$\|x^{\lambda,\varepsilon} - x\|_{\ell^1} = \mathcal{O}(\varepsilon). \tag{18.7}$$

In addition to the above error estimate one can give an a priori parameter rule which ensures that the unknown support of the sparse solution $x \in \ell^0$ is recovered exactly (cf. [35]).

Theorem 18.4 (Lower bound on α). *Let $x \in \ell^0$, $\operatorname{supp}(x) = I$, and $y^\varepsilon = Kx + \eta$ the noisy data. Assume that K is bounded and possesses the FBI property. If the following condition holds,*

$$\sup_{i \in I^{\complement}} \|(KP_I)^\dagger K e_i\|_{\ell^1} < 1, \tag{18.8}$$

then the parameter rule

$$\alpha > \frac{1 + \sup_{i \in I^{\complement}} \|(KP_I)^\dagger K e_i\|_{\ell^1}}{1 - \sup_{i \in I^{\complement}} \|(KP_I)^\dagger K e_i\|_{\ell^1}} \sup_{i \in \mathbb{Z}} |\langle \eta, K e_i \rangle| \tag{18.9}$$

ensures that the support of $x^{\lambda,\varepsilon}$ is contained in I.

Theorem 18.4 gives a lower bound on the regularization parameter λ to ensure $\operatorname{supp}(x^{\lambda,\varepsilon}) \subset \operatorname{supp}(x)$. To even guarantee $\operatorname{supp}(x^{\lambda,\varepsilon}) = \operatorname{supp}(x)$ we need an additional upper bound for λ. The following theorem from [35] leads to that purpose.

Theorem 18.5 (Error estimate). *Let the assumptions of Theorem 18.4 hold and choose λ according to (18.9). Then the following error estimate is valid:*

$$\|x - u^{\lambda,\varepsilon}\|_{\ell^\infty} \leq (\lambda + \sup_{i\in\mathbb{Z}} |\langle \eta, Ke_i\rangle|) \|(P_I K^* K P_I)^{-1}\|_{\ell^1,\ell^1}. \tag{18.10}$$

Remark 18.1. Due to the error estimate (18.10) we achieve a linear convergence rate measured in the ℓ^∞ norm. In finite dimensions all ℓ^p norms are equivalent, hence we also get an estimate for the ℓ^1 error:

$$\|x - x^{\lambda,\varepsilon}\|_{\ell^1} \leq (\lambda + \varepsilon\|K\|)\ |I|\ \|(P_I K^* K P_I)^{-1}\|_{\ell^1,\ell^1}.$$

Compared to the estimate (18.6) from Theorem 18.3, the quantities θ and $\|w\|$ are not present anymore. The role of $1/c$ is now played by $\|(P_I K^* K P_I)^{-1}\|_{\ell^1,\ell^1}$. However, if upper bounds on I or on its size (together with structural information on K) is available, the estimate can give a-priori checkable error estimates.

Theorem 18.6 (Exact recovery condition in the presence of noise). *Let $x \in \ell^0$ with $\operatorname{supp}(x) = I$ and $y^\varepsilon = Kx + \eta$ the noisy data with noise-to-signal ratio*

$$r_{\eta/u} := \frac{\sup_{i\in\mathbb{Z}} |\langle \eta, Ke_i\rangle|}{\min_{i\in I} |x_i|}.$$

Assume that the operator K is bounded and possesses the FBI property. Then the exact recovery condition in the presence of noise (εERC)

$$\sup_{i\in I^\complement} \|(KP_I)^\dagger Ke_i\|_{\ell^1} < 1 - 2r_{\eta/u}\|(P_I K^* K P_I)^{-1}\|_{\ell^1,\ell^1} \tag{18.11}$$

ensures that there is a suitable regularization parameter λ,

$$\frac{1 + \sup_{i\in I^\complement} \|(KP_I)^\dagger Ke_i\|_{\ell^1}}{1 - \sup_{i\in I^\complement} \|(KP_I)^\dagger Ke_i\|_{\ell^1}} \sup_{i\in\mathbb{Z}} |\langle \eta, Ke_i\rangle| < \lambda \tag{18.12}$$

$$\lambda < \frac{\min_{i\in I} |u_i^\diamond|}{\|(P_I K^* K P_I)^{-1}\|_{\ell^1,\ell^1}} - \sup_{i\in\mathbb{Z}} |\langle \eta, Ke_i\rangle|,$$

which provides exact recovery of I.

18.3 Compressive Sensing Principles for Ill-Posed Problems

Within this section we combine the concepts of compressive sensing and sparse recovery for solving inverse and ill-posed problems. To establish an adequate measurement model, we adapt an infinite dimensional compressed sensing setup that was invented in [22]. As the main result we provide recovery accuracy estimates for the computed sparse approximations in the language of [11] but now for the solution of the underlying inverse problem. One essential difference to the classical compressed sensing framework is the incorporation of joint sparsity measures allowing the treatment of infinite dimensional reconstruction spaces. Moreover, to tackle ill-posed operator equations we rely on constrained optimization formulations that are very close to elastic net type optimizations.

18.3.1 Compressive Sensing Model and Classical Results

Within this section we provide the standard reconstruction space, the compressive sensing model and repeat classical recovery results for finite-dimensional problems that can be established thanks to the restricted isometry property of the underlying sensing matrix.

Let X be a separable Hilbert space and $X_m \subset X$ the (possibly infinite dimensional) reconstruction space defined by

$$X_m = \left\{ x \in X,\, x = \sum_{\ell=1}^{m} \sum_{\lambda \in \Lambda} d_{\ell,\lambda} a_{\ell,\lambda},\;\; d \in (\ell_2(\Lambda))^m \right\},$$

where we assume that Λ is a countable index set and $\Phi_a = \{a_{\ell,\lambda},\, \ell = 1, \ldots, m\,,\, \lambda \in \Lambda\}$ forms a frame for X_m with frame bounds $0 < C_{\Phi_a} \leq C^{\Phi_a} < \infty$. Note that the reconstruction space X_m is a subspace of X with possibly large m. Typically we consider functions of the form $a_{\ell,\lambda} = a_\ell(\cdot - \lambda \mathscr{T})$, for some $\mathscr{T} > 0$. With respect to Φ_a we define the map $F_a : X_m \to (\ell_2(\Lambda))^m$ through $x \mapsto F_a x = (\{\langle x, a_{1,\lambda}\rangle\}_{\lambda\in\Lambda}, \ldots, \{\langle x, a_{m,\lambda}\rangle\}_{\lambda\in\Lambda})^T$. F_a is the analysis operator and its adjoint, given by $F_a^* : (\ell_2(\Lambda))^m \to X_m$ through $d \mapsto F_a^* d = \sum_{\ell=1}^m \sum_{\lambda\in\Lambda} d_{\ell,\lambda} a_{\ell,\lambda}$, is the so-called synthesis operator. Since Φ_a forms a frame, each $x \in X_m$ can be reconstructed from its moments $F_a x$ through $(F_a^* F_a)^{-1} F_a^*$. A special choice of analysis/sampling functions might relax the situation a bit. Assume we have another family of sampling functions Φ_v at our disposal fulfilling $F_v F_a^* = I$, then it follows with $x = F_a^* d$

$$y = F_v x = \begin{pmatrix} \{\langle x, v_{1,\lambda}\rangle\}_{\lambda\in\Lambda} \\ \vdots \\ \{\langle x, v_{m,\lambda}\rangle\}_{\lambda\in\Lambda} \end{pmatrix} = \begin{pmatrix} \{\langle F_a^* d, v_{1,\lambda}\rangle\}_{\lambda\in\Lambda} \\ \vdots \\ \{\langle F_a^* d, v_{m,\lambda}\rangle\}_{\lambda\in\Lambda} \end{pmatrix} = F_v F_a^* d = d\,, \tag{18.13}$$

i.e. the sensed values y equal d and therefore $x = F_a^* F_v x$. The condition $F_v F_a^* = I$ means nothing else than $\langle a_{\ell,\lambda}, v_{\ell',\lambda'} \rangle = \delta_{\lambda\lambda'}\delta_{\ell\ell'}$ for all $\lambda, \lambda' \in \Lambda$ and $\ell, \ell' = 1, \ldots, m$, i.e. Φ_v and Φ_a are biorthogonal to each other.

As we focus on reconstructing functions (or solutions of operator equations) x that have a sparse series expansion $x = F_a^* d$ with respect to Φ_a, i.e. the series expansion of x has only a very small number of non-vanishing coefficients $d_{\ell,\lambda}$, or that x is compressible (meaning that x can be well-approximated by a sparse series expansion), the theory of compressed sensing suggests to sample x at much lower rate as done in the classical setting mentioned above (there it was $m/\mathscr{T}$) while ensuring exact recovery of x (or recovery with overwhelming probability). The compressive sampling idea applied to the sensing situation (18.13) goes now as follows. Assume we are given a sensing matrix $A \in \mathbb{R}^{p\times m}$ with $p \ll m$. Then we construct p species of sampling functions through

$$\begin{pmatrix} s_{1,\lambda} \\ \vdots \\ s_{p,\lambda} \end{pmatrix} = A \begin{pmatrix} v_{1,\lambda} \\ \vdots \\ v_{m,\lambda} \end{pmatrix} \quad \text{for all } \lambda \in \Lambda\,. \tag{18.14}$$

As a simple consequence of (18.14), the following lemma holds true.

Lemma 18.1. *Assume for all $\lambda \in \Lambda$ the sampling functions $s_{1,\lambda}, \ldots, s_{p,\lambda}$ are chosen as in (18.14) and let y denote the exactly sensed data. If Φ_a and Φ_v are biorthogonal to each other, then $y = Ad$.*

Let d_λ denote the m-dimensional vector $(d_{1,\lambda}, \ldots, d_{m,\lambda})^T$ and y_λ the p-dimensional vector $(y_{1,\lambda}, \ldots, y_{p,\lambda})^T$, then Lemma 18.1 states that for each $\lambda \in \Lambda$ the measurement vectors are given by $y_\lambda = Ad_\lambda$. It has been shown in [12], that for each individual $\lambda \in \Lambda$ the solution d_λ^* to

$$\min_{d_\lambda \in \mathbb{R}^m} \|d_\lambda\|_{\ell_1} \quad \text{subject to} \quad y_\lambda = Ad_\lambda\,, \tag{18.15}$$

recovers d_λ exactly provided that d_λ is sufficiently sparse and the matrix A obeys a condition known as the *restricted isometry property*.

Definition 18.1 (restricted isometry property). For each integer $k = 1, 2, \ldots$, define the isometry constant δ_k of a sensing matrix A as the smallest number such that

$$(1 - \delta_k)\|x\|_{\ell_2}^2 \le \|Ax\|_{\ell_2}^2 \le (1 + \delta_k)\|x\|_{\ell_2}^2 \tag{18.16}$$

holds for all k-sparse vectors x. A vector is said to be k-sparse if it has at most k non-vanishing entries.

Theorem 18.7 (noiseless recovery, Candès [11]). *Assume* $\delta_{2k} < \sqrt{2}-1$. *Then for each* $\lambda \in \Lambda$ *the solution* d_λ^* *to* (18.15) *obeys*

$$\|d_\lambda^* - d_\lambda\|_{\ell_1} \leq C_0 \|d_\lambda^k - d_\lambda\|_{\ell_1} \tag{18.17}$$

$$\|d_\lambda^* - d_\lambda\|_{\ell_2} \leq C_0 k^{-1/2} \|d_\lambda^k - d_\lambda\|_{\ell_1} \tag{18.18}$$

for some constant C_0 *(that can be explicitly computed) and* d_λ^k *denoting the best* k*-term approximation. If* d_λ *is* k*-sparse, the recovery is exact.*

This result can be extended to the more realistic scenario in which the measurements are contaminated by noise. Then we have to solve

$$\min_{d_\lambda \in \mathbb{R}^m} \|d_\lambda\|_{\ell_1} \quad \text{subject to} \quad \|y_\lambda^\delta - A d_\lambda\|_{\ell_2} \leq \delta \,. \tag{18.19}$$

Theorem 18.8 (noisy recovery, Candès [11]). *Assume* $\delta_{2k} < \sqrt{2} - 1$ *and* $\|y_\lambda^\delta - y_\lambda\|_{\ell_2} \leq \delta$. *Then for each* $\lambda \in \Lambda$ *the solution* d_λ^* *to* (18.19) *obeys*

$$\|d_\lambda^* - d_\lambda\|_{\ell_2} \leq C_0 k^{-1/2} \|d_\lambda^k - d_\lambda\|_{\ell_1} + C_1 \delta \tag{18.20}$$

with the same constant C_0 *as before and some* C_1 *(that can be explicitly computed).*

18.3.2 Infinite Dimensional Regime and Joint Sparsity Measures

In the previous subsection we have summarized results that apply for all individual sensing scenarios, i.e. that hold true for all individual $\lambda \in \Lambda$. But as the index set Λ is possibly of infinite cardinality, we are faced with the problem of recovering infinitely many unknown vectors d_λ for which the (essential) support can be different. Therefore, the determination of d by solving for each λ an individual optimization problem is numerically not feasible.

For a simultaneous treatment of all individual optimization problems, we have to restrict the set of all possible solutions d_λ. One quite natural restriction is that all d_λ share a joint sparsity pattern. Introducing support sets $\mathscr{I} \subset \{1, \ldots, m\}$, the reconstruction space is given through

$$X_k = \left\{ x \in X,\, x = \sum_{\ell \in \mathscr{I}, |\mathscr{I}| = k} \sum_{\lambda \in \Lambda} d_{\ell,\lambda} a_{\ell,\lambda},\;\; d \in (\ell_2(\Lambda))^m \right\} , \tag{18.21}$$

i.e. only k out of m sequences $\{d_{\ell,\lambda}\}_{\lambda\in\Lambda}$ do not vanish. The space X_k is no longer a subspace since two different x might correspond to two different support sets $\mathscr{I}$ and therefore its sum is not contained in X_k. The space X_k can be seen as a union of (shift invariant) subspaces.

To solve the recovery problem we propose a constrained optimization approach. Let therefore the linear sensing operator T be given by $T : (\ell_2(\Lambda))^m \to (\ell_2(\Lambda))^p$ via $T\,d = T(\{d_{1,\lambda}\}_{\lambda\in\Lambda},\dots,\{d_{m,\lambda}\}_{\lambda\in\Lambda})^T = (\{(Ad_\lambda)^1\}_{\lambda\in\Lambda},\dots,\{(Ad_\lambda)^p\}_{\lambda\in\Lambda})^T$. For the purpose of identifying the support set $\mathscr{I}$ we restrict the minimization of $\|y^\delta - Td\|^2_{(\ell_2(\Lambda))^p}$ to the sub-domain

$$B(\Psi_{1,2},R) = \{d \in (\ell_2(\Lambda))^m : \Psi_{1,2}(d) \le R\}\,,$$

where $\Psi_{q,r}$ is a joint sparsity measure defined by $\Psi_{q,r}(d) = \left(\sum_{\ell=1}^m \left(\sum_{\lambda\in\Lambda} |d_{\ell,\lambda}|^r\right)^{\frac{q}{r}}\right)^{\frac{1}{q}}$. This measure forces the solution d for reasonably small chosen q (e.g. $1 \le q < 2$) to have non-vanishing rows $\{d_{\ell,\lambda}\}_{\lambda\in\Lambda}$ only if $\|\{d_{\ell,\lambda}\}_{\lambda\in\Lambda}\|_{\ell_r(\Lambda)}$ is large enough. Consequently, the optimization reads then as

$$\min_{d\in B(\Psi_{1,2},R)} \|y^\delta - Td\|^2_{(\ell_2(\Lambda))^p}\,, \tag{18.22}$$

where the minimizing element in $B(\Psi_{1,2},R)$ can be approached by

$$d^{n+1} = \mathscr{P}\left(d^n + \frac{\gamma}{C}T^*(y^\delta - Td^n)\right)\,, \tag{18.23}$$

where $\gamma > 0$ is a step-length control (determined below) and $\mathscr{P}$ is the ℓ_2-projection on $B(\Psi_{1,2},R)$, which can be realized by the sequence-valued generalized soft-shrinkage operator. To control the speed of convergence we introduce conditions on γ.

Definition 18.2. We say that the sequence $\{\gamma^n\}_{n\in\mathbb{N}}$ satisfies Condition *(B)* with respect to the sequence $\{d^n\}_{n\in\mathbb{N}}$ if there exists n_0 such that:

(B1) $\sup\{\gamma^n; n\in\mathbb{N}\} < \infty$ and $\inf\{\gamma^n; n\in\mathbb{N}\} \ge 1$

(B2) $\gamma^n\|Td^{n+1} - Td^n\|^2_{(\ell_2(\Lambda))^p} \le C\,\|d^{n+1} - d^n\|^2_{(\ell_2(\Lambda))^m} \quad \forall n \ge n_0\,.$

Proposition 18.4. *If for arbitrarily chosen d^0 assume d^{n+1} is given by*

$$d^{n+1} = \mathscr{P}\left(d^n + \frac{\gamma^n}{C}T^*(y^\delta - Td^n)\right)\,, \tag{18.24}$$

with γ^n satisfying Condition (B) *with respect to $\{d^n\}_{n\in\mathbb{N}}$, the sequence of residuals $\|y^\delta - Td^n\|^2_{(\ell_2(\Lambda))^p}$ is monotonically decreasing and $\{d^n\}_{n\in\mathbb{N}}$ converges in norm towards d^*, where d^* fulfills the necessary condition for a minimum of* (18.22).

18.3.3 Compressive Sensing and Recovery for Ill-Posed Problems

The objective in the sensing scenario for ill-posed problems is again to recover x, but now we only have access to Kx and K is supposed to be a linear (possibly ill-posed) and bounded operator between Hilbert spaces X and Y.

The data y are obtained by sensing Kx through $F_s : Y \to (\ell_2(\Lambda))^p$, i.e. $y = F_s Kx = F_s K F_a^* d$. Similarly to Lemma 18.1, we have the following result.

Lemma 18.2. *Assume for all $\lambda \in \Lambda$ the sampling functions $s_{1,\lambda}, \ldots, s_{p,\lambda}$ are chosen as in (18.14). Then $y = AF_{K^* v} F_a^* d = AF_v F_{Ka}^* d$.*

An ideal choice to guarantee recovery within the compressive sampling framework would be to ensure $F_{K^* v} F_a^* = F_v F_{Ka}^* = Id$. For normalized systems Φ_a and Φ_v and ill-posed operators K this is impossible to achieve. The simplest case is that we have systems Φ_a and Φ_v at our disposal that diagonalize K, i.e. $\langle Ka_{\ell,\lambda}, v_{\ell',\lambda'} \rangle = \kappa_{\ell,\lambda} \delta_{\lambda'\lambda} \delta_{\ell'\ell}$. One prominent example is the so-called wavelet-vaguelette decomposition with respect to K. If Φ_a and Φ_v diagonalize K, then the structure of the sensing operator is $TD : (\ell_2(\Lambda))^m \to (\ell_2(\Lambda))^p$, where $(TD)(\{d_{1,\lambda}\}_{\lambda\in\Lambda}, \ldots, \{d_{m,\lambda}\}_{\lambda\in\Lambda}) = (\{(AD_\lambda d_\lambda)^1\}_{\lambda\in\Lambda}, \ldots, \{(AD_\lambda d_\lambda)^p\}_{\lambda\in\Lambda})$, and D is defined by λ-dependant blocks D_λ of size $m \times m$, $D_\lambda = \mathrm{diag}(\kappa_{1,\lambda}, \kappa_{2,\lambda}, \ldots, \kappa_{m,\lambda})$.

Let us first consider the sensing problems for each individual label λ (which are m-dimensional recovery problems),

$$y_\lambda^\delta = AD_\lambda d_\lambda + z_\lambda \quad \text{with} \quad \|z_\lambda\| \le \delta \,. \tag{18.25}$$

Since K is ill-posed, the sensing matrix AD_λ obeys no longer the restricted isometry property. Therefore, we propose to minimize the stabilized constrained optimization problem

$$\min_{d_\lambda \in B(\ell_1, R)} \|y_\lambda^\delta - AD_\lambda d_\lambda\|_{\ell_2}^2 + \alpha \|d_\lambda\|_{\ell_2}^2 \,, \tag{18.26}$$

where $B(\ell_1, R) = \{d_\lambda \in \ell_2 : \|d_\lambda\|_{\ell_1} \le R\}$. Let us define $L^2 := D_\lambda A^* A D_\lambda + \alpha I$, if A fulfills the restricted isometry property (18.16), then the operator L obeys a restricted isometry condition of the following form,

$$(\kappa_{min}^2(1 - \delta_k) + \alpha)\|d_\lambda\|_{\ell_2}^2 \le \|Ld_\lambda\|_{\ell_2}^2 \le (\kappa_{max}^2(1 + \delta_k) + \alpha)\|d_\lambda\|_{\ell_2}^2 \,, \tag{18.27}$$

for all k-sparse vectors d_λ and where κ_{max} denotes the largest and κ_{min} the smallest eigenvalue of D_λ.

Theorem 18.9 (Finite dimensions). *Assume R is such that $d_\lambda \notin B(\ell_1, R)$ and that*

$$0 \le \delta_{2k} < \frac{(1 + \sqrt{2})\kappa_{min}^2 - \kappa_{max}^2 + \sqrt{2}\alpha}{(1 + \sqrt{2})\kappa_{min}^2 + \kappa_{max}^2} \,. \tag{18.28}$$

Then the minimizer d_λ^* of (18.26) satisfies

$$\|d_\lambda^* - d_\lambda\|_{\ell_2} \le C_0 k^{-1/2} \|d_\lambda^k - d_\lambda\|_{\ell_1} + C_1 \|L(d_\lambda^\dagger - d_\lambda)\|_{\ell_2} + C_2\delta + C_3\sqrt{\alpha}R\,, \tag{18.29}$$

where $d_\lambda^\dagger$ is the $B(\ell_1, R)$-best approximate solution, d_λ^k the best k-term approximation, and where the constants C_0, C_1, C_2, and C_3 are given explicitly.

As (18.28) serves as a condition for δ_{2k} and α at the same time, it turns out that the choice of α influences the choice of a suitable sensing matrix A and vice versa.

Let us now investigate the full infinite dimensional measurement model,

$$y^\delta = (TD)d + z \;\text{ with }\; \|z\|_{(\ell_2(\Lambda))^m} \le \delta\,.$$

We propose to solve the following optimization problem,

$$\min_{d \in B(\Psi_{1,2}, R)} \|y^\delta - (TD)d\|^2_{(\ell_2(\Lambda))^p} + \alpha \|d\|^2_{(\ell_2(\Lambda))^m}\,. \tag{18.30}$$

For the minimizing element the following error estimate hold true.

Theorem 18.10 (Infinite dimensions). *Assume R is such that $d \notin B(\Psi_{1,2}, R)$ and δ_{2k} is as in Theorem 18.9. Then the minimizer d^* of (18.30) satisfies*

$$\|d^* - d\|_{(\ell_2(\Lambda))^m} \le C_0 k^{-1/2} \Psi_{1,2}(d^k - d) + C_1 \|L(d^\dagger - d)\|_{(\ell_2(\Lambda))^m} + C_2\delta + C_3\sqrt{\alpha}R\,.$$

The minimizing elements can be iteratively approximated by

$$d_\lambda^{n+1} = \mathscr{P}\left(D_\lambda A^*(y_\lambda^\delta - AD_\lambda d_\lambda^n)\frac{\gamma^n}{C} + \left(1 - \frac{\alpha\gamma^n}{C}\right) d_\lambda^n\right)$$

for problem (18.26) and for the full infinite dimensional case by

$$d^{n+1} = \mathscr{P}\left(D^*T^*(y^\delta - TDd^n)\frac{\gamma^n}{C} + \left(1 - \frac{\alpha\gamma^n}{C}\right) d^n\right)\,.$$

The norm convergence is ensured by Proposition 18.4.

References

1. Adcock, B., Hansen, A.C.: Generalized sampling and infinite dimensional compressed sensing. Technical report NA2011/02, DAMTP, University of Cambridge (2012)
2. Adcock, B., Hansen, A.C., Herrholz, E., Teschke, G.: Generalized sampling, infinite-dimensional compressed sensing, and semi-random sampling for asymptotically incoherent dictionaries (2012, submitted)

3. Adcock, B., Hansen, A.C., Herrholz, E., Teschke, G.: Generalized sampling: extension to frames and inverse and ill-posed problems. Inverse Probl. **29**(1), 015,008 (2013)
4. Adcock, B., Hansen, A.C., Roman, B., Teschke, G.: Generalized sampling: stable reconstructions, inverse problems and compressed sensing over the continuum. Adv. Imaging Electron Phys. **182**, 1–51 (2008)
5. Bonesky, T., Bredies, K., Lorenz, D.A., Maass, P.: A generalized conditional gradient method for nonlinear operator equations with sparsity constraints. Inverse Probl. **23**, 2041–2058 (2007)
6. Bredies, K., Lorenz, D., Maass, P.: A generalized conditional gradient method and its connection to an iterative shrinkage method. Comput. Optim. Appl. **42**(2), 173–193 (2009)
7. Bredies, K., Lorenz, D.A.: Linear convergence of iterative soft-thresholding. J. Fourier Anal. Appl. **14**(5–6), 813–837 (2008)
8. Candès, E., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
9. Candès, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
10. Candès, E., Tao, T.: Near optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
11. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. C. R. Acad. Sci. **346**(9–10), 589–592 (2008)
12. Candès, E.J., Tao, T.: Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
13. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM Rev. **43**(1), 129–159 (2001)
14. Daubechies, I., Defrise, M., De Mol, C.: An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Commun. Pure Appl. Math. **57**(11), 1413–1457 (2004)
15. Daubechies, I., Teschke, G., Vese, L.: Iteratively solving linear inverse problems with general convex constraints. Inverse Probl. Imaging **1**(1), 29–46 (2007)
16. Daubechies, I., Teschke, G., Vese, L.: On some iterative concepts for image restoration. Adv. Imaging Electron Phys. **150**, 1–51 (2008)
17. Denis, L., Lorenz, D.A., Dennis, T.: Greedy solution of ill-posed problems: error bounds and exact inversion. Inverse Probl. **25**(11), 115,017 (24pp) (2009)
18. Donoho, D.: Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
19. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. Proc. Natl. Acad. Sci. **100**, 2197–2202 (2003)
20. Donoho, D.L., Hou, X.: Uncertainty priciples and ideal atomic decomposition. IEEE Trans. Inf. Theory **47**, 2845–2862 (2001)
21. Dossal, C., Mallat, S.: Sparse spike deconvolution with minimum scale. In: Proceedings of SPARS05, Rennes (2005)
22. Eldar, Y.C.: Compressed sensing of analog signals in shift-invariant spaces. IEEE Trans. Signal Process. **57**(8), 2986–2997 (2009)
23. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer, Dordrecht (1996)
24. Feuer, A., Nemirovski, A.: On sparse representations in unions of bases. IEEE Trans. Inf. Theory **49**, 1579–1581 (2003)
25. Fornasier, M.: Domain decomposition methods for linear inverse problems with sparsity constraints. Inverse Probl. **23**, 2505–2526 (2007)
26. Fornasier, M., Rauhut, H.: Recovery algorithms for vector valued data with joint sparsity constraint. SIAM J. Numer. Anal. **46**(2), 577–613 (2008)
27. Grasmair, M., Haltmeier, M., Scherzer, O.: Sparse regularization with ℓ^q penalty term. Inverse Probl. **24**(5), 055,020 (13pp) (2008)
28. Gribonval, R., Nielsen, M.: Beyond sparsity: recovering structured representations by ℓ^1 minimization and greedy algorithms. Adv. Comput. Math. **28**(1), 23–41 (2008)
29. Groetsch, C.W.: The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind. Pitman, Boston (1984)

30. Groetsch, C.W.: Inverse Problems in the Mathematical Sciences. Vieweg, Braunschweig (1993)
31. Herrholz, E., Teschke, G.: Compressive sensing principles and iterative sparse recovery for inverse and ill-posed problems. Inverse Probl. **26**, 125,012 (2010)
32. Klann, E., Kuhn, M., Lorenz, D.A., Maass, P., Thiele, H.: Shrinkage versus deconvolution. Inverse Probl. **23**, 2231–2248 (2007)
33. Laska, J., Kirolos, S., Duarte, M., Ragheb, T., Baraniuk, R., Massoud, Y.: Theory and implementation of an analog-to-information converter using random demodulation. In: Proceedings of IEEE International Symposium on Circuits and Systems (ISCAS), New Orleans (2007)
34. Lorenz, D.A.: Convergence rates and source conditions for Tikhonov regularization with sparsity constraints. J. Inverse Ill-Posed Probl. **16**(5), 463–478 (2008)
35. Lorenz, D.A., Schiffler, S., Trede, D.: Beyond convergence rates: exact inversion with Tikhonov regularization with sparsity constraints. Inverse Probl. **27**, 085,009 (2011)
36. Mallat, S., Davis, G., Zhang, Z.: Adaptive time-frequency decompositions. SPIE J. Opt. Eng. **33**(7), 2183–2191 (1994)
37. Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)
38. Mishali, M., Eldar, Y.C.: Blind multi-band signal reconstruction: compressed sensing for analog signals. IEEE Trans. Signal Process. **57**(3), 993–1009 (2009)
39. Morozov, V.A.: Methods for Solving Incorrectly Posed Problems. Springer, New York (1984)
40. Pati, Y., Rezaiifar, R., Krishnaprasad, P.: Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In: Proceedings of 27th Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, vol. 1, pp. 40–44 (1993)
41. Ramlau, R., Teschke, G.: A Tikhonov-based projection iteration for nonlinear ill-posed problems with sparsity constraints. Numer. Math. **104**(2), 177–203 (2006)
42. Rauhut, H., Schass, K., Vandergheynst, P.: Compressed sensing and redundant dictionaries. IEEE Trans. Inf. Theory **54**(5), 2210–2219 (2008)
43. Soulez, F., Denis, L., Fournier, C., Thiébaut, É., Goepfert, C.: Inverse problem approach for particle digital holography: accurate location based on local optimisation. J. Opt. Soc. Am. A **24**(4), 1164–1171 (2007)
44. Teschke, G.: Multi-frame representations in linear inverse problems with mixed multi-constraints. Appl. Computat. Harmon. Anal. **22**, 43–60 (2007)
45. Teschke, G., Borries, C.: Accelerated projected steepest descent method for nonlinear inverse problems with sparsity constraints. Inverse Probl. **26**, 025,007 (2010)
46. Tropp, J.A.: Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inf. Theory **50**(10), 2231–2242 (2004)

Chapter 19
Low-Rank Dynamics

Christian Lubich

Abstract This note reviews differential equations on manifolds of matrices or tensors of low rank. They serve to approximate, in a low-rank format, large time-dependent matrices and tensors that are either given explicitly via their increments or are unknown solutions of differential equations. Furthermore, low-rank differential equations are used in novel algorithms for eigenvalue optimisation, for instance in robust-stability problems.

19.1 Introduction

Low-rank approximation of hugh matrices and tensors is a basic model reduction technique in a wide variety of applications ranging from quantum physics to information retrieval. In this paper we review the low-rank approximation of *time-dependent* matrices and tensors, using differential equations to update low-rank approximations to given matrices or tensors, and to approximate solutions to matrix or tensor differential equations in a data-reduced, low-rank format. In Sect. 19.2 we recapitulate the Dirac–Frenkel time-dependent variational approximation principle in its abstract form. At every time instant, the time derivative is projected onto the tangent space of an approximation manifold at the current approximation. In Sect. 19.3 we consider the dynamical low-rank approximation of matrices, and in Sect. 19.4 we discuss a suitable numerical integrator for the corresponding differential equations on the low-rank manifold. In Sect. 19.5 we consider dynamical approximation of tensors in various formats: Tucker tensors, tensor trains, and hierarchical Tucker tensors. Section 19.6 considers briefly the MCTDH method of quantum dynamics, which uses the Dirac-Frenkel time-dependent variational principle in its original setting, the time-dependent Schrödinger equation. Finally, in Sect. 19.7 we consider differential equations for matrices of very low rank (rank 1, 2 or 4) that are used to solve optimisation problems for eigenvalues of structured

C. Lubich (✉)
University of Tübingen, Auf der Morgenstelle 10, 72076 Tübingen, Germany
e-mail: lubich@na.uni-tuebingen.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_19

or unstructured matrices, such as computing the distance to instability of a stable complex or real matrix and solving further matrix nearness problems.

19.2 Projecting onto the Tangent Space: The Dirac–Frenkel Time-Dependent Variational Approximation Principle

The dynamical low-rank approximations of matrices, tensors and multivariate functions as discussed below are instances of a general approximation principle that was first used and described, for a particular example, by Dirac [6] and Frenkel [7] in the early days of quantum mechanics. This time-dependent variational principle can be viewed, from a numerical analysis viewpoint, as a nonlinear Galerkin method in which time derivatives are projected onto the tangent space of an approximation manifold at the current approximation. In this preparatory section we give an abstract description.

Let $\mathscr{H}$ be a (real or complex, finite- or infinite-dimensional) Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$ given by $\|v\|^2 = \langle v, v \rangle$ for $v \in \mathscr{H}$. Let $\mathscr{M} \subset \mathscr{H}$ be a submanifold on which approximations to a time-dependent function $a(t) \in \mathscr{H}$, $0 \le t \le \bar{t}$, are restricted to lie. While a best approximation $x(t) \in \mathscr{M}$ to $a(t)$ on $\mathscr{M}$ is determined by the condition, pointwise in time,

$$x(t) \in \mathscr{M} \quad \text{such that} \quad \|x(t) - a(t)\| = \min, \tag{19.1}$$

we here determine an approximation $y(t) \in \mathscr{M}$ from the condition that its time derivative $\dot{y}(t) = (dy/dt)(t)$, which lies in the tangent space $T_{y(t)}\mathscr{M}$ of $\mathscr{M}$ at $y(t)$, should satisfy

$$\dot{y}(t) \in T_{y(t)}\mathscr{M} \quad \text{such that} \quad \|\dot{y}(t) - \dot{a}(t)\| = \min. \tag{19.2}$$

While the two conditions above look superficially similar, they are fundamentally different in that (19.1) is in general a non-linear, non-convex optimisation problem pointwise for every time t, whereas (19.2) requires a *linear* projection of $\dot{a}(t)$ onto the tangent space to determine $\dot{y}(t)$ and yields a differential equation on the approximation manifold $\mathscr{M}$. An equivalent condition is

$$\dot{y}(t) \in T_{y(t)}\mathscr{M} \quad \text{such that} \quad \langle \dot{y}(t) - \dot{a}(t), v \rangle = 0 \quad \forall\, v \in T_{y(t)}\mathscr{M}, \tag{19.3}$$

which can be viewed as a Galerkin condition on the state-dependent space $T_{y(t)}\mathscr{M}$. (In the case of a complex Hilbert space the minimisation condition in (19.2) yields the real part of the inner product in (19.3). This condition remains equivalent to (19.3) – not just the real part – if the tangent space is complex linear, i.e., with $v \in T_{y(t)}\mathscr{M}$ also $iv \in T_{y(t)}\mathscr{M}$.)

Let $P(y) : \mathscr{H} \to T_y\mathscr{M}$ denote the orthogonal projection onto the tangent space at $y \in \mathscr{M}$. Then, (19.2) and (19.3) can be written compactly as

$$\dot{y} = P(y)\dot{a}. \tag{19.4}$$

This differential equation on the approximation manifold $\mathscr{M}$ must be complemented with an initial value $y(0) \in \mathscr{M}$, ideally the best approximation $x(0) \in \mathscr{M}$. Solving this initial value problem yields the approximation $y(t) \in \mathscr{M}$ on the prescribed time interval. This approach can be viewed as a continuous-time updating technique for an approximation to $a(t)$ on $\mathscr{M}$ which uses only the increments $\dot{a}(t)$, starting from an initial approximation.

In contrast to (19.1), condition (19.2) (or equivalently, (19.3) or (19.4)) extends readily to the situation where $a(t) \in \mathscr{H}$ is not given explicitly, but is the unknown solution of a differential equation on $\mathscr{H}$,

$$\dot{a} = f(a), \quad \text{with given initial value } a(0).$$

In this situation an approximation $y(t) \in \mathscr{M}$ to $a(t) \in \mathscr{H}$ is determined, without prior computation of $a(t)$, from the following extension of (19.2),

$$\dot{y}(t) \in T_{y(t)}\mathscr{M} \quad \text{such that} \quad \|\dot{y}(t) - f(y(t))\| = \min, \tag{19.5}$$

or equivalently of (19.3),

$$\dot{y}(t) \in T_{y(t)}\mathscr{M} \quad \text{such that} \quad \langle \dot{y}(t) - f(y(t)), v\rangle = 0 \quad \forall\, v \in T_{y(t)}\mathscr{M}, \tag{19.6}$$

or equivalently of (19.4),

$$\dot{y} = P(y)f(y). \tag{19.7}$$

The above approach was first taken by Dirac [6] in a particular situation for approximating solutions of the multi-particle time-dependent Schrödinger equation by a method that is nowadays known as the time-dependent Hartree–Fock method. For further properties and uses of the time-dependent variational approximation principle in quantum dynamics we refer to Kramer and Saraceno [23] and Lubich [26]. Quasi-optimality results for the variational approximation were first shown in [25], in the context of the time-dependent Schrödinger equation.

In the following sections we consider in some detail the following cases of particular interest:

- Dynamical low-rank approximation of matrices: $\mathscr{H} = \mathbf{R}^{m\times n}$ and $\mathscr{M}$ is a manifold of matrices of fixed rank $r < \min\{m, n\}$ (Sects. 19.3 and 19.4);
- Dynamical low-rank approximation of tensors: $\mathscr{H} = \mathbf{R}^{n_1\times\cdots\times n_d}$ and $\mathscr{M}$ is a manifold of tensors of fixed rank in the Tucker format or tensor train format or hierarchical Tucker format (Sect. 19.5);

- The multi-configuration time-dependent Hartree (MCTDH) method of multi-particle quantum dynamics: $\mathscr{H} = L^2(\mathbf{R}^d)$, $\mathscr{M}$ is a manifold of multivariate functions of fixed rank in the Tucker format (or hierarchical Tucker format), and the differential equation to be approximated is the time-dependent Schrödinger equation (Sect. 19.6).

19.3 Dynamical Low-Rank Approximation of Matrices

In this section we follow the lines of Koch and Lubich [20] and present their setting and main results.

19.3.1 Rank-r Matrices and Their Tangent Matrices

Let $\mathscr{M}_r$ denote the manifold of real rank-r matrices of dimension $m \times n$. Every $Y \in \mathscr{M}_r$ can be written, in a non-unique way, as

$$Y = USV^T, \tag{19.8}$$

where $U \in \mathbf{R}^{m\times r}$ and $V \in \mathbf{R}^{n\times r}$ have orthonormal columns and $S \in \mathbf{R}^{r\times r}$ is nonsingular. The singular value decomposition yields S diagonal, but no special form of S is assumed in the following. As a substitute for the non-uniqueness in the decomposition (19.8), we use a unique decomposition in the tangent space. Every tangent matrix $\delta Y \in T_Y\mathscr{M}_r$ is of the form

$$\delta Y = \delta U S V^T + U \delta S V^T + U S \delta V^T, \tag{19.9}$$

where δS, δU and δV turn out to be uniquely determined by δY under the orthogonality constraints

$$U^T \delta U = 0, \quad V^T \delta V = 0. \tag{19.10}$$

It is found that

$$\begin{aligned} \delta S &= U^T \delta Y V, \\ \delta U &= (I - UU^T)\delta Y V S^{-1}, \\ \delta V &= (I - VV^T)\delta Y^T U S^{-T}. \end{aligned} \tag{19.11}$$

Formulas (19.9) and (19.11) establish an isomorphism between the subspace

$$\{(\delta S, \delta U, \delta V) \in \mathbf{R}^{r\times r} \times \mathbf{R}^{m\times r} \times \mathbf{R}^{n\times r} : U^T \delta U = 0,\ V^T \delta V = 0\}$$

and the tangent space $T_Y\mathscr{M}_r$.

19.3.2 Differential Equations for Dynamical Low-Rank Approximation

Given time-dependent $m \times n$ matrices $A(t)$, which depend in a continuously differentiable way on t, the minimisation condition (19.2) on the tangent space of the rank-r manifold $\mathscr{M}_r$ is equivalent to finding $\dot{Y} \in T_Y\mathscr{M}_r$ (we omit the argument t) satisfying

$$\langle \dot{Y} - \dot{A},\, \delta Y \rangle = 0 \quad \text{for all } \delta Y \in T_Y\mathscr{M}_r. \tag{19.12}$$

This formulation is used to derive differential equations for the factors in the representation (19.8).

Theorem 19.1. *For $Y = USV^T \in \mathscr{M}_r$ with nonsingular $S \in \mathbf{R}^{r\times r}$ and with $U \in \mathbf{R}^{m\times r}$ and $V \in \mathbf{R}^{n\times r}$ having orthonormal columns, condition (19.12) is equivalent to $\dot{Y} = \dot{U}SV^T + U\dot{S}V^T + US\dot{V}^T$, where*

$$\begin{aligned} \dot{S} &= U^T\dot{A}V \\ \dot{U} &= (I_m - UU^T)\dot{A}VS^{-1} \\ \dot{V} &= (I_n - VV^T)\dot{A}^T US^{-T}. \end{aligned} \tag{19.13}$$

When $A(t)$ is not given explicitly, but is the unknown solution of a matrix differential equation $\dot{A} = F(A)$, then $\dot{A}$ is replaced by $F(Y)$.

The differential equations for U, S, V are solved numerically; see Nonnenmacher and Lubich [32] for numerical experiments for applications ranging from the compression of time-varying term-document matrices and of series of images to the computation of blow-up in reaction-diffusion equations. A suitable integrator, which does not suffer from a possible ill-conditioning of S, is described in Sect. 19.4. That integrator starts from the formulation (19.4), viz.,

$$\dot{Y} = P(Y)\dot{A},$$

where $P(Y)$ is the orthogonal projection onto $T_Y\mathscr{M}_r$. Using the above result, in [20] an explicit expression for this tangent space projection at $Y = USV^T$ is found to be

$$P(Y)Z = ZVV^T - UU^TZVV^T + UU^TZ. \tag{19.14}$$

Since UU^T is the orthogonal projector onto the range $\mathscr{R}(Y)$ of $Y = USV^T$, and VV^T is the orthogonal projector onto the range $\mathscr{R}(Y^T)$, this formula can be rewritten as

$$P(Y)Z = ZP_{\mathscr{R}(Y^T)} - P_{\mathscr{R}(Y)}ZP_{\mathscr{R}(Y^T)} + P_{\mathscr{R}(Y)}Z. \tag{19.15}$$

19.3.3 Curvature Bounds

To obtain approximation estimates for the dynamical low-rank approximation, a key is to bound the curvature of the rank-r manifold $\mathscr{M}_r$ near a given matrix in $\mathscr{M}_r$. The following bounds are obtained in [20]. Here and in the following the matrix norm $\|\cdot\| = \|\cdot\|_F$ is the Frobenius norm, and $\|\cdot\|_2$ denotes the matrix norm induced by the Euclidean vector norm. We write $P^\perp(Y) = \mathrm{Id} - P(Y)$.

Lemma 19.1. *Let the rank-r matrix $X \in \mathscr{M}_r$ be such that its smallest non-zero singular value satisfies $\sigma_r(X) \geq \rho > 0$, and let $Y \in \mathscr{M}_r$ with $\|Y - X\| \leq \frac{1}{8}\rho$. Then, the following bounds hold: for all $B \in \mathbf{R}^{m\times n}$,*

$$\|\big(P(Y) - P(X)\big)B\| \leq 8\rho^{-1}\,\|Y - X\| \cdot \|B\|_2, \tag{19.16}$$

$$\|P^\perp(Y)(Y - X)\| \leq 4\rho^{-1}\,\|Y - X\|^2. \tag{19.17}$$

Note that the curvature bound is inversely proportional to the smallest non-zero singular value ρ. When the rank is increased in order to attain a better accuracy of approximation, the smallest nonzero singular value may become small, so that the opposing effects of better approximability and higher curvature in the approximation manifold need to be balanced carefully. This is a major challenge in the analysis of the approximation properties of the dynamical low-rank approximation.

19.3.4 Approximation Estimates

We state two results from [20]. The first result shows quasi-optimality over a time interval whose length is inversely proportional to a lower bound of the rth singular value of $A(t)$.

Theorem 19.2. *Suppose that a continuously differentiable best approximation $X(t) \in \mathscr{M}_r$ to $A(t)$ exists for $0 \leq t \leq \bar{t}$, and that $\|\dot{A}(t)\|_2 \leq \mu$ for $0 \leq t \leq \bar{t}$. Let the rth singular value of $A(t)$ have the lower bound $\sigma_r(A(t)) \geq \rho > 0$, and assume that the best-approximation error is bounded by $\|X(t) - A(t)\| \leq \frac{1}{16}\rho$ for $0 \leq t \leq \bar{t}$. Then, the approximation error of (19.12) with initial value $Y(0) = X(0)$ is bounded in the Frobenius norm by*

$$\|Y(t) - X(t)\| \leq 2\beta\, e^{\beta t} \int_0^t \|X(s) - A(s)\|\, ds \qquad \text{with } \beta = 8\mu\rho^{-1}$$

for $t \leq \bar{t}$ and as long as the right-hand side is bounded by $\frac{1}{8}\rho$.

Remarkably, small errors over longer time intervals are obtained even in cases of small singular values and when there is no gap in the singular values. Suppose that $A(t)$ can be written as

$$A(t) = X(t) + E(t), \qquad 0 \le t \le \bar{t}, \tag{19.18}$$

where $X(t) \in \mathscr{M}_r$ and the derivative of the remainder term is bounded by

$$\|\dot{E}(t)\| \le \varepsilon \tag{19.19}$$

with a small $\varepsilon > 0$. Suppose further that $X(t) \in \mathscr{M}_r$ with $\sigma_r(X(t)) \ge \rho > 0$ has a decomposition

$$X(t) = U_0(t) S_0(t) V_0(t)^T \qquad \text{for } 0 \le t \le \bar{t}, \tag{19.20}$$

with nonsingular $S_0(t) \in \mathbf{R}^{r\times r}$, and with $U_0(t) \in \mathbf{R}^{m\times r}$ and $V_0(t) \in \mathbf{R}^{n\times r}$ having orthogonal columns, such that the following bounds are valid for $0 \le t \le \bar{t}$:

$$\left\| \frac{d}{dt} S_0^{-1}(t) \right\|_2 \le c_1 \rho^{-1}, \qquad \|\dot{U}_0(t)\|_2 \le c_2, \quad \|\dot{V}_0(t)\|_2 \le c_2. \tag{19.21}$$

Under these conditions an $O(\varepsilon)$ error over times $O(1)$ is obtained even with $\rho \sim \varepsilon$.

Theorem 19.3. *Under the above conditions and for $\varepsilon \le c_0\rho$, the approximation error of the dynamical low-rank approximation (19.12) with initial value $Y(0) = X(0)$ is bounded by*

$$\|Y(t) - X(t)\| \le 2t\varepsilon \quad \text{for} \quad t \le t^*,$$

where $t^* \le \bar{t}$ depends only on c_0, c_1, and c_2.

We remark that both Theorems 19.2 and 19.3 rely heavily on Lemma 19.1.

19.4 A Projector-Splitting Integrator for Dynamical Low-Rank Approximation

To integrate the differential equations of dynamical low-rank approximation numerically, Lubich and Oseledets [27] propose a method that is based on splitting the projector (19.14). This method turns out to have remarkable robustness properties. It improves on a previously proposed integrator by Khoromskij, Oseledets and Schneider [18] which is based on splitting according to the differential equations for U, S, V.

It is observed in [27] that each of the differential equations for $Y = USV^T$ obtained by splitting the projector (19.15),

$$\dot{Y} = \dot{A} P_{\mathscr{R}(Y^T)}, \quad \dot{Y} = -P_{\mathscr{R}(Y)} \dot{A} P_{\mathscr{R}(Y^T)}, \quad \dot{Y} = P_{\mathscr{R}(Y)} \dot{A},$$

can be solved explicitly, and solving them in the indicated order over a time step yields the following algorithm. Given a factorisation (19.8) of the rank-r matrix $Y_0 = U_0 S_0 V_0^T$ and denoting the increment $\Delta A = A(t_1) - A(t_0)$, proceed as follows:

1. Set $K_1 = U_0 S_0 + \Delta A\, V_0$ and compute the factorisation $U_1 \hat{S}_1 = K_1$, with U_1 having orthonormal columns and with an $r \times r$ matrix $\hat{S}_1$ (e.g., using a QR factorisation).
2. Set $\tilde{S}_0 = \hat{S}_1 - U_1^T\, \Delta A\, V_0$.
3. Set $L_1 = V_0 \tilde{S}_0^T + \Delta A^T U_1$ and compute the factorisation $V_1 S_1^T = L_1$, with V_1 having orthonormal columns and with an $r \times r$ matrix S_1 (using QR).

The algorithm computes a factorisation of the rank-r matrix

$$Y_1 = U_1 S_1 V_1^T,$$

which is taken as an approximation to $Y(t_1)$.

This algorithm is of approximation order 1, and by composing it with the adjoint method that performs the above substeps in the reverse order, one obtains a second-order method. Using the standard technique of composing several time steps of suitably chosen lengths (see, e.g., [15, Chap. V.3]), one obtains methods of arbitrary order of approximation. The method is also readily extended to the low-rank approximation of the unknown solution of matrix differential equations $\dot{A} = F(A)$, in its simplest version replacing ΔA just by $\Delta t\, F(Y_0)$.

The above integrator has surprising robustness properties under ill-conditioning of the matrix factor S, whose inverse appears in the differential equations (19.13). While standard integrators such as Runge–Kutta methods applied to (19.13) break down as S approaches a singular matrix, the integrator given above has the following exactness property, which is proved in [27].

Theorem 19.4. *Suppose that $A(t)$ has rank at most r for all t. With the initial value $Y_0 = A(t_0)$, the above splitting algorithm is then exact: $Y_1 = A(t_1)$.*

When $A(t)$ is a perturbation of a matrix of rank $q < r$, the favourable behaviour of the above integrator persists. With a small parameter ε, assume that (with primes just as notational symbols),

$$A(t) = A'(t) + \varepsilon A''(t) \quad \text{with } \operatorname{rank}(A'(t)) = q < r,$$

where A' and A'' and their derivatives are bounded independently of ε. Suppose that the qth singular value of $A'(t)$ is larger than some positive constant and factorise

$$A'(t) = U'(t) S'(t) V'(t)^T$$

with $U'(t)$ and $V'(t)$ having q orthonormal columns and with an invertible $q \times q$ matrix $S'(t)$.

We apply the splitting integrator for the dynamical rank-r approximation of $A(t)$ with starting value

$$Y_0 = A'(0) + \varepsilon A_0'', \qquad \operatorname{rank}(Y_0) = r,$$

where A_0'' is bounded independently of ε. We compare the result of the rank-r algorithm with that of the rank-q algorithm starting from

$$\bar{Y}_0 = A'(0) + \varepsilon \bar{A}_0'', \qquad \operatorname{rank}(\bar{Y}_0) = q < r.$$

The following perturbation result is shown in [27].

Theorem 19.5. *In the above situation, let Y_n and $\bar{Y}_n$ denote the results of n steps of the splitting integrator for the rank-r approximation and rank-q approximation, respectively, applied with step size Δt. Then, as long as $n\Delta t \le \bar{t}$,*

$$\|Y_n - \bar{Y}_n\| \le C(\varepsilon + \Delta t),$$

where C is independent of n, Δt and ε (but depends on $\bar{t}$).

By the standard error estimates of splitting methods, the integration error of the rank-q approximation is $\bar{Y}_n - \bar{Y}(t_n) = O(\Delta t)$, uniformly in ε. Furthermore, it follows from the over-approximation lemma in [20] that the difference of the rank-r and rank-q approximations is bounded by $Y(t) - \bar{Y}(t) = O(\varepsilon)$.

The above result is in marked contrast to standard integrators, such as explicit or implicit Runge–Kutta methods, which break down as $\varepsilon \to 0$. Numerical experiments in [27] illustrate the favourable behaviour of the projector-splitting integrator.

19.5 Dynamical Low-Rank Approximation of Tensors

In this section the setting is that of Sect. 19.2 in the space of tensors $\mathscr{H} = \mathbf{R}^{n_1 \times \cdots \times n_d}$, with the inner product of two tensors given as the Euclidean inner product of the vectors that carry the entries of the tensors. The norm of a tensor is the Euclidean norm of the corresponding vector. We consider various approximation manifolds $\mathscr{M} \subset \mathscr{H}$: the manifolds of tensors of fixed rank in the Tucker format, in the tensor train format and the hierarchical Tucker format.

19.5.1 Dynamical Tensor Approximation in the Tucker Format

Here we look for an approximation of time-dependent tensors

$$A(t) \in \mathbf{R}^{n_1 \times \ldots n_d} \quad \text{with entries } A(k_1, \ldots, k_d; t)$$

by tensors $Y(t) \in \mathbf{R}^{n_1 \times \dots n_d}$ of the form

$$Y(k_1, \dots, k_d; t) = \sum_{j_1=1}^{r_1} \cdots \sum_{j_d=1}^{r_d} s_{j_1,\dots,j_d}(t)\, u_{j_1}^{(1)}(k_1; t) \dots u_{j_d}^{(d)}(k_d; t),$$

where for each $i = 1, \dots, d$, we have $r_i < n_i$ (and typically $r_i \ll n_i$). For a fixed t, this is called a tensor in the *Tucker format of multilinear rank* $r = (r_1, \dots, r_d)$ (see [12, 22]) if the following conditions are satisfied for each $i = 1, \dots, d$:

- The ith matrix unfolding of the core tensor $S(t) \in \mathbf{R}^{r_1 \times \cdots \times r_d}$, which aligns all entries corresponding to the subscript j_i (with $1 \le j_i \le r_i$) lexicographically in the j_i-th row, has full rank r_i.
- The vectors $u_1^{(i)}(t), \dots, u_{r_i}^{(i)}(t) \in \mathbf{R}^{n_i}$ are mutually orthonormal.

The set of all tensors in the Tucker format of rank $r = (r_1, \dots, r_d)$ is a manifold, which we denote again by $\mathscr{M}_r$. In dimension $d = 2$, tensors in the Tucker format of rank (r, r) correspond to matrices of rank r in the factorisation (19.8).

The dynamical approximation of tensors in the Tucker format has been studied in Koch and Lubich [21]. The theory extends that for the matrix case described in Sect. 19.3. To describe the results of that paper, it is convenient to use the shorthand notation (here we omit the argument t)

$$Y = S \times_1 U_1 \times_2 U_2 \cdots \times_d U_d = S \times_{i=1}^{d} U_i,$$

where U_i is the $n_i \times r_i$ matrix with columns $u_{j_i}^{(i)}$ $(j_i = 1, \dots, r_i)$.

Tangent tensors in $T_Y \mathscr{M}_r$ are of the form

$$\delta Y = \delta S \times_{i=1}^{d} U_i + \sum_{i=1}^{d} S \times_i \delta U_i \times_{\substack{l=1 \\ l \neq i}}^{d} U_l,$$

where δS and δU_i are shown to be determined uniquely under the gauge conditions

$$U_i^T \delta U_i = 0 \quad (i = 1, \dots, d).$$

Then, δS and δU_i are given by the following formulae (cf. (19.11)):

$$\begin{aligned} \delta S &= \delta Y \times_{i=1}^{d} U_i^T \\ \delta U_i &= (I_{n_i} - U_i U_i^T) \left[\delta Y \times_{l \neq i} U_l^T\right]_{(i)} S_{(i)}^{\dagger}, \end{aligned} \tag{19.22}$$

where the subscript (i) denotes the ith matrix unfolding and $S_{(i)}^{\dagger} = S_{(i)}^T (S_{(i)} S_{(i)}^T)^{-1}$ is the pseudo-inverse.

A similar form is taken by the differential equations for $S(t)$ and $U_i(t)$ in the factorisation of $Y(t) \in \mathscr{M}_r$ determined from the projection of the time derivative onto the tangent space,

$$\langle \dot{Y} - \dot{A}, \delta Y \rangle = 0 \quad \text{for all } \delta Y \in T_Y \mathscr{M}_r. \tag{19.23}$$

Theorem 19.6. *For a tensor in the Tucker format $Y = S \times_{i=1}^{d} U_i$, condition (19.23) is equivalent to*

$$\dot{Y} = \dot{S} \times_{i=1}^{d} U_i + \sum_{i=1}^{d} S \times_i \dot{U}_i \times_{\substack{l=1 \\ l \neq i}}^{d} U_l \,, \tag{19.24}$$

where the factors in the decomposition satisfy the system of differential equations

$$\begin{aligned} \dot{S} &= \dot{A} \times_{i=1}^{d} U_i^T \\ \dot{U}_i &= (I_{n_i} - U_i U_i^T) \left[\dot{A} \times_{l \neq i} U_l^T \right]_{(i)} S_{(i)}^{\dagger}. \end{aligned} \tag{19.25}$$

It is shown in [21] that the curvature bound of Lemma 19.1 extends to tensors in the Tucker format of rank r. The corresponding bounds are very similar to those of the matrix case, now with ρ denoting a lower bound of the nonzero singular values of all the matricisations $S_{(i)}$ of the core tensor. With the curvature bounds at hand, the approximation results of Theorems 19.2 and 19.3 are extended to the Tucker tensor case. We omit the precise formulations of these results and refer to [21] for the details.

19.5.2 Dynamical Tensor Approximation in the Tensor Train Format and Hierarchical Tucker Format

In the Tucker format, the memory requirements still grow exponentially with the dimension d. In the last decade, tensor formats with a substantially reduced growth of data with the dimension have been developed both in mathematics and physics. The book by Hackbusch [12] gives an excellent account of the various tensor formats.

A tensor in the *tensor train* (TT) format, or a *matrix product state* in the terminology of physics, has entries of the form

$$Y(k_1, \ldots, k_d) = G_1(k_1) \ldots G_d(k_d),$$

where the $G(k_i)$ are $r_{i-1} \times r_i$ matrices, with $r_0 = r_d = 1$. Matrix product states have become very popular in quantum physics in the last decade after it was realised that they are an appropriate framework for algorithms of renormalisation group theory; see Verstraete, Murg and Cirac [36] and Schollwöck [34]. We further refer to Haegeman, Osborne and Verstraete [14] in a time-dependent setting. The tensor train format was independently proposed and studied by Oseledets [33] in the mathematical literature.

The set of tensor trains for which the stacked matrices $(G_i(1), \dots, G_i(n_i))$ and $(G_i(1)^T, \dots, G_i(n_i)^T)$ are of full rank r_{i-1} and r_i, respectively, form a manifold. Properties of this TT manifold and of its tangent space are studied in Holtz, Rohwedder and Schneider [16].

A tensor in the *hierarchical Tucker* (HT) format is built up using a binary tree with the following data:

- For each leaf $i = 1, \dots, d$, there are r_i orthonormal vectors $u_1^{(i)}, \dots, u_{r_i}^{(i)} \in \mathbf{R}^{n_i}$.
- For each inner node τ of the tree with children nodes τ_1 and τ_2, there is a 3-tensor $B^\tau = (b^\tau_{j,j_1,j_2})$ of dimension $r_\tau \times r_{\tau_1} \times r_{\tau_2}$.

Define $Y^i_{j_i} = u^{(i)}_{j_i}$ for the leaves, and for an inner node τ with children τ_1 and τ_2 set

$$Y^\tau_j = \sum_{j_1,j_2} b^\tau_{j,j_1,j_2}\, Y^{\tau_1}_{j_1} \otimes Y^{\tau_2}_{j_2},$$

descending the tree from the leaves to the root. To ensure that Y^τ_j for $j = 1, \dots, r_\tau$ remain orthonormal at each node τ except the root, it is further required that the transfer tensors B^τ satisfy, in their first matrix unfolding, the orthonormality condition $B^\tau_{(1)}(B^\tau_{(1)})^T = I_{r_\tau}$.

Hierarchical Tucker tensors were first described and used in the chemical physics literature [2, 37] and are proposed and thoroughly discussed in the mathematical literature by Hackbusch and Kühn [13]; see also Lubich [26, p.45] for a brief account. The set of HT tensors for which all the transfer 3-tensors are of full Tucker rank, forms a manifold that has been studied by Uschmajew and Vandereycken [35].

The dynamical approximation in the HT and TT formats has very recently been studied independently by Arnold and Jahnke [1] and Lubich, Rohwedder, Schneider and Vandereycken [29]. These papers give the HT versions of the differential equations for dynamical approximation and derive curvature bounds similar to Lemma 19.1, with ρ now a lower bound of the nonzero singular values of matricisations of the transfer 3-tensors. These bounds allow one to derive approximation results similar to those known for the matrix and Tucker tensor cases.

The projector-splitting integrator of Sect. 19.5 is extended to the TT format by Lubich, Oseledets and Vandereycken [28], with similarly favourable properties as in the matrix case, and it can also be extended to the HT format (B. Vandereycken, personal communication).

19.6 The MCTDH Method for Quantum Dynamics

Much of the mathematical work on dynamical low-rank approximation was initially motivated by a model reduction method in molecular quantum dynamics, the *multi-configuration time-dependent Hartree* (MCTDH) method that was proposed by Meyer, Manthe and Cederbaum [30] and has developed into the standard tool for

accurately computing the quantum dynamics of small molecules; see the book by Meyer, Gatti and Worth [31].

The MCTDH method uses the Dirac–Frenkel time-dependent variational principle on the Hilbert space $\mathscr{H} = L^2(\mathbf{R}^d)$ for the Schrödinger equation

$$i\frac{\partial \psi}{\partial t} = -\Delta\psi + V\psi, \qquad x \in \mathbf{R}^d,\ t > 0,$$

with an approximation manifold $\mathscr{M}$ that consists of functions

$$u(x_1, \dots, x_d) = \sum_{j_1=1}^{r_1} \dots \sum_{j_d=1}^{r_d} a_{j_1 \dots j_d}\, \phi_{j_1}^{(1)}(x_1) \cdot \ldots \cdot \phi_{j_d}^{(d)}(x_d)$$

with orthonormal single-particle functions $\phi_1^{(i)}, \dots, \phi_{r_i}^{(i)}$ and a core tensor $(a_{j_1 \dots j_d})$ of full multilinear rank $(r_1, \dots, r_N)$. This is the Tucker format in an infinite-dimensional setting.

The Dirac–Frenkel time-dependent variational principle yields a nonlinearly coupled system of ordinary differential equations for the core tensor and low-dimensional Schrödinger equations for the single-particle functions. Well-posedness and regularity properties for this coupled system have been derived in Koch and Lubich [19]. A variational splitting integrator, which is explicit and unconditionally stable, has been proposed and analysed in [24].

An interesting theoretical question concerns error bounds of the MCTDH approximation with respect to the solution of the Schrödinger equation and convergence as the ranks $r_i \to \infty$. A quasi-optimality result of the MCTDH approximation, which bounds the MCTDH error in terms of the best-approximation error on the MCTDH manifold $\mathscr{M}$ given above, was shown for *fixed* ranks r_i in [25]. However, the length of the time intervals on which the estimates are meaningful, shrinks to 0 as $r_i \to \infty$, because the curvature bounds increase beyond any limit. While a naive expectation might suggest that one gets an ever better approximation as more and more terms are added, there are two obstructions:

- The approximation properties of the basis of tensor products of the time-dependent single-particle functions, which are themselves solution of a nonlinear system of partial differential equations, are not known a priori.
- As more terms are added, more and more *nearly irrelevant* terms are added that lead to small singular values of matrix unfoldings of the core tensor.

Conte and Lubich [5] derive an error bound that nevertheless yields a small error over a fixed time interval even with small singular values of the matricisations of the core tensor, in the spirit of Theorem 19.3. We refer to [5] for the precise assumptions and statement of the result.

19.7 Low-Rank Differential Equations for Structured Matrix Nearness Problems

Beyond the approximation of hugh time-dependent matrices and tensors, differential equations on low-rank manifolds have recently shown up unexpectedly in optimisation problems for eigenvalues of large structured or unstructured matrices, of which the following is a typical example:

> Given a matrix with all eigenvalues in the left complex half-plane, find a nearest matrix with an eigenvalue on the imaginary axis.

This asks for the distance to instability of a stable matrix; see Byers [4] and Burke, Lewis and Overton [3]. The problem can be posed for general complex matrices or be restricted to real matrices or companion matrices or non-positive matrices or matrices with some other structure. The notion of "nearest" depends on the matrix norm chosen, usually the matrix 2-norm or the Frobenius norm.

The above problem is closely related to finding extremal (here: rightmost) points in pseudospectra: for $\varepsilon > 0$, a chosen norm $\|\cdot\|$, a set of square matrices $\mathscr{S}$, and a matrix $A \in \mathscr{S}$, the (structured) ε-pseudospectrum is the set (cf. [17])

$$\Lambda_\varepsilon(A, \|\cdot\|, \mathscr{S}) = \{\lambda \in \mathbf{C} : \lambda \text{ is an eigenvalue of some matrix } B \in \mathscr{S} \text{ with } \|B - A\| \le \varepsilon\}.$$

The approach taken in Guglielmi and Lubich [8,9] for complex and for real matrices is to determine matrices $E(t)$ of unit norm such that the rightmost eigenvalue of $A + \varepsilon E(t)$ increases monotonically with t and tends to a (locally) rightmost point in the complex or real ε-pseudospectrum as $t \to \infty$. It turns out that in the extremal point, the extremiser E_* is of rank 1 in the complex case, and of rank at most 2 in the real case. This motivates to search for a differential equation on the manifolds of rank-1 and rank-2 matrices, respectively, which has the extremiser as a stationary point.

For example, in the real case and for the matrix 2-norm, such a differential equation is derived in [9] for a rank-2 matrix $E(t)$ with both nonzero singular values equal to 1. This is factorised as $E = UQV^T$ with $U, V \in \mathbf{R}^{n\times 2}$ having orthogonal columns and with an orthogonal 2×2 matrix Q. The differential equations for U, V, Q read

$$\begin{aligned} \dot U &= (I - UU^T)XY^T VQ^T \\ \dot V &= (I - VV^T)YX^T UQ \\ \dot Q &= Q\ \mathrm{skew}(Q^T U^T XY^{TV}). \end{aligned} \tag{19.26}$$

Here, X and Y are $n \times 2$ matrices containing the real and imaginary parts of the left and right eigenvectors x and y, of unit norm and with $x^* y > 0$, corresponding to a non-real eigenvalue λ of $A + \varepsilon E$ with $E = UQV^T$. It is shown that $\mathrm{Re}\,\lambda(t)$

increases with growing t and that $\lambda(t)$ runs, for $t \to \infty$, into a point on the boundary of the real ε-pseudospectrum that has a tangent parallel to the imaginary axis. This system of low-rank differential equations can be solved efficiently also for large sparse matrices A, since the differential equation requires only the leading eigenvalue of the low-rank perturbation $A+\varepsilon E(t)$ and its left and right eigenvectors.

To determine the distance to instability and the associated extremising matrix, one then optimises over ε until the leading eigenvalue hits the imaginary axis.

In a similar vein, matrix nearness problems for Hamiltonian and symplectic matrices are approached via rank-2 and rank-4 differential equations in Guglielmi, Kressner and Lubich [10, 11].

References

1. Arnold, A., Jahnke, T.: On the approximation of high-dimensional differential equations in the hierarchical Tucker format. BIT (2013). doi:10.1007/s10543-013-0444-2
2. Beck, M.H., Jäckle, A., Worth, G.A., Meyer, H.-D.: The multiconfiguration time-dependent Hartree (MCTDH) method: a highly efficient algorithm for propagating wavepackets. Phys. Rep. **324**, 1–105 (2000)
3. Burke, J.V., Lewis, A.S., Overton, M.L.: Optimization and pseudospectra, with applications to robust stability. SIAM J. Matrix Anal. Appl. **25**(1), 80–104 (2003)
4. Byers, R.: A bisection method for measuring the distance of a stable matrix to the unstable matrices. SIAM J. Sci. Stat. Comput. **9**, 875–881 (1988)
5. Conte, D., Lubich, C.: An error analysis of the multi-configuration time-dependent Hartree method of quantum dynamics. ESAIM M2AN **44**, 759–780 (2010)
6. Dirac, P.A.M.: Note on exchange phenomena in the Thomas atom. Proc. Camb. Philos. Soc. **26**, 376–385 (1930)
7. Frenkel, J.: Wave Mechanics. Advanced General Theory. Clarendon Press, Oxford (1934)
8. Guglielmi, N., Lubich, C.: Differential equations for roaming pseudospectra: paths to extremal points and boundary tracking. SIAM J. Numer. Anal. **49**, 1194–1209 (2011) and Erratum/addendum. SIAM J. Numer. Anal. **50**, 977–981 (2012)
9. Guglielmi, N., Lubich, C.: Low-rank dynamics for computing extremal points of real pseudospectra. SIAM J. Matrix Anal. Appl. **34**, 40–66 (2013)
10. Guglielmi, N., Kressner, D., Lubich, C.: Low-rank differential equations for Hamiltonian matrix nearness problems. Technical report (2013)
11. Guglielmi, N., Kressner, D., Lubich, C.: Computing extremal points of symplectic pseudospectra and solving symplectic matrix nearness problems. Technical report (2013)
12. Hackbusch, W.: Tensor Spaces and Numerical Tensor Calculus. Springer, Berlin (2012)
13. Hackbusch, W., Kühn, S.: A new scheme for the tensor representation. J. Fourier Anal. Appl. **15**, 706–722 (2009)
14. Haegeman, J., Osborne, T., Verstraete, F.: Post-matrix product state methods: to tangent space and beyond. Phys. Rev. B **88**, 075133 (2013)
15. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations. Springer Series in Computational Mathematics, vol. 31, 2nd edn. Springer, Berlin (2006)
16. Holtz, S., Rohwedder, T., Schneider, R.: On manifolds of tensors of fixed TT-rank. Numer. Math. **120**, 701–731 (2012)
17. Karow, M., Kokiopoulou, E., Kressner, D.: On the computation of structured singular values and pseudospectra. Syst. Control Lett. **59**, 122–129 (2010)

18. Khoromskij, B., Oseledets, I., Schneider, R.: Efficient time-stepping scheme for dynamics on TT-manifolds. Report 24/2012, Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig (2012)
19. Koch, O., Lubich, C.: Regularity of the multi-configuration time-dependent Hartree approximation in quantum molecular dynamics. ESAIM M2AN **41**, 315–332 (2007)
20. Koch, O., Lubich, C.: Dynamical low rank approximation. SIAM J. Matrix Anal. Appl. **29**, 434–454 (2007)
21. Koch, O., Lubich, C.: Dynamical tensor approximation. SIAM J. Matrix Anal. Appl. **31**, 2360–2375 (2010)
22. Kolda, T.G., Bader, B.W.: Tensor decompositions and applications. SIAM Rev. **51**, 455–500 (2009)
23. Kramer, P., Saraceno, M.: Geometry of the time-dependent variational principle in quantum mechanics. Lecture Notes in Physics, vol. 140. Springer, Berlin (1981)
24. Lubich, C.: A variational splitting integrator for quantum molecular dynamics. Appl. Numer. Math. **48**, 355–368 (2004)
25. Lubich, C.: On variational approximations in quantum molecular dynamics. Math. Comput. **74**, 765–779 (2005)
26. Lubich, C.: From Quantum to Classical Molecular Dynamics. Reduced Models and Numerical Analysis. European Mathematical Society, Zurich (2008)
27. Lubich, C., Oseledets, I.: A projector-splitting integrator for dynamical low-rank approximation. BIT (2013). doi:10.1007/s10543-013-0454-0
28. Lubich, C., Oseledets, I., Vandereycken, B.: Time integration of tensor trains (in preparation). E-print arXiv:1407.2042 [math.NA]
29. Lubich, C., Rohwedder, T., Schneider, R., Vandereycken, B.: Dynamical approximation by hierarchical Tucker and tensor-train tensors. SIAM J. Matrix Anal. Appl. **34**, 470–494 (2013)
30. Meyer, H.-D., Manthe, U., Cederbaum, L.S.: The multi-configurational time-dependent Hartree approach. Chem. Phys. Lett. **165**, 73–78 (1990)
31. Meyer, H.-D., Gatti, F., Worth, G.A. (eds.): Multidimensional Quantum Dynamics: MCTDH Theory and Applications. Wiley, New York (2009)
32. Nonnenmacher, A., Lubich, C.: Dynamical low-rank approximation: applications and numerical experiments. Math. Comput. Simul. **79**, 1346–1357 (2008)
33. Oseledets, I.V.: Tensor-train decomposition. SIAM J. Sci. Comput. **33**, 2295–2317 (2011)
34. Schollwöck, U.: The density-matrix renormalization group in the age of matrix product states. Ann. Phys. **326**, 96–192 (2011)
35. Uschmajew, A., Vandereycken, B.: The geometry of algorithms using hierarchical tensors. Linear Algebra Appl. **439**, 133–166 (2013)
36. Verstraete, F., Murg, V., Cirac, V.I.: Matrix product states, projected entangled pair states, and variational renormalization group methods for quantum spin systems. Adv. Phys. **57**, 143–224 (2008)
37. Wang, H., Thoss, M.: Multilayer formulation of the multiconfiguration time-dependent Hartree theory. J. Chem. Phys. **119**, 1289–1299 (2003)

Chapter 20
Computation of Expectations by Markov Chain Monte Carlo Methods

Erich Novak and Daniel Rudolf

Abstract Markov chain Monte Carlo (MCMC) methods are a very versatile and widely used tool to compute integrals and expectations. In this short survey we focus on error bounds, rules for choosing the burn in, high dimensional problems and tractability versus curse of dimension.

20.1 Introduction

Consider the following example. We want to compute

$$\mathbb{E}_G(f) = \frac{1}{\mathrm{vol}_d(G)} \int_G f(x)\,\mathrm{d}x,$$

where f belongs to some class of functions and G belongs to some class of sets. We assume that $G \subset \mathbb{R}^d$ is measurable with $0 < \mathrm{vol}_d(G) < \infty$, where vol_d denotes the Lebesgue measure. Thus, we want to compute the expected value of f with respect to the uniform distribution on G.

The input (f, G) is given by an oracle: For $x \in G$ we can compute $f(x)$ and G is given by a membership oracle, i.e. we are able to check whether any $x \in \mathbb{R}^d$ is in G or not. We always assume that G is convex and will work with the class

$$\mathscr{G}_{r,d} = \{G \subset \mathbb{R}^d : G \text{ is convex},\ B_d \subset G \subset rB_d\}, \tag{20.1}$$

where $r \geq 1$ and $rB_d = \{x \in \mathbb{R}^d : |x| \leq r\}$ is the Euclidean ball with radius r.

A first approach might be a simple acceptance/rejection method. The idea is to generate a point in rB_d according to the uniform distribution and if it is in G it is accepted, otherwise it is rejected. If $x_1, \ldots, x_n \in G$ are the accepted points then we output the mean value of the $f(x_i)$. However, this method does not work reasonably since the acceptance probability can be extremely small, it can be r^{-d}.

E. Novak (✉) • D. Rudolf
Friedrich Schiller University of Jena, Ernst-Abbe-Platz 2, 07743 Jena, Germany
e-mail: erich.novak@uni-jena.de; daniel.rudolf@uni-jena.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_20

It seems that all known efficient algorithms for this problem use Markov chains. The idea is to find a sampling procedure that approximates a sample with respect to the uniform distribution in G. More precisely, we run a Markov chain to approximate the uniform distribution for any $G \in \mathscr{G}_{r,d}$. Let $X_1, X_2, \dots, X_{n+n_0}$ be the first $n + n_0$ steps of such a Markov chain. Then

$$S_{n,n_0}(f, G) = \frac{1}{n} \sum_{j=1}^{n} f(X_{j+n_0})$$

is an approximation of $\mathbb{E}_G(f)$. The additional parameter n_0 is called burn-in and, roughly spoken, is the number of steps of the Markov chain to get close to the uniform distribution.

20.2 Approximation of Expectations by MCMC

20.2.1 Preliminaries

We provide the basics of Markov chains. For further reading we refer to the paper [14] of Roberts and Rosenthal which surveys various results about Markov chains on general state spaces.

A Markov chain is a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ which satisfies the Markov property. For $i \in \mathbb{N}$, the conditional distribution of X_{i+1} depends only on X_i and not on $(X_1, \dots, X_{i-1})$,

$$\mathbb{P}(X_{i+1} \in A \mid X_1, \dots, X_i) = \mathbb{P}(X_{i+1} \in A \mid X_i).$$

By $\mathscr{B}(G)$ we denote the Borel σ-algebra of G. Let ν be a distribution on $(G, \mathscr{B}(G))$ and let $K\colon G \times \mathscr{B}(G) \to [0, 1]$ be a *transition kernel*, i.e. $K(x, \cdot)$ is a probability measure for each $x \in G$ and $K(\cdot, A)$ is a $\mathscr{B}(G)$-measurable real-valued function for each $A \in \mathscr{B}(G)$. A transition kernel and a distribution ν give rise to a Markov chain $(X_n)_{n\in\mathbb{N}}$ in the following way. Assume that the distribution of X_1 is given by ν. Then, for $i \geq 2$ and a given $X_{i-1} = x_{i-1}$, we have X_i with distribution $K(x_{i-1}, \cdot)$, that is, for all $A \in \mathscr{B}(G)$, the conditional probability that $X_i \in A$ is given by $K(x_{i-1}, A)$. We call such a sequence of random variables a Markov chain with transition kernel K and initial distribution ν.

In the whole paper we only consider Markov chains with *reversible* transition kernel, we assume that there exists a probability measure π on $\mathscr{B}(G)$ such that

$$\int_A K(x, B)\, \pi(\mathrm{d}x) = \int_B K(x, A)\, \pi(\mathrm{d}x), \quad A, B \in \mathscr{B}(G).$$

In particular any such π is a stationary distribution of K, i.e.,

$$\pi(A) = \int_G K(x, A)\, \pi(\mathrm{d}x), \quad A \in \mathscr{B}(G).$$

Further, the transition kernel induces an operator on functions and an operator on measures given by

$$Pf(x) = \int_G f(y)\, K(x, \mathrm{d}x), \quad \text{and} \quad \nu P(A) = \int_G K(x, A)\, \nu(\mathrm{d}x),$$

where f is π-integrable and ν is absolutely continuous with respect to π. One has

$$\mathbb{E}[f(X_n) \mid X_1 = x] = P^{n-1} f(x) \quad \text{and} \quad \mathbb{P}_\nu(X_n \in A) = \nu P^{n-1}(A),$$

for $x \in G$, $A \in \mathscr{B}(G)$ and $n \in \mathbb{N}$, where ν in $\mathbb{P}_\nu$ indicates that X_1 has distribution ν. By the reversibility with respect to π we have $\frac{d(\nu P)}{d\pi}(x) = P(\frac{d\nu}{d\pi})(x)$, where $\frac{d\nu}{d\pi}$ denotes the density of ν with respect to π.

Further, for $p \in [1, \infty)$ let $L_p = L_p(\pi)$ be the space of measurable functions $f: G \to \mathbb{R}$ which satisfy

$$\|f\|_p = \left(\int_G |f(x)|^p \pi(\mathrm{d}x) \right)^{1/p} < \infty.$$

The operator $P: L_p \to L_p$ is linear and bounded and by the reversibility $P: L_2 \to L_2$ is self-adjoint.

The goal is to quantify the speed of convergence, if it converges at all, of νP^n to π for increasing $n \in \mathbb{N}$. For this we use the *total variation distance* between two probability measures ν, μ on $(G, \mathscr{B}(G))$ given by

$$\|\nu - \mu\|_{\mathrm{tv}} = \sup_{A \in \mathscr{B}(G)} |\nu(A) - \mu(A)|.$$

It is helpful to consider the total variation distance as an L_1-norm, see for example [14, Proposition 3, p. 28].

Lemma 20.1. *Assume the probability measures ν, μ have densities $\frac{d\nu}{d\pi}, \frac{d\mu}{d\pi} \in L_1$, then $\|\nu - \mu\|_{\mathrm{tv}} = \frac{1}{2} \left\| \frac{d\nu}{d\pi} - \frac{d\mu}{d\pi} \right\|_1$.*

Now we ask for an upper bound of $\|\nu P^n - \pi\|_{\mathrm{tv}}$.

Lemma 20.2. *Let ν be a probability measure on $(G, \mathscr{B}(G))$ with $\frac{d\nu}{d\pi} \in L_1$ and let $S(f) = \int_G f(x)\, \pi(\mathrm{d}x)$. Then, for any $n \in \mathbb{N}$ holds*

$$\|\nu P^n - \pi\|_{\mathrm{tv}} \le \|P^n - S\|_{L_1 \to L_1} \frac{1}{2} \left\| \frac{d\nu}{d\pi} - 1 \right\|_1 \le \|P^n - S\|_{L_1 \to L_1}$$

and

$$\|\nu P^n - \pi\|_{\mathrm{tv}} \le \|P^n - S\|_{L_2 \to L_2} \frac{1}{2} \left\| \frac{d\nu}{d\pi} - 1 \right\|_2 .$$

Proof. By Lemma 20.1, by $P^n 1 = 1$ and by the reversibility, in particular $\frac{d(\nu P^n)}{d\pi}(x) = P^n(\frac{d\nu}{d\pi})(x)$, we have

$$2\,\|\nu P^n - \pi\|_{\mathrm{tv}} = \left\| \frac{d(\nu P^n)}{d\pi} - 1 \right\|_1 = \left\| P^n \left(\frac{d\nu}{d\pi} - 1 \right) \right\|_1 = \left\| (P^n - S) \left(\frac{d\nu}{d\pi} - 1 \right) \right\|_1 .$$

Note that the last equality comes from $S(\frac{d\nu}{d\pi} - 1) = 0$.

Observe that for $\nu = \pi$ the left-hand side and also the right-hand side of the estimates are zero.

Let us consider $\|P^n - S\|_{L_2 \to L_2}$. Because of the reversibility with respect to π we obtain the following, see for example [16, Lemma 3.16, p. 45].

Lemma 20.3. *For $n \in \mathbb{N}$ we have*

$$\|P^n - S\|_{L_2 \to L_2} = \|(P - S)^n\|_{L_2 \to L_2} = \|P - S\|^n_{L_2 \to L_2} .$$

The last two lemmata motivate the following two convergence properties of transition kernels.

Definition 20.1 (L_1-exponential convergence). Let $\alpha \in [0, 1)$ and $M \in (0, \infty)$. Then the transition kernel K is L_1-exponentially convergent with (α, M) if

$$\|P^n - S\|_{L_1 \to L_1} \le \alpha^n M, \quad n \in \mathbb{N}. \tag{20.2}$$

A Markov chain with transition kernel K is called L_1-exponentially convergent if there exist an $\alpha \in [0, 1)$ and $M \in (0, \infty)$ such that (20.2) holds.

Definition 20.2 (L_2-spectral gap). We say that a transition kernel K and its corresponding Markov operator P have an L_2-spectral gap if

$$\mathrm{gap}(P) = 1 - \|P - S\|_{L_2 \to L_2} > 0.$$

If the transition kernel has an L_2-spectral gap, then by Lemmas 20.2 and 20.3 we have that

$$\|\nu P^n - \pi\|_{\mathrm{tv}} \le (1 - \mathrm{gap}(P))^n \left\| \frac{d\nu}{d\pi} - 1 \right\|_2 .$$

Next, we define other convergence properties which are based on the total variation distance.

Definition 20.3 (uniform ergodicity and geometric ergodicity). Let $\alpha \in [0, 1)$ and $M : G \to (0, \infty)$. Then the transition kernel K is called *geometrically ergodic with* $(\alpha, M(x))$ if one has for π-almost all $x \in G$ that

$$\|K^n(x, \cdot) - \pi\|_{\text{tv}} \leq M(x)\,\alpha^n, \quad n \in \mathbb{N}. \tag{20.3}$$

If the inequality (20.3) holds with a bounded function $M(x)$, i.e.

$$\sup_{x \in G} M(x) \leq M' < \infty,$$

then K is called *uniformly ergodic with* (α, M').

Now we state several relations between the different properties. Since we assume that the transition kernel is reversible with respect to π we have the following:

$$\begin{array}{ccc} \text{uniformly ergodic} & \Longleftrightarrow & L_1\text{-exponentially convergent} \\ \text{with } (\alpha, M) & & \text{with } (\alpha, 2M) \\ \Downarrow & & \Downarrow \\ \text{geometrically ergodic} & & L_2\text{-spectral gap} \geq \\ \text{with } (\alpha, M(x)) & & 1 - \alpha. \end{array} \tag{20.4}$$

The fact that uniform ergodicity implies geometric ergodicity is obvious. For the proofs of the other relations and further details we refer to [16, Proposition 3.23, Proposition 3.24]. Further, if the transition kernel is φ-irreducible, for details we refer to [13] and [15], then

$$\begin{array}{ccc} \text{geometrically ergodic} & \Longleftrightarrow & L_2\text{-spectral gap} \geq \\ \text{with } (\alpha, M(x)) & & 1 - \alpha. \end{array} \tag{20.5}$$

20.2.2 Mean Square Error Bounds of MCMC

The goal is to compute

$$S(f) = \int_G f(x)\,\pi(\mathrm{d}x).$$

We use an average of a finite Markov chain sample as approximation of the mean, i.e. we approximate $S(f)$ by

$$S_{n,n_0}(f) = \frac{1}{n} \sum_{j=1}^{n} f(X_{j+n_0}).$$

The number n determines the number of function evaluations of f. The number n_0 is the *burn-in* or *warm up* time. Intuitively, it is the number of steps of the Markov chain to get close to the stationary distribution π.

We study the mean square error of S_{n,n_0}, given by

$$e_\nu(S_{n,n_0}, f) = (\mathbb{E}_{\nu,K}|S_{n,n_0}(f) - S(f)|)^{1/2},$$

where ν and K indicate the initial distribution and transition kernel. We start with the case $\nu = \pi$, where the initial distribution is the stationary distribution.

Lemma 20.4. *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with transition kernel K and initial distribution π. We define*

$$\Lambda = \sup\{\alpha : \alpha \in \operatorname{spec}(P - S)\},$$

where $\operatorname{spec}(P - S)$ *denotes the spectrum of the operator $P - S : L_2 \to L_2$, and assume that $\Lambda < 1$. Then*

$$\sup_{\|f\|_2\le 1} e_\pi(S_{n,n_0}, f)^2 \le \frac{2}{n(1-\Lambda)}.$$

For a proof of this result we refer to [16, Corollary 3.27]. Let us discuss the assumptions and implications of Lemma 20.4. First, note that for the simple Monte Carlo method we have $\Lambda = 0$. In this case we get (up to a constant of 2) what we would expect. Further, note that $\operatorname{gap}(P) = 1 - \|P - S\|_{L_2\to L_2}$ and

$$\|P - S\|_{L_2\to L_2} = \sup\{|\alpha| : \alpha \in \operatorname{spec}(P - S)\},$$

so that $\operatorname{gap}(P) \le 1 - \Lambda$. This also implies that if $P : L_2 \to L_2$ is positive semidefinite we obtain $\operatorname{gap}(P) = 1 - \Lambda$. Thus, whenever we have a lower bound for the spectral gap we can apply Lemma 20.4 and can replace $1 - \Lambda$ by $\operatorname{gap}(P)$. Further note if $\gamma \in [0, 1)$, $M \in (0, \infty)$ and the transition kernel is L_1-exponentially convergent with (γ, M) then we have, using (20.4), that $\operatorname{gap}(P) \ge 1 - \gamma$.

Now we ask how $e_\nu(S_{n,n_0}, f)$ behaves depending on the initial distribution. The idea is to decompose the error in a suitable way. For example in a bias and variance term. However, we want to have an estimate with respect to $\|f\|_2$ and in this setting the following decomposition is more convenient:

$$e_\nu(S_{n,n_0}, f)^2 = e_\pi(S_{n,n_0}, f)^2 + \text{rest},$$

where rest denotes an additional term such that equality holds. Then, we estimate the remainder term and use Lemma 20.4 to obtain an error bound. For further details of the proof of the following error bound we refer to [16, Theorem 3.34 and Theorem 3.41].

Theorem 20.1. *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with reversible transition kernel K and initial distribution ν. Further, let*

$$\Lambda = \sup\{\alpha : \alpha \in \operatorname{spec}(P - S)\},$$

where $\operatorname{spec}(P - S)$ *denotes the spectrum of the operator $P - S : L_2 \to L_2$, and assume that $\Lambda < 1$. Then*

$$\sup_{\|f\|_p \le 1} e_\nu(S_{n,n_0}, f)^2 \le \frac{2}{n(1-\Lambda)} + \frac{2\, C_\nu \gamma^{n_0}}{n^2(1-\gamma)^2} \tag{20.6}$$

holds for $p = 2$ and for $p = 4$ under the following conditions

1. *For $p = 2$, $\frac{d\nu}{d\pi} \in L_\infty$ and a transition kernel K which is L_1-exponentially convergent with (γ, M) where $C_\nu = M \left\| \frac{d\nu}{d\pi} - 1 \right\|_\infty$;*
2. *For $p = 4$, $\frac{d\nu}{d\pi} \in L_2$ and $1 - \gamma = \operatorname{gap}(P) > 0$ where $C_\nu = 64 \left\| \frac{d\nu}{d\pi} - 1 \right\|_2$.*

Let us discuss the results. If the transition kernel is L_1-exponentially ergodic, then we have an explicit error bound for integrands $f \in L_2$ whenever the initial distribution has a density $\frac{d\nu}{d\pi} \in L_\infty$. However, in general it is difficult to provide explicit values γ and M such that the transition kernel is L_1-exponentially convergent with (γ, M). This motivates to consider transition kernel which satisfy a weaker convergence property, such as the existence of an L_2-spectral gap. In this case we have an explicit error bound for integrands $f \in L_4$ whenever the initial distribution has a density $\frac{d\nu}{d\pi} \in L_2$. Thus, by assuming a weaker convergence property of the transition kernel we obtain a weaker result in the sense that f must be in L_4 rather than L_2. However, with respect to $\frac{d\nu}{d\pi}$ we do not need boundedness anymore, it is enough that $\frac{d\nu}{d\pi} \in L_2$.

In Theorem 20.1 we provided explicit error bounds and we add in passing that also other error bounds are known, see [1, 4, 5, 16].

If we want to have an error of $\varepsilon \in (0, 1)$ it is still not clear how to choose n and n_0 to minimize the total amount of steps $n + n_0$. How should we choose the burn-in n_0? Let $e(n, n_0)$ be the right hand side of (20.6) and assume that $\Lambda = \gamma$. Further, assume that we have computational resources for $N = n + n_0$ steps of the Markov chain. We want to get an n_{opt} which minimizes $e(N - n_0, n_0)$. In [16, Lemma 2.26] the following is proven: For all $\delta > 0$ and large enough N and C_ν the number n_{opt} satisfies

$$n_{\text{opt}} \in \left[\frac{\log C_\nu}{\log \gamma^{-1}}, (1+\delta) \frac{\log C_\nu}{\log \gamma^{-1}} \right].$$

Further note that $\log \gamma^{-1} \ge 1 - \gamma$. Thus, in this setting $n_{\text{opt}} = \lceil \frac{\log C_\nu}{1-\gamma} \rceil$ is a reasonable and almost optimal choice for the burn-in.

20.3 Application of the Error Bound and Limitations of MCMC

First, we briefly introduce a technique to prove a lower bound of the spectral gap if the Markov operator of a transition kernel is positive semidefinite on L_2. The following result, known as *Cheeger's inequality*, is in this form due to Lawler and Sokal [6].

Proposition 20.1. *Let K be a reversible transition kernel, which induces a Markov operator $P\colon L_2 \to L_2$. Then*

$$\frac{\varphi^2}{2} \le 1 - \Lambda \le 2\varphi,$$

where $\Lambda = \sup\{\alpha\colon \alpha \in \operatorname{spec}(P - S)\}$ and

$$\varphi = \inf_{0<\pi(A)\le 1/2} \frac{\int_A K(x, A^c)\,\pi(\mathrm{d}x)}{\pi(A)}$$

is the conductance of K.

Now we state different applications of Theorem 20.1.

20.3.1 Hit-and-Run Algorithm

We consider the example of Sect. 20.1. Let $G \in \mathscr{G}_{r,d}$, see (20.1), and let μ_G be the uniform distribution in G. We define

$$\mathscr{F}_{r,d} = \{(f, G)\colon G \in \mathscr{G}_{r,d},\ f \in L_4(\mu_G),\ \|f\|_4 \le 1\}. \tag{20.7}$$

The goal is to approximate

$$S(f, \mathbf{1}_G) = \frac{1}{\mathrm{vol}_d(G)} \int_G f(x)\,\mathrm{d}x,$$

where $(f, G) \in \mathscr{F}_{r,d}$. The hit-and-run algorithm defines a Markov chain which satisfies the assumptions of Theorem 20.1. A step from $x \in G$ of the hit-and-run algorithm works as follows

1. Choose a direction, say θ, uniformly distributed on the sphere ∂B_d.
2. Choose the next state, say $y \in G$, uniformly distributed in $G \cap \{x + \theta r\colon r \in \mathbb{R}\}$.

After choosing a direction θ one samples the next state $y \in G$ with respect to the uniform distribution in the line determined by the current state x and the direction

θ restricted to G. The random number, say $u \in [0,1]$, for the second part is chosen independently of the first part and also all steps are independent.

Lovaśz and Vempala prove in [7, Theorem 4.2, p. 993] a lower bound of the conductance φ, see Proposition 20.1 for the definition of the conductance.

Proposition 20.2. *Let $G \in \mathscr{G}_{r,d}$. Then, the conductance of the hit-and-run algorithm is bounded from below by $2^{-25}(dr)^{-1}$.*

It is known that the hit-and-run algorithm induces a positive semidefinite Markov operator, say H, see [19]. By Proposition 20.1 we obtain

$$\operatorname{gap}(H) \geq \frac{2^{-51}}{(dr)^2}$$

and Theorem 20.1 implies the following error bound for the class $\mathscr{F}_{r,d}$, see (20.1) and (20.7).

Theorem 20.2. *Let ν be the uniform distribution on B_d. Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with transition kernel, given by the hit-and-run algorithm, and initial distribution ν. Let*

$$n_0 = \lceil 4.51 \cdot 10^{15} d^2 r^2 (d \log r + 4.16) \rceil.$$

Then

$$\sup_{(f,G)\in\mathscr{F}_{r,d}} e_\nu(S_{n,n_0}, (f, \mathbf{1}_G)) \leq 9.5 \cdot 10^7 \frac{dr}{\sqrt{n}} + 6.4 \cdot 10^{15} \frac{d^2 r^2}{n}.$$

This result states that the number of oracle calls for f and G to obtain an error $\varepsilon > 0$ is bounded by $\kappa\, d^2 r^2 (\varepsilon^{-2} + d \log r)$, for an explicit constant $\kappa > 0$. Hence the computation of $S(f, \mathbf{1}_G)$ on the class $\mathscr{F}_{r,d}$ is polynomially tractable, see [10–12]. The tractability result can be extended also to other classes of functions, see [17]. Note that we applied the second statement of Theorem 20.1. It is known that the hit-and-run algorithm is L_1-exponentially ergodic with (γ, M), for some $\gamma \in (0,1)$ and $M \in (0,\infty)$. But the best known numbers γ and M are exponentially bad in terms of the dimension, see [20].

20.3.2 Metropolis-Hastings Algorithm

Let $G \subset \mathbb{R}^d$ and $\rho\colon G \to (0,\infty)$, where ρ is integrable with respect to the Lebesgue measure. We define the distribution π_ρ on $(G, \mathscr{B}(G))$ by

$$\pi_\rho(A) = \frac{\int_A \rho(x)\,\mathrm{d}x}{\int_G \rho(x)\,\mathrm{d}x}, \qquad A \in \mathscr{B}(G).$$

The goal is to compute

$$S(f,\rho) = \int_G f(x)\,\pi_\rho(\mathrm{d}x) = \frac{\int_G f(x)\rho(x)\,\mathrm{d}x}{\int_G \rho(x)\,\mathrm{d}x}$$

for functions $f\colon G \to \mathbb{R}$ which are integrable with respect to π_ρ.

The *Metropolis-Hastings algorithm* defines a Markov chain which approximates π_ρ. We need some further notations. Let $q\colon G \times G \to [0,\infty]$ be a function such that $q(x,\cdot)$ is Lebesgue integrable for all $x \in G$ with $\int_G q(x,y)\,\mathrm{d}y \le 1$. Then

$$Q(x,A) = \int_A q(x,y)\,\mathrm{d}y + \mathbf{1}_A(x)\left(1 - \int_G q(x,y)\,\mathrm{d}y\right), \quad x \in G,\ A \in \mathscr{B}(G),$$

is a transition kernel and we call $q(\cdot,\cdot)$ *transition density*. The idea is to modify Q, such that π_ρ gets a stationary distribution of the modification. We propose a state with Q and with a certain probability, which depends on ρ, the state is accepted. Let $\alpha(x,y)$ be the acceptance probability

$$\alpha(x,y) = \begin{cases} 1 & \text{if } q(x,y)\rho(x) = 0, \\ \min\{1, \frac{q(y,x)\rho(y)}{q(x,y)\rho(x)}\} & \text{otherwise.} \end{cases}$$

The transition kernel of the Metropolis-Hastings algorithm is

$$K_\rho(x,A) = \int_A \alpha(x,y)\,q(x,y)\mathrm{d}y + \mathbf{1}_A(x)\left[1 - \int_G \alpha(x,y)\,q(x,y)\mathrm{d}y\right]$$

for $x \in G$ and $A \in \mathscr{B}(G)$. The transition kernel K_ρ is reversible with respect to π_ρ. From the current state $x \in G$ a single transition of the algorithm works as follows:

1. Sample a proposal state $y \in G$ with respect to $Q(x,\cdot)$.
2. With probability $\alpha(x,y)$ return y, otherwise reject y and return x.

Again, all steps are done independently of each other. If $q(x,y) = q(y,x)$, i.e. q is symmetric, then K_ρ is called *Metropolis algorithm* and if $q(x,y) = \eta(y)$ for a function $\eta\colon G \to (0,\infty)$ for all $x, y \in G$, then K_ρ is called *independent Metropolis algorithm*.

Let $G \subset \mathbb{R}^d$ be bounded and for $C \ge 1$ let

$$\mathscr{R}_C = \{\rho\colon G \to (0,\infty) \mid 1 \le \rho(x) \le C\}. \tag{20.8}$$

Thus, for any $\rho \in \mathscr{R}_C$ holds $\sup\rho/\inf\rho \le C$. If $\rho\colon G \to (0,\infty)$ satisfies $\sup\rho/\inf\rho \le C$, then

$$\frac{\|\rho\|_\infty}{C} \le \rho(x) \le C \inf\rho.$$

Thus, $C \cdot \rho / \|\rho\|_\infty \in \mathscr{R}_C$. We consider an independent Metropolis algorithm. The proposal transition kernel is

$$Q(x, A) = \mu_G(A) = \frac{\mathrm{vol}_d(A)}{\mathrm{vol}_d(G)}, \quad A \in \mathscr{B}(G),$$

i.e. a state is proposed with the uniform distribution in G. Then

$$K_\rho(x, A) = \int_A \alpha(x, y) \frac{\mathrm{d}y}{\mathrm{vol}_d(G)} + \mathbf{1}_A(x) \left(1 - \int_G \alpha(x, y) \frac{\mathrm{d}y}{\mathrm{vol}_d(G)}\right),$$

where $\alpha(x, y) = \min\{1, \rho(y)/\rho(x)\}$. The transition operator $P_\rho \colon L_2(\pi_\rho) \to L_2(\pi_\rho)$, induced by K_ρ, is positive semidefinite. For details we refer to [19]. Thus, $\mathrm{gap}(P_\rho) = 1 - \Lambda_\rho$, with $\Lambda_\rho = \Lambda$. Further, for $\rho \in \mathscr{R}_C$ Theorem 2.1 of [9] provides a criterion for uniform ergodicity of the independent Metropolis algorithm. Namely, K_ρ is uniformly ergodic with $(\gamma, 1)$ for $\gamma = 1 - C^{-1}/\mathrm{vol}_d(G)$. Thus, by (20.4) we have that it is L_1-exponentially ergodic with $(\gamma, 2)$. Further, by (20.4) we obtain

$$1 - \Lambda_\rho = \mathrm{gap}(P_\rho) \geq \frac{C^{-1}}{\mathrm{vol}_d(G)}.$$

Let

$$\mathscr{F}_{C,d} = \{(f, \rho) \colon \rho \in \mathscr{R}_C,\ f \in L_2(\pi_\rho),\ \|f\|_2 \leq 1\}. \tag{20.9}$$

We apply Theorem 20.1 and obtain for the class $\mathscr{F}_{C,d}$ (see (20.8) and (20.9))

Theorem 20.3. *Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with transition kernel, given by the Metropolis algorithm with proposal μ_G, and initial distribution μ_G. Let*

$$n_0 = \lceil C \,\mathrm{vol}_d(G) \log(2C) \rceil .$$

Then

$$\sup_{(f,\rho)\in\mathscr{F}_{C,d}} e_\nu(S_{n,n_0}, (f, \rho))^2 \leq \frac{2\,C\,\mathrm{vol}_d(G)}{n} + \frac{4\,C^2\,\mathrm{vol}_d(G)^2}{n^2}.$$

The upper bound in Theorem 20.3 does not depend on the dimension d, as long as $\mathrm{vol}_d(G)$ and C do not depend on d. In some applications, however, the upper bound is rather useless since $C = C_d$ is exponentially large in d. Assume, for example, that

$$\rho(x) = \exp(-\alpha |x|^2), \tag{20.10}$$

i.e. ρ is the non-normalized density of a $N(0, \sqrt{2\alpha^{-1}})$ random variable. We consider scaled versions of ρ. If $G = B_d$, then $\exp(\alpha)\rho \in \mathscr{R}_{\exp(\alpha)}$ and if $G = [-1,1]^d$, then $\exp(\alpha d)\rho \in \mathscr{R}_{\exp(\alpha d)}$. This is bad, since C, for example $\exp(\alpha)$ or $\exp(\alpha d)$, might depend exponentially on α and d.

This example shows that we would greatly prefer an upper bound where C is replaced by a power of $\log C$. However, on the class $\mathscr{F}_{C,d}$ this is not possible. The same proof as in [8, Theorem 1] leads to the following lower bound for *all* randomized algorithms.

Theorem 20.4. *Any randomized algorithm S_n that uses n values of f and ρ satisfies the lower bound*

$$\sup_{(f,\rho)\in\mathscr{F}_{C,d}} e(S_n,(f,\rho)) \geq \frac{\sqrt{2}}{6}\begin{cases}\sqrt{\frac{C}{2n}} & 2n \geq C-1,\\ \frac{3C}{C+2n-1} & 2n < C-1.\end{cases}$$

The class $\mathscr{F}_{C,d}$ is too large. Thus the error bound is not satisfying. In the following we prove a much better upper bound for a smaller class of densities. Let $G = B_d$ and let ρ be log-concave, i.e. for all $\lambda \in (0,1)$ and for all $x, y \in B_d$ we have

$$\rho(\lambda x + (1-\lambda)y) \geq \rho(x)^{\lambda}\rho(y)^{1-\lambda}. \tag{20.11}$$

Then let

$$\mathscr{R}_{\alpha,d} = \{\rho\colon B_d \to (0,\infty) \mid \rho \text{ is log-concave}, \, |\log\rho(x) - \log\rho(y)| \leq \alpha|x-y|\}. \tag{20.12}$$

We consider log-concave densities where $\log\rho$ is Lipschitz continuous with constant α. Note that the setting is more restrictive compared to the previous one. The goal is to get an upper error bound which is polynomially in α and d. We consider a *Metropolis algorithm based on a ball walk*. For $\delta > 0$ the transition kernel of the δ ball walk is

$$B_\delta(x,A) = \frac{\mathrm{vol}_d(A \cap B_\delta(x))}{\mathrm{vol}_d(B_\delta(0))} + \mathbf{1}_A(x)\left(1 - \frac{\mathrm{vol}_d(G \cap B_\delta(x))}{\mathrm{vol}_d(B_\delta(0))}\right), \quad x \in G,\ A \in \mathscr{B}(G),$$

where $B_\delta(x)$ denotes the Euclidean ball with radius δ around x. Let $K_{\rho,\delta}$ be the transition kernel of the Metropolis algorithm with ball walk proposal B_δ, let $P_{\rho,\delta}$ be the corresponding transition operator and let $\Lambda_{\rho,\delta}$ be the largest element of the spectrum of $P_{\rho,\delta} - S\colon L_2(\pi_\rho) \to L_2(\pi_\rho)$.

In [8, Corollary 1] the following result is proven.

Proposition 20.3. *Let $\rho \in \mathscr{R}_{\alpha,d}$ and let $\delta = \min\{1/\sqrt{d+1}, \alpha^{-1}\}$. Then, the conductance of $K_{\rho,\delta}$ is bounded from below by*

$$\frac{0.0025}{\sqrt{d+1}} \min\left\{\frac{1}{\sqrt{d+1}}, \frac{1}{\alpha}\right\}.$$

By Propositions 20.1 and 20.3 we have a lower bound of $1 - \Lambda_{\rho,\delta}$. However, to apply Theorem 20.1 we need a lower bound on $\mathrm{gap}(P_{\rho,\delta})$. Let $\tilde{K}_{\rho,\delta}$ be the transition kernel of the lazy version of $K_{\rho,\delta}$, i.e. for $x \in G$ and $A \in \mathscr{B}(G)$ holds $\tilde{K}_{\rho,\delta}(x, A) = (K_{\rho,\delta}(x, A) + \mathbf{1}_A(x))/2$. In words, $\tilde{K}_{\rho,\delta}$ can be described as follows: With probability $1/2$ stay at the current state and with probability $1/2$ do one step with $K_{\rho,\delta}$. This transition kernel induces a positive semidefinite operator $\tilde{P}_{\rho,\delta}: L_2(\pi_\rho) \to L_2(\pi_\rho)$ with

$$\mathrm{gap}(\tilde{P}_{\rho,\delta}) = \frac{1}{2}(1 + \Lambda_{\rho,\delta}).$$

Let

$$\mathscr{F}_{\alpha,d} = \{(f, \rho): \rho \in \mathscr{R}_{\alpha,d},\ f \in L_4(\pi_\rho),\ \|f\|_4 \le 1\}, \tag{20.13}$$

and recall that $\mathscr{R}_{\alpha,d}$ is defined in (20.12). Note that we assumed $G = B_d$. Now we can apply Theorem 20.1 for the lazy Metropolis algorithm with ball walk proposal $\tilde{K}_{\rho,\delta}$.

Theorem 20.5. *Let ν be the uniform distribution on B_d and let us assume that $\delta = \min\{1/\sqrt{d+1}, \alpha^{-1}\}$. Let $(X_n)_{n\in\mathbb{N}}$ be a Markov chain with transition kernel $\tilde{K}_{\rho,\delta}$, i.e. the lazy version of the Metropolis algorithm with ball walk proposal B_δ, and initial distribution ν. Let*

$$n_0 = \lceil 5.92 \cdot 10^6 (d+1) \max\{\alpha^2, d+1\}(2\alpha + 4.16) \rceil.$$

Then

$$\sup_{(f,G)\in\mathscr{F}_{\alpha,d}} e_\nu(S_{n,n_0}, (f, \rho)) \le 1{,}089 \frac{\sqrt{d+1}\max\{\alpha, \sqrt{d+1}\}}{\sqrt{n}} + 8.38 \cdot 10^5 \frac{(d+1)\max\{\alpha^2, d+1\}}{n}.$$

The last theorem states that the number of oracle calls of f and ρ to obtain an error $\varepsilon > 0$ is bounded by $\kappa\, d \max\{\alpha^2, d\}(\varepsilon^2 + \alpha)$. Hence the computation of $S(f, \rho)$ is polynomially tractable. Note that $\mathscr{R}_{\alpha,d}$ might be interpreted as a subclass of $\mathscr{R}_C$ with $C = \exp(2\alpha)$ and $G = B_d$, since $\rho \in \mathscr{R}_{\alpha,d}$ implies $\exp(2\alpha)\rho/\|\rho\|_\infty \in \mathscr{R}_{\exp(2\alpha)}$. Thus, by Theorem 20.5 we obtain that the number of oracle calls to get an error ε also depends polynomially on $\log C$, since $C = \exp(2\alpha)$.

20.4 Open Problems and Related Comments

- We do not know whether an error bound as in Theorem 20.1 holds for $f \in L_2$ if $\text{gap}(P) > 0$.
- In [18] error bounds of S_{n,n_0} for $f \in L_p$ with $1 < p \leq 2$ are proven. Then one needs a new error criterion, here the absolute mean error

$$\mathbb{E}_{\nu,K}|S_{n,n_0}(f) - S(f)|$$

  is used. If the Markov chain is L_1-exponentially convergent, then the error bound decreases with $n^{1/p-1}$. For a Markov chain with L_2-spectral gap a similar error bound is shown.
- The tractability results in Theorems 20.2 and 20.5 are nice since the degree of the polynomial is small. Nevertheless, the upper bound is not really useful because of the huge constants. Is it possible to prove these or similar results with much smaller constants?
- A related question would be the construction of Markov chain quasi-Monte Carlo methods, see [2, 3]. Here the idea is to derandomize the Markov chain by using a carefully constructed deterministic sequence of numbers to obtain a sample $x_1, \ldots, x_{n+n_0}$. However, explicit constructions with small error bounds are not known.

References

1. Belloni, A., Chernozhukov, V.: On the computational complexity of MCMC-based estimators in large samples. Ann. Stat. **37**(4), 2011–2055 (2009)
2. Chen, S., Dick, J., Owen, A.: Consistency of Markov chain quasi-Monte Carlo on continuous state spaces. Ann. Stat. **39**(2), 673–701 (2011)
3. Dick, J., Rudolf, D., Zhu, H.: Discrepancy bounds for uniformly ergodic Markov chain quasi-Monte Carlo (2013). http://arxiv.org/abs/1303.2423
4. Joulin, J.A., Ollivier, Y.: Curvature, concentration and error estimates for Markov chain Monte Carlo. Ann. Probab. **38**(6), 2418–2442 (2010)
5. Latuszynski, K., Miasojedow, B., Niemiro, W.: Nonasymptotic bounds on the estimation error of MCMC algorithms. Bernoulli **19**(5A), 2033–2066 (2013)
6. Lawler, G.F., Sokal, A.D.: Bounds on the l^2 spectrum for Markov chains and Markov Processes: a generalization of cheeger's inequality. Trans. Am. Math. Soc. **309**(2), 557–580 (1988)
7. Lovász, L., Vempala, S.: Hit-and-run from a corner. SIAM J. Comput. **35**(4), 985–1005 (2006)
8. Mathé, P., Novak, E.: Simple Monte Carlo and the metropolis algorithm. J. Complex. **23**(4–6), 673–696 (2007)
9. Mengersen, K., Tweedie, R.: Rates of convergence of the hastings and metropolis algorithms. Ann. Stat. **24**(1), 101–121 (1996)
10. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume I: Linear Information. EMS Tracts in Mathematics, vol. 6. European Mathematical Society (EMS), Zürich (2008)

11. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume II: Standard Information for Functionals. EMS Tracts in Mathematics, vol. 12. European Mathematical Society (EMS), Zürich (2010)
12. Novak, E., Woźniakowski, H.: Tractability of Multivariate Problems. Volume III: Standard Information for Operators. EMS Tracts in Mathematics, vol. 18. European Mathematical Society (EMS), Zürich (2012)
13. Roberts, G.O., Rosenthal, J.S.: Geometric ergodicity and hybrid Markov chains. Electron. Commun. Probab. **2**, 13–25 (1997)
14. Roberts, G.O., Rosenthal, J.S.: General state space Markov chains and MCMC algorithms. Probab. Surv. **1**, 20–71 (2004)
15. Roberts, G.O., Tweedie, R.L.: Geometric l^2 and l^1 convergence are equivalent for reversible Markov chains. J. Appl. Probab. **38A**, 37–41 (2001)
16. Rudolf, D.: Explicit error bounds for Markov chain Monte Carlo. Diss. Math. **485**, 93 (2012)
17. Rudolf, D.: Hit-and-run for numerical integration. In: Monte Carlo and Quasi-Monte Carlo Methods 2012. Springer Proceedings in Mathematics & Statistics, vol. 65, pp. 597–612. Springer, Berlin/Heidelberg (2013)
18. Rudolf, D., Schweizer, N.: Error bounds of MCMC for functions with unbounded stationary variance (2013). http://arxiv.org/abs/1312.4344
19. Rudolf, D., Ullrich, M.: Positivity of hit-and-run and related algorithms. Electron. Commun. Probab. **18**, 1–8 (2013)
20. Smith, R.L.: Efficient Monte Carlo procedures for generating points uniformly distributed over bounded regions. Oper. Res. **32**, 1296–1308 (1984)

Chapter 21
Regularity, Complexity, and Approximability of Electronic Wavefunctions

Harry Yserentant

Abstract The electronic Schrödinger equation describes the motion of N electrons under Coulomb interaction forces in a field of clamped nuclei. The solutions of this equation, the electronic wavefunctions, depend on $3N$ variables, three spatial dimensions for each electron. Approximating the solutions is thus inordinately challenging, and it is conventionally believed that a reduction to simplified models, such as those of the Hartree-Fock method or density functional theory, is the only tenable approach. The situation is, however, more complicated: the regularity of the solutions, which increases with the number of electrons, the decay behavior of their mixed derivatives, and the antisymmetry enforced by the Pauli principle contribute properties that allow these functions to be approximated with an order of complexity which comes arbitrarily close to that of a system of two or even only one electron.

21.1 Introduction

The approximation of high-dimensional functions, might they be given explicitly or implicitly as solutions of differential equations, represents one of the grand challenges of applied mathematics. High-dimensional problems arise in many fields of application such as data analysis and statistics or machine learning and computational finance, but first of all in the sciences. One of the most notorious and complicated problems of this type is the Schrödinger equation. The Schrödinger equation forms the basis of quantum mechanics and is of fundamental importance for our understanding of atoms and molecules. It links chemistry to physics and describes a system of electrons and nuclei that interact by Coulomb attraction and repulsion forces. As proposed by Born and Oppenheimer in the early times of quantum theory, the much slower motion of the nuclei is mostly separated from that of the electrons. This results in the electronic Schrödinger equation, the problem to find the eigenvalues and eigenfunctions of the differential operator

H. Yserentant (✉)
Technische Universität Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
e-mail: yserentant@math.tu-berlin.de

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5_21

$$H = -\frac{1}{2}\sum_{i=1}^{N}\Delta_i - \sum_{i=1}^{N}\sum_{\nu=1}^{K}\frac{Z_\nu}{|x_i - a_\nu|} + \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\frac{1}{|x_i - x_j|}, \tag{21.1}$$

written here in dimensionless form or atomic units. It acts on functions with arguments $x_1, \ldots, x_N$ in $\mathbb{R}^3$ that are associated with the positions of given N electrons; the positions a_ν of the nuclei are kept fixed. The positive values Z_ν are the charges of the nuclei in multiples of the electron charge.

Because of its high-dimensionality, it seems to be completely hopeless to attack the electronic Schrödinger equation directly. Dirac, one of the fathers of quantum theory, commented this with the often quoted words, "the underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble." This situation did not much change since then, and depending on what one understands by soluble, it will never change. Dirac continued, "it therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation." Physicists and chemists followed Dirac's advice and invented, during the previous decades, a whole raft of such methods of steadily increasing sophistication. The most prominent are the Hartree-Fock method and its many variants, extensions, and successors, and the newer density functional based methods, that are based on the observation that the ground state energy is completely determined by the electron density. Modern quantum-chemical approximation methods are based on deep insights into the nature of atoms and molecules. They are used with much success and form the basis of a steadily expanding branch of chemistry. Their power and efficiency are impressive. There is, however, no real mathematical explanation for their often amazing accuracy. From the perspective of a mathematician, all these methods have a decisive drawback. They either simplify the basic equation and suffer from a priori modeling errors, or it is unclear how the accuracy can be systematically improved without that the effort truly explodes with the number of particles. In other words, they are no true, unbiased discretizations of the Schrödinger equation in the sense of numerical analysis.

Several groups in the priority program tried to change this unsatisfying situation and to develop tools aiming to overcome the described complexity barriers. The present article surveys some theoretical results [7] and [6, 8, 9] by the author and his coworkers on the mixed regularity of the electronic wavefunctions and the decay behavior of their mixed derivatives. On the basis of these analytical properties and the antisymmetry of the wavefunctions enforced by the Pauli principle, one can surprisingly construct simple, sparse grid or hyperbolic cross like expansions of the wavefunctions whose convergence rate, measured in terms of the number of basis functions involved, astonishingly does not deteriorate with the number of electrons. It comes close to that for the case of two or even only of one single electron [8, 10].

21.2 The Variational Form of the Equation

The solution space of the electronic Schrödinger equation is the Hilbert space H^1 that consists of the one time weakly differentiable, square integrable functions

$$u : (\mathbb{R}^3)^N \to \mathbb{R} : (x_1, \ldots, x_N) \to u(x_1, \ldots, x_N) \tag{21.2}$$

with square integrable first-order weak derivatives. The norm on H^1 is composed of the L_2-norm $\|\cdot\|_0$ and the H^1-seminorm, the L_2-norm of the gradient. In the language of physics, H^1 is the space of the wavefunctions for which the total position probability remains finite and the expectation value of the kinetic energy can be given a meaning. Let $\mathscr{D}$ be the space of the infinitely differentiable functions (21.2) with bounded support. The functions in $\mathscr{D}$ form a dense subset of H^1. Let

$$V(x) \;=\; -\sum_{i=1}^{N}\sum_{\nu=1}^{K}\frac{Z_\nu}{|x_i - a_\nu|} \;+\; \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{N}\frac{1}{|x_i - x_j|}$$

be the potential in the Schrödinger operator (21.1). The basic observation is that there is a constant $\theta > 0$ such that for all functions u and v in the space $\mathscr{D}$

$$\int V u\, v \,\mathrm{d}x \;\leq\; \theta\, \|u\|_0 \|\nabla v\|_0 \tag{21.3}$$

in terms of the L_2-norm of u and of the gradient of v holds. The proof of this estimate is based on the three-dimensional Hardy inequality. The expression

$$a(u, v) = (Hu, v)$$

defines therefore a H^1-bounded bilinear form on $\mathscr{D}$, where $(\cdot, \cdot)$ denotes the L_2-inner product. It can be uniquely extended to a bounded bilinear form on H^1. In this setting, a function $u \neq 0$ in H^1 is an eigenfunction of the electronic Schrödinger operator (21.1) for the eigenvalue λ if the relation

$$a(u, v) = \lambda(u, v) \tag{21.4}$$

holds for all test functions $v \in H^1$. The weak form (21.4) of the eigenvalue equation $Hu = \lambda u$ particularly fixes the behavior of the eigenfunctions at the singularities of the interaction potential and at infinity. For normed u, $a(u, u)$ is the expectation value of the total energy. One can deduce from the estimate (21.3) that the total energy of the system is bounded from below. Hence one is allowed to define the constant

$$\Lambda \;=\; \inf\big\{a(u, u) \,\big|\, u \in \mathscr{D},\ \|u\|_0 = 1\big\},$$

the minimum energy that the system can attain. Its counterpart is the ionization threshold. To prepare its definition lets

$$\Sigma(R) = \inf\big\{a(u,u) \,\big|\, u \in \mathscr{D}(R),\ \|u\|_0 = 1\big\},$$

where $\mathscr{D}(R)$ consists of those functions in $\mathscr{D}$ for which $u(x) = 0$ for $|x| \le R$. One can show that the constants $\Sigma(R)$ are bounded from above by the value zero. As they are monotonously increasing in R, one can therefore define the constant

$$\Sigma^* = \lim_{R\to\infty} \Sigma(R) \le 0,$$

the energy threshold above which at least one electron has moved arbitrarily far away from the nuclei, the ionization threshold. We restrict ourselves here to the case that $\Lambda < \Sigma^*$, that is, that it is energetically more advantageous for the electrons to stay in the vicinity of the nuclei than to fade away at infinity. In other words, we consider cases where the nuclei can bind all electrons. The ionization threshold is the bottom of the essential spectrum of the Schrödinger operator. This article discusses properties of eigenfunctions for eigenvalues below the ionization threshold. Such eigenfunctions u decay exponentially in the L_2-sense. That means there is a constant $\gamma > 0$, depending on the distance of the eigenvalue under consideration to the bottom of the essential spectrum, such that the functions

$$x \quad \to \quad \exp\left(\gamma \sum_{i=1}^{N} |x_i|\right) u(x) \tag{21.5}$$

are square integrable. Details and references to the literature can be found in [8].

21.3 The Mixed Regularity of the Wavefunctions

The regularity of the electronic wavefunctions increases in a sense with the number of electrons, the reason being that the interaction potential is composed of two-particle interactions of a very specific form. To describe this behavior, we introduce a scale of norms that are given by

$$|||u|||^2_{\vartheta,m} = \int \left\{1 + \sum_{i=1}^{N} |\omega_i|^2\right\}^m \left\{\prod_{i=1}^{N} \left(1 + |\omega_i|^2\right)\right\}^{\vartheta} |\hat{u}(\omega)|^2 \, \mathrm{d}\omega \tag{21.6}$$

and are defined on the Hilbert spaces $H^{\vartheta,m}_{\mathrm{mix}}$ that consist of the square integrable functions (21.2) for which these expressions remain finite. The $\omega_i \in \mathbb{R}^3$ forming together the variable $\omega \in (\mathbb{R}^3)^N$ can be associated with the momentums of the electrons; the expressions $|\omega_i|$ are their euclidean norms. For nonnegative integer values m and ϑ, the norms measure the L_2-norm of weak partial derivatives. The parameter m measures the isotropic smoothness that does not distinguish between

different directions, and the parameter ϑ the mixed smoothness in direction of the three-dimensional coordinate spaces of the electrons. The spaces L_2 and H^1 are special cases of such spaces, with indices $m = 0$ and $\vartheta = 0$ respectively $m = 1$ and $\vartheta = 0$. A function in $H^{1,0}_{\text{mix}}$ possesses weak partial derivatives of order N in L_2.

It has been proved in [8] that the physically admissible eigenfunctions of the electronic Schrödinger operator (21.1), those with symmetry properties as enforced by the Pauli principle, and their exponentially weighted counterparts (21.5) as well possess a large number of square integrable mixed derivatives, sufficiently many to show that they are contained in the space $H^{\vartheta,1}_{\text{mix}}$ for $\vartheta = 1/2$. This result has been improved substantially in [6]. It has been shown there that the eigenfunctions of the electronic Schrödinger operator and their exponentially weighted counterparts (21.5) are, independent of these symmetry properties, contained in

$$H^{1,0}_{\text{mix}} \cap \bigcap_{\vartheta<3/4} H^{\vartheta,1}_{\text{mix}}. \tag{21.7}$$

The bound $3/4$ is optimal and can, except for special cases, probably neither be reached nor improved further. This shows the example of the function

$$u(x) = \Big(1 + \frac{1}{2}\,|x_1 - x_2|\Big)\,\exp\Big(-\frac{1}{4}\,|x_1|^2 - \frac{1}{4}\,|x_2|^2\Big)$$

the ground state of the so-called hookium or harmonium atom, an artificial two-electron system with the Hamiltonian

$$-\frac{1}{2}\,\Delta + \frac{1}{8}\,|x|^2 + \frac{1}{|x_1 - x_2|}$$

in which the potential of the nucleus is replaced by that of a harmonic oscillator.

It is however, possible to prove a higher regularity in weighted L_2-spaces [6]. To describe the corresponding result, we first introduce the set

$$\mathscr{A} = \big\{(\alpha_1, \ldots, \alpha_N)\,\big|\,\alpha_i \in \mathbb{Z}^3_{\geq 0},\ \alpha_{i,1} + \alpha_{i,2} + \alpha_{i,3} \leq 1\big\}, \tag{21.8}$$

of multi-indices α and the weight function

$$W(x) = \min\big\{\min_{i<j} |x_i - x_j|,\ 1\big\}.$$

The weak derivatives $\mathrm{D}^\alpha u$, $\alpha \in \mathscr{A}$, of the eigenfunctions u of the Schrödinger operator, which exist by the result above and are square integrable, are then themselves one times weakly differentiable outside the singular set where the positions x_i of two or more electrons coincide and it holds

$$W^\mu \nabla \mathrm{D}^\alpha u \in L_2$$

for all exponents $\mu > 1/2$. Again, the bound $\mu = 1/2$ cannot be reached.

21.4 The Transcorrelated Formulation and the Regularity Proof

A physically admissible wavefunction, that is compatible with the Pauli principle, vanishes where two electrons of same spin meet, a fact which counterbalances the singular behavior of the derivatives of the interaction potential. Our original proof [7, 8] utilized this fact. The more recent proof in [6] is based on the multiplicative splitting of the eigenfunction u under consideration into a more regular part

$$u_0(x) \;=\; \exp\Big(-\sum_{i<j}\phi(x_i-x_j)\Big)\,u(x) \tag{21.9}$$

in $H^{1,1}_{\mathrm{mix}}$ and a universal factor that covers the electron-electron singularities. This kind of splitting can be traced back to the work of Hylleraas [5] in the early years of quantum mechanics and plays an important role in quantum chemistry. It has been used in [1] and [4] to study the Hölder regularity of the eigenfunctions. There is a lot of freedom in the choice of the function ϕ. It needs only to be of the form

$$\phi(x) \;=\; \tilde\phi(|x|), \quad \tilde\phi'(0) \;=\; \frac{1}{2},$$

where $\tilde\phi : [0,\infty) \to \mathbb{R}$ is an infinitely differentiable function behaving sufficiently well at infinity. The regularity is therefore determined by that of the explicitly known factor from (21.9) that describes the behavior of the solutions at the singular points of the electron-electron interaction potential.

The proof starts from the weak form

$$\frac{1}{2}\int \nabla u_0\cdot\nabla v\,\mathrm{d}x \;+\; s(u_0,v) \;=\; \lambda(u_0,v), \quad v\in H^1, \tag{21.10}$$

of the so-called transcorrelated equation for the regular part (21.9) of the wavefunctions. The bilinear form $s(u,v)$ is nonsymmetric and originates from the zero- and first-order parts of this equation. Its coefficient functions are by one order less singular than those of the original operator (21.1). An explicit representation of this bilinear form can be found in [9]. The key to our regularity proof is the estimate

$$\tilde s\,(u,v) \;\le\; (1+\Omega^{-1})\,\kappa\,\|\!|u|\!\|_{1,0}\|\!|v|\!\|_{1,1} \tag{21.11}$$

for the bilinear form $\tilde s\,(u,v) \;=\; s(u,\mathscr{L}v)$ on the space of infinitely differentiable functions with compact support. The high-order differential operator

$$\mathscr{L} \;=\; \sum_{\alpha\in\mathscr{A}}(-1)^{|\alpha|}\Omega^{-2|\alpha|}\mathrm{D}^{2\alpha} \;=\; \prod_{i=1}^{N}(\mathrm{I}-\Omega^{-2}\Delta_i)$$

is built up of the even-order partial derivatives $\mathrm{D}^{2\alpha}$ for the multi-indices α in the set (21.8). The prefactors reflect the different scaling behavior of the derivatives; the constant Ω will be fixed later. The differential operator induces the norms given by

$$|||u|||_{1,0}^2 = \sum_{\alpha\in\mathscr{A}} \Omega^{-2|\alpha|} \|\mathrm{D}^{2\alpha} u\|_0^2, \quad |||v|||_{1,1}^2 = \sum_{\alpha\in\mathscr{A}} \Omega^{-2|\alpha|} \|\mathrm{D}^{2\alpha} v\|_1^2, \tag{21.12}$$

that are scaled variants of the norms on the spaces $H^{1,0}_{\mathrm{mix}}$ and $H^{1,1}_{\mathrm{mix}}$. The constant κ does not depend on the scaling parameter Ω and the positions of the nuclei. A proof of the estimate (21.11) for the case $\Omega = 1$ can be found in [9]. It is based on estimates of the single integral expressions of which the bilinear form $\tilde{s}\,(u, v)$ is composed and relies on careful integration by parts across the singularities of the coefficient functions involved and the three-dimensional Hardy inequality. It is no big deal to incorporate the scaling parameter Ω into the estimates of the single terms.

The estimate (21.11) allows it to extend the bilinear form $\tilde{s}\,(u, v)$ from the space of the infinitely differentiable functions with compact support to a bounded bilinear form on the corresponding Hilbert spaces with finite norms (21.12) and to transform equation (21.10) into a high-order equation, basically by replacing the test functions v by test functions $\mathscr{L} v$, v rapidly decreasing. If one chooses a suitable scaling parameter Ω of order κ and proceeds as in [8] or [9], one finally obtains the desired regularity theorem and can show that the regularized parts (21.9) of the eigenfunctions for eigenvalues below the essential spectrum are contained in $H^{1,1}_{\mathrm{mix}}$. Moreover,

$$\int \left\{1 + \sum_{i=1}^{N} \left|\frac{\omega_i}{\Omega}\right|^2\right\} \left\{\prod_{i=1}^{N} \left(1 + \left|\frac{\omega_i}{\Omega}\right|^2\right)\right\} |\hat{u}_0(\omega)|^2 \,\mathrm{d}\omega \;\le\; 4\,\mathrm{e}\,\|u_0\|_0^2, \tag{21.13}$$

with Euler's number $\mathrm{e} = \exp(1)$. The constant Ω fixes a characteristic lengthscale. It is independent of the eigenfunction under consideration and depends basically only on the number of electrons. The same type of estimate holds for the correspondingly regularized parts of the exponentially weighted eigenfunctions (21.5).

The spaces $H^{\vartheta,1}_{\mathrm{mix}}$, $0 < \vartheta < 1$, are interpolation spaces between H^1 and $H^{1,1}_{\mathrm{mix}}$. Interpolation spaces are defined with help of the K-functional. The K-functional of a function $u \in H^1$ in a version adapted to the given setting is

$$K(t, u) = \inf_{v\in H^{1,1}_{\mathrm{mix}}} \left\{ \|u - v\|_1^2 + t^2\, |||v|||_{1,1}^2 \right\}^{1/2}.$$

The faster $K(t, u)$ tends to zero for $t \to 0+$ the smoother u is. The K-functional is needed to define the interpolation norm of a function $u \in H^1$, which is given by

$$\|u\|^2 = \int_0^\infty [\, t^{-\vartheta} K(t, u)\,]^2 \,\frac{\mathrm{d}t}{t}.$$

It remains finite if and only if u is contained in the space $H^{\vartheta,1}_{\mathrm{mix}}$ and is in this case a fixed multiple of the original norm on this space. The proof that the eigenfunctions under consideration and their exponentially weighted counterparts are contained in the spaces (21.7) is based on this characterization of the fractional order spaces.

The complete the proof of this regularity property, one needs to estimate the speed with which the K-functional of functions

$$u(x) \;=\; \exp\Bigl(\sum_{i<j}\phi(x_i - x_j)\Bigr)u_0(x)$$

tends to zero as t goes to zero, where $u_0 \in H^{1,1}_{\mathrm{mix}}$ is here at first not nearer specified. The approximating functions v in the definition of the K-functional are for this constructed smoothing the function ϕ in the exponent properly. It turns out [6] that

$$K(t,u) \;\lesssim\; |\ln(t)|^{1/2}\, t^{3/4}\, |\!|\!| u_0 |\!|\!|_{1,1}, \quad t \to 0+,$$

and u is thus contained in $H^{\vartheta,1}_{\mathrm{mix}}$ for $\vartheta < 3/4$. As the $H^{1,0}_{\mathrm{mix}}$-norm of u can be estimated by that of u_0, the functions under consideration are therefore contained in the space (21.7). This holds in particular for the solutions of the electronic Schrödinger equation and their exponentially weighted counterparts (21.5).

21.5 The Radial-Angular Decomposition

An interesting consequence of these regularity properties is the following observation. Consider a complete L_2-orthonormal system

$$\phi_{n\ell m}(x) \;=\; \frac{1}{r}\, f_{n\ell}(r)\, Y^m_\ell(x), \quad n,\ell = 0,1,\ldots,\;\; m = -\ell,\ldots,\ell, \tag{21.14}$$

of functions from $\mathbb{R}^3$ to $\mathbb{R}$, where $r = |x|$ has been set and the Y^m_ℓ are the spherical harmonics. The joint eigenfunctions of the harmonic oscillator and the angular momentum operators L^2 and L_3 known from quantum mechanics represent an example of such a system. Every square integrable function $u : (\mathbb{R}^3)^N \to \mathbb{R}$ can then be expanded into an orthogonal series

$$u(x) \;=\; \sum_{n,\ell,m} a(n,\ell,m) \prod_{i=1}^{N} \phi_{n_i \ell_i m_i}(x_i),$$

where n, ℓ, and m are multi-indices here with components n_i, ℓ_i, and m_i. With help of this expansion, one can define the L_2-orthogonal projections

$$\big(Q(\ell,m)u\big)(x) \;=\; \sum_{n} a(n,\ell,m) \prod_{i=1}^{N} \phi_{n_i \ell_i m_i}(x_i)$$

in which the angular parts are kept fixed and the sum extends only over the radial parts. These projections can in fact be defined without recourse to the given expansion. They are not only L_2-orthogonal but also orthogonal as projections of many other L_2-like Sobolev spaces into themselves. For functions in H^1,

$$\|u\|_1^2 \;=\; \sum_{\ell,m} \|Q(\ell,m)u\|_1^2.$$

The point is that the weighted norm defined by the expression

$$|||u|||^2 \;=\; \sum_{\ell,m} \Big\{ \prod_{i=1}^{N} \Big(1 + \ell_i\,(\ell_i+1)\Big) \Big\}\, \|Q(\ell,m)u\|_1^2$$

can be estimated by the $H^{1,1}_{\mathrm{mix}}$-norm of the exponentially weighted counterpart (21.5) of u, as long this exponentially weighted counterpart is in this space. The proof uses that the functions (21.14) are eigenfunctions of the square of the angular momentum operator; see [8]. Interpolation theory thus shows that for $\vartheta < 3/4$ the expressions

$$\sum_{\ell,m} \Big\{ \prod_{i=1}^{N} \Big(1 + \ell_i\,(\ell_i+1)\Big) \Big\}^{\vartheta}\, \|Q(\ell,m)u\|_1^2$$

remain finite for the eigenfunctions u of the electronic Schrödinger operator for eigenvalues below the bottom of the essential spectrum. Thus only comparatively few of the angular parts make a significant contribution to these eigenfunctions. At least for atoms this is not truly surprising and reflects their shell structure.

21.6 Sparse Grids, Hyperbolic Cross Spaces, and Antisymmetry

Electrons have an internal property called spin that behaves similar to angular momentum. The spin of an electron can attain the two half-integer values $\pm 1/2$. Correspondingly, the complete wavefunctions are of the form

$$\psi : (\mathbb{R}^3)^N \times \{-1/2,\, 1/2\}^N \to \mathbb{R} : (x,\sigma) \to \psi(x,\sigma), \tag{21.15}$$

that is, depend not only on the positions $x_i \in \mathbb{R}^3$, but also on the spins $\sigma_i = \pm 1/2$ of the electrons. The Pauli principle, one of the fundamental principles of quantum mechanics, states that only those eigenfunctions are admissible that change their sign under a simultaneous exchange of the positions x_i and x_j and the spins σ_i and σ_j of two electrons i and j, that is, are antisymmetric in the sense that

$$\psi(Px, P\sigma) = \operatorname{sign}(P)\psi(x, \sigma)$$

holds for arbitrary simultaneous permutations $x \to Px$ and $\sigma \to P\sigma$ of the electron positions and spins. The Pauli principle forces the admissible wavefunctions to vanish where $x_i = x_j$ and $\sigma_i = \sigma_j$ for $i \neq j$, that is, that the probability that two electrons i and j with the same spin meet is zero. The admissible solutions of the scalar Schrödinger equation (21.4) are those that are components

$$u : (\mathbb{R}^3)^N \to \mathbb{R} : x \to \psi(x, \sigma)$$

of an antisymmetric wavefunction (21.15). They are antisymmetric under all permutations of the particles that leave the spin vector σ invariant.

We explain the interplay of these symmetry properties and the mixed regularity of the wavefunctions for the approximation of the solutions of the Schrödinger equation in the following by means of a simple model problem, the approximation of functions u that are odd and 2π-periodic in every coordinate direction on the axiparallel cube $Q = [0, \pi]^d$ by the tensor products

$$\phi(k, x) = \prod_{i=1}^{d} \phi_{k_i}(x_i), \quad \phi_{k_i}(\xi) = \sqrt{\frac{2}{\pi}} \sin(k_i \xi) \tag{21.16}$$

of one-dimensional trigonometric polynomials labeled by the components k_i of the multi-indices k. These tensor products form a complete L_2-orthonormal system on the cube Q. Functions of the given kind that are square integrable over Q can therefore be expanded into a multivariate Fourier series

$$u(x) = \sum_k \hat{u}(k)\phi(k, x). \tag{21.17}$$

We measure the speed of convergence of this series in the sense of the L_2-norm which reads in terms of the expansion coefficients

$$\|u\|_0^2 = \sum_k |\hat{u}(k)|^2.$$

The speed of convergence of the series is thus determined by the speed with which the expansion coefficients decay. At this place the mixed regularity comes into play.

Consider functions u that possess corresponding weak partial derivatives and set

$$|u|_{1,\mathrm{mix}}^2 = \int_Q \left| \frac{\partial^d u}{\partial x_1 \dots \partial x_d} \right|^2 \mathrm{d}x$$

or, in terms of the expansion coefficients,

$$|u|_{1,\mathrm{mix}}^2 = \sum_k \left(\prod_{i=1}^d k_i \right)^2 |\hat{u}(k)|^2.$$

Let u_ε be the function represented by the finite part of the series (21.17) that extends over the multi-indices k inside the hyperboloid or hyperbolic cross given by

$$\prod_{i=1}^d k_i < \frac{1}{\varepsilon}. \tag{21.18}$$

The L_2-error can then be estimated as

$$\|u - u_\varepsilon\|_0 \leq \varepsilon |u|_{1,\mathrm{mix}}$$

and tends like $\mathcal{O}(\varepsilon)$ to zero as ε goes to zero. The dimension n of the space spanned by the functions (21.16) for which (21.18) holds increases for ε tending to zero like

$$n \sim |\log \varepsilon|^{d-1} \varepsilon^{-1}.$$

This shows that already a comparatively slow growth of the smoothness with the space dimension can help to reduce the complexity of a high-dimensional approximation problem substantially, an observation that forms the basis of the sparse grid or hyperbolic cross techniques. Due to the logarithmic term, the applicability of such methods is, however, in general limited to moderate space dimensions.

Symmetry properties as enforced by the Pauli principle represent a possibility to escape from this curse of dimensionality without forcing up the smoothness requirements further, a fact that has first been noted by Hackbusch [2]. Consider functions u that are antisymmetric with respect to the exchange of their variables, i.e., that

$$u(Px) = \mathrm{sign}(P) u(x)$$

holds for all permutation matrices P. It is not astonishing that such symmetry properties are immediately reflected in the expansion (21.17). Let

$$\tilde{\phi}(k, x) = \frac{1}{\sqrt{d!}} \sum_P \mathrm{sign}(P) \phi(k, Px) \tag{21.19}$$

be the renormalized, antisymmetric parts of the functions (21.16), where the sums extend over the $d!$ permutation matrices P of order d. The antisymmetrized functions (21.19) can be written as determinants

$$\frac{1}{\sqrt{d!}} \begin{vmatrix} \phi_{k_1}(x_1) & \dots & \phi_{k_d}(x_1) \\ \vdots & \ddots & \vdots \\ \phi_{k_1}(x_d) & \dots & \phi_{k_d}(x_d) \end{vmatrix}$$

and evaluated in this way. For the functions u in the given symmetry class, many terms in the expansion (21.17) can be combined. It finally collapses into

$$u(x) \;=\; \sum_{k_1>\ldots>k_d} \big(u, \tilde\phi(k,\cdot)\big)\tilde\phi(k,x),$$

where the expansion coefficients are the L_2-inner products of u with the corresponding functions (21.19). The number of basis functions needed to reach a given accuracy is reduced by more than the factor $d!$, a significant gain for larger dimensions d.

It remains to count the number of the sequences $k_1 > k_2 > \ldots > k_d$ of natural numbers that satisfy the condition (21.18) and with that also the number of basis function (21.19) needed to reach the accuracy $\mathcal{O}(\varepsilon)$. To study the asymptotic behavior of the number of these sequences in dependence of the dimension d and of ε, it suffices when we restrict ourselves to the case $\varepsilon = 1/2^L$, with positive integers L. That is, we have to give bounds for the number of sequences $k_1 > \ldots > k_d$ for which

$$\prod_{i=1}^{d} k_i \;\le\; 2^L. \tag{21.20}$$

The problem to estimate this number has to do with the prime factorization of integers. To simplify this problem, we group the numbers k_i into levels and decompose the space of the trigonometric polynomials correspondingly. Let

$$\ell(k_i) \;=\; \max\big\{\, \ell \in \mathbb{Z} \;\big|\; 2^{\,\ell} \le k_i \,\big\}.$$

An upper bound for the number of these sequences is then the number $a(d,L)$ of the sequences $k_1 > k_2 > \ldots > k_d$ of natural numbers for which

$$\prod_{i=1}^{d} 2^{\,\ell(k_i)} \;\le\; 2^L.$$

The numbers $a(d,L)$ can be calculated recursively; see [8] for details. A crude estimate yields $a(d,L) = 0$ if $L + 1 < d$. Thus

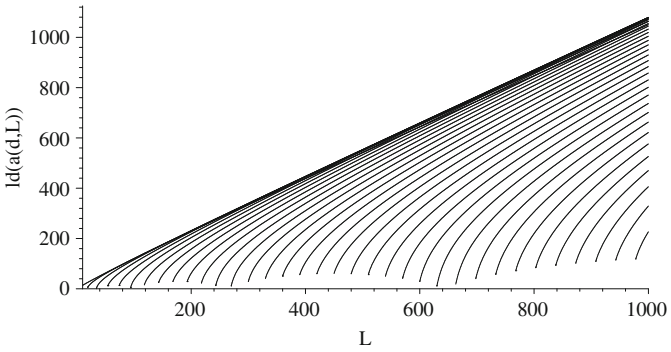


Fig. 21.1 The numbers $a^*(L)$ and $a(d, L)$ for $d = 10, 15, 20, \ldots, 175$

$$a^*(L) := \max_{d \geq 1} a(d, L) = \max_{d \leq L+1} a(d, L). \tag{21.21}$$

Figure 21.1 shows, in logarithmic scale, how the $a(d, L)$ behave compared to their joint least upper bound $a^*(L)$. It becomes obvious from this picture that this upper bound exceeds the actual dimensions for larger d by many orders of magnitude, the more the more the number d of variables increases. The joint least upper bound that is independent of d for the number of the sequences $k_1 > \ldots > k_d$ of natural numbers k_i for which (21.20) holds grows at least like $\sim 2^L$ since already for the case $d = 1$, there are 2^L such "sequences", namely those with values $k_1 = 1, \ldots, 2^L$. Figure 21.1 suggests conversely that the upper bound (21.21) for the number of these sequences does not grow much faster than $\sim 2^L$. This is in fact the case since the number of the decreasing infinite sequences $k_1 \geq k_2 \geq k_3 \geq \ldots$ of natural numbers for which

$$\prod_{i=1}^{\infty} 2^{\ell(k_i)} \leq 2^L, \tag{21.22}$$

with L a given nonnegative integer, is bounded by

$$\sum_{\ell=0}^{L} p(\ell)\, 2^{\ell}, \tag{21.23}$$

where $p(\ell)$ denotes the partition number of ℓ, the number of possibilities of representing ℓ as sum of nonnegative integers without regard to the order. To show this, we observe that the number of these sequences is bounded by the number of sequences $k_1, k_2, k_3, \ldots$ of natural numbers for which their levels $\ell(k_1), \ell(k_2), \ldots$ decrease and that satisfy (21.22). We show that the expression (21.23) counts the number of these sequences. Let the integers $\ell_i = \ell(k_i)$ first

be given. As there are 2^{ℓ_i} natural numbers k_i for which $\ell(k_i) = \ell_i$, namely $k_i = 2^{\ell_i}, \dots, 2^{\ell_i+1} - 1$, there are

$$\prod_{i=1}^{\infty} 2^{\ell_i} = 2^{\ell}, \quad \ell = \sum_{i=1}^{\infty} \ell_i,$$

sequences $k_1, k_2, k_3, \dots$ for which the $\ell(k_i)$ attain the prescribed values ℓ_i. The problem thus reduces to the question how many decreasing sequences of nonnegative integers ℓ_i exist that sum up to values $\ell \le L$, i.e., for which

$$\sum_{i=1}^{\infty} \ell_i = \ell.$$

This number is by definition the partition number $p(\ell)$ of the nonnegative integer ℓ. Every sequence $k_1 > k_2 > \dots > k_d$ of natural numbers for which (21.20) holds can obviously be expanded to an infinite, decreasing sequence $k_1 \ge k_2 \ge k_3 \ge \dots$ of natural numbers that satisfies (21.22) by setting all $k_i = 1$ for $i > d$. The sum (21.23) represents therefore also an upper bound for the number of these sequences.

The partition number plays a big role in combinatorics. Hardy and Ramanujan [3] have shown that it behaves asymptotically like

$$p(\ell) \sim \frac{\exp\left(\pi \sqrt{2\ell/3}\right)}{\ell}$$

as ℓ goes to infinity. We conclude that the upper bound (21.21) for the number of determinants needed to reach an error $\le 2^{-L}|u|_{1,\mathrm{mix}}$ behaves like

$$a^*(L) = (2^L)^{1+\delta(L)}, \quad 0 \le \delta(L) \le cL^{-1/2},$$

where c is a constant that depends neither on L nor on the space dimension d or the function u. Using the representation (21.21) of $a^*(L)$ and the recursively calculated values $a(d, L)$, the exponents $1 + \delta(L)$ can be calculated. They decay for L ranging from 10 to 1,000 monotonously from 1.406 to 1.079. For $L = 100$, $1 + \delta(L) = 1.204$. That is, the error tends faster to zero in the number n of determinants than

$$\sim \frac{1}{n^{1-\vartheta}}$$

for any given ϑ in the interval $0 < \vartheta < 1$. Not only does the convergence rate deteriorate neither with the dimension nor the number of variables, it behaves asymptotically almost as in the one-dimensional case. Similar results hold for partially antisymmetric functions as they occur in quantum mechanics.

21.7 Eigenfunction and Wavelet Expansions

The constructions sketched in the previous section transfer to the more complicated case of the expansion of the solutions of the electronic Schrödinger equation into correspondingly antisymmetrized tensor products of three-dimensional Hermite functions or other eigenfunctions of three-dimensional Schrödinger-like operators as in [8] or wavelets as in [10]. Indeed, it finally turns out that the convergence rate measured in terms of the number of basis functions involved does not deteriorate with the number of electrons and comes close to that for the two- or even one-particle case. We do not explicate the partly technical details here but explain how one can utilize the intermediate smoothness of the exponentially weighted solutions (21.5) to obtain optimal convergence rates.

Let e^F be exponential factor in (21.5). The argumentation starts from functions v whose exponentially weighted counterparts $\mathrm{e}^F v$ are located in $H^{1,1}_{\mathrm{mix}}$, that is, have in contrast to the solutions of the Schrödinger equation full mixed regularity. The essential observation is that the norm $|||\mathrm{e}^F v|||_{1,1}$ can be estimated by the sum of the weighted L_2-norms $\|\mathrm{e}^F \mathrm{D}^\alpha v\|_0$ of the involved derivatives $\mathrm{D}^\alpha v$ of v and vice versa. This comes from the special structure of the function F. The norm $|||\mathrm{e}^F v|||_{1,1}$ measures therefore the exponentially weighted L_2-norms of the involved derivatives of v. It is therefore reasonable to start from a sequence $T_n : H^1 \to H^1$, $n = 1, 2, \ldots$, of linear approximation operators that are uniformly H^1-bounded and to require that

$$\|v - T_n v\|_1 \lesssim n^{-q} \, |||\mathrm{e}^F v|||_{1,1} \tag{21.24}$$

for all functions $v \in H^1$ for which $\mathrm{e}^F v \in H^{1,1}_{\mathrm{mix}}$. The constant $q > 0$ is an unspecified convergence rate also depending on what n means. These assumptions form a proper framework for sparse grid-like approximation methods as those mentioned above modeled after the example from the last section. Another example is the expansion into tensor products of three-dimensional functions with given angular parts as described in Sect. 21.5. The range of the operators T_n is in this case infinite dimensional. The exponential factor is the tribute paid to the infinite extension of the domain. The assumption (21.24) implies for the functions $u \in H^1$ whose exponentially weighted counterparts $\mathrm{e}^F u$ are in $H^{\vartheta,1}_{\mathrm{mix}}$ for some ϑ between 0 and 1, the error estimate

$$\|u - T_n u\|_1 \lesssim n^{-\vartheta q} \, |||\mathrm{e}^F u|||_{\vartheta,1}.$$

The proof utilizes the characterization of the spaces $H^{\vartheta,1}_{\mathrm{mix}}$ as interpolation spaces.

We conclude that for the case of the solutions u of the Schrödinger equation the H^1-error $\|u - T_n u\|_1$ tends faster to zero as $n^{-\vartheta q}$ for any $\vartheta < 3/4$. An estimate directly based on an estimate of their K-functional even shows that

$$\|u - T_n u\|_1 \lesssim \sqrt{\ln(n)} \, n^{-3/4\, q}$$

so that up to the logarithmic term only the factor $3/4$ gets lost compared to the case of full mixed regularity. The estimate is optimal, at least up to the logarithmic factor, and can in general not be improved further.

References

1. Fournais, S., Hoffmann-Ostenhof, M., Hoffmann-Ostenhof, T., Østergard Sørensen, T.: Sharp regularity estimates for Coulombic many-electron wave functions. Commun. Math. Phys. **255**, 183–227 (2005)
2. Hackbusch, W.: The efficient computation of certain determinants arising in the treatment of Schrödinger's equation. Computing **67**, 35–56 (2000)
3. Hardy, G., Ramanujan, S.: Asymptotic formulae in combinatory analysis. Proc. Lond. Math. Soc. **17**, 75–115 (1918)
4. Hoffmann-Ostenhof, M., Hoffmann-Ostenhof, T., Østergard Sørensen, T.: Electron wavefunctions and densities for atoms. Ann. Henri Poincaré **2**, 77–100 (2001)
5. Hylleraas, E.: Neue Berechnung der Energie des Heliums im Grundzustande, sowie des tiefsten Terms von Ortho-Helium. Z. Phys. **54**, 347–366 (1929)
6. Kreusler, H., Yserentant, H.: The mixed regularity of electronic wave functions in fractional order and weighted Sobolev spaces. Numer. Math. **121**, 781–802 (2012)
7. Yserentant, H.: On the regularity of the electronic Schrödinger equation in Hilbert spaces of mixed derivatives. Numer. Math. **98**, 731–759 (2004)
8. Yserentant, H.: Regularity and Approximability of Electronic Wave Functions. Lecture Notes in Mathematics, vol. 2000. Springer, Heidelberg/Dordrecht/London/New York (2010)
9. Yserentant, H.: The mixed regularity of electronic wave functions multiplied by explicit correlation factors. ESAIM M2AN **45**, 803–824 (2011)
10. Zeiser, A.: Wavelet approximation in weighted Sobolev spaces of mixed order with applications to the electronic Schrödinger equation. Constr. Approx. **35**, 293–322 (2012)

Index

Adaptivity, 189
  adaptive multiwavelet methods, 321
  adaptive tensor sampling, 196, 202
  adaptive wavelet scheme, 104
Alternating directions fitting, 208
Alternating linear scheme (ALS), 254
Alternating minimisation, 208
Approximation
  best k-term approximation, 374
  dynamic programming, 2
  linear approximation, 101
  manifold, 382
  nonlinear approximation, 101
  quasi-best approximation, 199
  rate, 79
  trigonometric reconstruction, 348
Atoms, 367

Backward stochastic differential equation, 7
Balanced graph cuts, 264, 273, 275
Bermudan option, 2
Besov space, 90
  Besov scale, 84
Biorthogonal systems, 373
Black-Scholes PDE, 322
Born-Oppenheimer approximation, 413
Burn-in, 402
Butterfly graph, 351
Butterfly scheme, 351
  error estimate, 352

Canonical (CP) format, 197
Cheeger's inequality, 404
Chemical master equation (CME), 306
Clarke subdifferential, 266
Collateralized Debt Obligations (CDOs), 321, 338
Compressive sensing, 366, 372
  error estimates, 374, 377
  finite dimensional, 372
  ill-posed problems, 376
  infinite dimensional, 366, 374
  sensing operator, 373, 375
Constrained optimization, 375
  accelerated iteration, 375
  generalized soft-shrinkage, 375
  projected iteration, 375
Credit value adjustment, 20
Cross approximation, 203
Curvature of low-rank matrix manifold, 386

Data-sparse approximation, 196
Dictionary, 367
Dimension tree, 199
Dimensionality reduction, 292
Dirac–Frenkel variational principle, 255, 382
Distance to instability, 394
Domain decomposition, 71
Doob decomposition, 4
Dynamic programming, 1
Dynamical low-rank approximation, 384
Dynamical tensor approximation, 389

Easy path wavelet transform (EPWT), 282
  decomposition, 283
  diameter condition, 289

S. Dahlke et al. (eds.), *Extraction of Quantifiable Information from Complex Systems*, Lecture Notes in Computational Science and Engineering 102, DOI 10.1007/978-3-319-08159-5

path construction, 284, 286
reconstruction, 284
region condition, 289
scattered data denoising, 287
Eigenproblem
nonlinear, critical point, 272
Eigenvalue optimisation, 394
Elliptic boundary value problem, 103
Embryogenesis, 54
Euler scheme
drift-implicit, 123
jump-adapted, 126
Evolution equation, 56
model, 54
solvability, 60
Exact recovery condition (ERC), 366, 368
Neumann ERC, 368, 369, 371
in the presence of noise, 369
Exact relaxation, 265
g-expectation, 18
L_1-exponential convergence, 400
Exponential decay
of electronic wavefunctions, 416
Extension operator, 75
hestenes extension, 77

Fast Fourier algorithms
FFT, 349
HCFFT, 349
NFFT, 349
Fast Fourier transform (FFT), 349
generated set, 350
inverse, 353
rank-1 lattice, 350
Fichera corner domain, 79
Financial products, 321
Finite basis injectivity property, 370
Fourier coefficients, 347
Fourier matrix, 348
equispaced, 353
nonequispaced, 353
Vandermonde structure, 358
Frame
recovery, 372
sensing, sampling, 372
Frequency index set, 347
l_p ball, 356
difference set, 355
full grid, 348
hyperbolic cross, 349, 357
energy-norm based, 349, 357
Function spaces, 55
solution spaces, 55

Gaussian processes, 183
Geometric ergodicity, 401

Heston model, 109, 110, 118
Hierarchical rank, 200
Hierarchical singular value decomposition, 202
Hierarchical tensor format, 199
Hierarchical tensor representation, 243, 391
Hierarchical Tucker format, 244, 321, 330, 391
High dimensions, 195, 321
High-Order Singular Value Decomposition (HOSVD), 199, 242
Hit-and-run algorithm, 404
HT format, 244, 391
Hyperbolic cross Fourier transform, 349
condition number, 354
fast (HCFFT), 349
inverse, 354

Ill-posed problem, 365
compressed sensing, 376
error estimates, 370, 377
Implicit function theorem, 62
Independent component analysis, 298
Inverse problem, 365

Joint sparsity measure, 374

K-sparse, 373
Kallianpur–Striebel formula, 166
Kullback–Leibler distance, 298

L-shaped domain, 79
Least-squares Monte Carlo, 8
Linear differential operator, 56
Lovasz extension, 264, 274
Low-rank tensor, 197

Malliavin calculus, 109, 111, 118, 119
Markov chain, 321, 397
Matching pursuit, 366
orthogonal, 367
Matricisation, 198, 242
Matrix differential equation, 385
Matrix nearness problems, 394
Matrix product representation, 245
Matrix product state, 245, 391
Maximum likelihood, 164

Metropolis-Hastings algorithm, 405
Minimal subspace, 241
Mixed regularity
of electronic wavefunctions, 417
Modulation manifold, 294
Modulation map, 294
Monte Carlo methods, 1
MPS format, 200
Multi-configuration time-dependent Hartree (MCTDH) method, 392
Multidimensional scaling, 293
Multilevel Monte Carlo, 11, 109, 111
central limit theorem, 129
discontinuous payoff, 118
Lévy SDE, 125
Multivariate trigonometric polynomial, 347
evaluation, 347, 348
reconstruction, 348, 352, 353, 355

Nested approximation, 204
Nestedness property, 201
Noise-to-signal ratio, 369, 371
Non-adaptive sampling, 206
Non-negative matrix factorization, 298
Non-negative PCA, 297
Nonequispaced fast Fourier transform (NFFT), 349
Nonlinear dimensionality reduction, 293
Nonlinear option pricing, 2
Nonlinear SOR, 208

Operator equation, 365
Optimal stopping problem, 4
Optimisation, 254, 394
convex optimisation, 1
nonconvex, 265

Parabolic regularity, 61
Parameter-dependent PDE, 196
Parameter identification, 53
Parameter rule, 370
Parameter-to-state map, 61
differentiability, 62
Lipschitz derivative, 64, 66
properties, 63
Parametric diffusion, 205
Parametric PDE, 205
Parametrix method, 114
Partial differential equation, 321
Pauli principle, 414
Payoff-splitting, 111, 118, 122
Primal-dual approach, 3
Principal component analysis, 293
Projector-splitting integrator, 387, 392
Prony method, 360
Propensity function, 305
Pseudospectrum, 394

Quantization of SDEs, 111, 114
Itô–Taylor step, 114, 115, 118
minimal errors, 113, 114
support reduction, 115, 116
Quantum dynamics, 392

Radial-angular decomposition
of electronic wavefunctions, 421
Random functions, 98
Besov regularity, 99
wavelet expansions, 99
Random sampling, 359
Random tensor, 208
Reaction channel, 303, 304
Reconstructing generated set, 358
Reconstructing rank-1 lattice, 355
component–by–component construction, 355, 356
Regularization, 365, 370
Reinforcement learning, 181
Restricted isometry property, 370, 373
Reversibility, 398
Riemannian manifold, 253
Robust stability, 394
Rough path theory, 164

Sampling set, 347
generated set, 350
random nodes, 359
rank-1 lattice, 350
sparse grid, 349
Scattered data denoising, 287
Schrödinger equation, 393, 413
electronic, 414
Semi-Lagrangian scheme, 190
Sensing operator, 373, 375
Sieve, 14
Signal detection, 295
Signal separation, 295
Slit domain, 79
Sparse grid approximation
of antisymmetric functions, 423
of electronic wavefunctions, 427
Sparse grids, 187

Sparse trigonometric recovery, 353, 360
  condition number, 359
Sparsity
  joint sparsity measure, 374
  sparse recovery, 366
  sparse representation, 365
  sparsity constraint, 366, 369
SPDE, 89
  Hölder-Besov regularity, 95
  spatial Besov regularity, 92
  weighted Sobolev regularity, 93
L_2-spectral gap, 400
Stable convergence, 128
Stochastic differential equation, 110
  driven by Lévy process, 124
  non-Lipschitz coefficients, 118, 120
Stochastic dynamic program, 1
Stochastic filtering, 162
Stochastic parabolic space, 87
Stoichiometric vector, 305
Subspace approximation, 240
Superposition operators, 57

Tangent space, 249
Tangent space projection, 382, 385
Tensor, 195, 238, 256
  completion, 196, 207
  compression, 246
  manifold, 248, 389
  product, 238, 239
  rank, 197
  sampling, 196
  space, 239
  train, 245, 391
Tensor train (TT) format, 200, 245, 391
Time-dependent variational principle, 382
Total variation distance, 399
Transition kernel, 398
Tree adaptivity, 206
Tree agglomeration, 206
Tucker format, 198, 240, 389
Tucker rank, 198, 241

Uniform ergodicity, 401

Value function, 182
Variational splitting integrator, 393

Wavelet basis, 72, 90
  wavelets, 71, 72
Weak solution, 60
Weighted Sobolev space, 73, 86
  embedding, 87

Editorial Policy

1. Volumes in the following three categories will be published in LNCSE:

i) Research monographs
ii) Tutorials
iii) Conference proceedings

Those considering a book which might be suitable for the series are strongly advised to contact the publisher or the series editors at an early stage.

2. Categories i) and ii). Tutorials are lecture notes typically arising via summer schools or similar events, which are used to teach graduate students. These categories will be emphasized by Lecture Notes in Computational Science and Engineering. **Submissions by interdisciplinary teams of authors are encouraged.** The goal is to report new developments – quickly, informally, and in a way that will make them accessible to non-specialists. In the evaluation of submissions timeliness of the work is an important criterion. Texts should be well-rounded, well-written and reasonably self-contained. In most cases the work will contain results of others as well as those of the author(s). In each case the author(s) should provide sufficient motivation, examples, and applications. In this respect, Ph.D. theses will usually be deemed unsuitable for the Lecture Notes series. Proposals for volumes in these categories should be submitted either to one of the series editors or to Springer-Verlag, Heidelberg, and will be refereed. A provisional judgement on the acceptability of a project can be based on partial information about the work: a detailed outline describing the contents of each chapter, the estimated length, a bibliography, and one or two sample chapters – or a first draft. A final decision whether to accept will rest on an evaluation of the completed work which should include

– at least 100 pages of text;
– a table of contents;
– an informative introduction perhaps with some historical remarks which should be accessible to readers unfamiliar with the topic treated;
– a subject index.

3. Category iii). Conference proceedings will be considered for publication provided that they are both of exceptional interest and devoted to a single topic. One (or more) expert participants will act as the scientific editor(s) of the volume. They select the papers which are suitable for inclusion and have them individually refereed as for a journal. Papers not closely related to the central topic are to be excluded. Organizers should contact the Editor for CSE at Springer at the planning stage, see *Addresses* below.

In exceptional cases some other multi-author-volumes may be considered in this category.

4. Only works in English will be considered. For evaluation purposes, manuscripts may be submitted in print or electronic form, in the latter case, preferably as pdf- or zipped ps-files. Authors are requested to use the LaTeX style files available from Springer at http://www.springer.com/gp/authors-editors/book-authors-editors/manuscript-preparation/5636 (Click on LaTeX Template → monographs or contributed books).

For categories ii) and iii) we strongly recommend that all contributions in a volume be written in the same LaTeX version, preferably LaTeX2e. Electronic material can be included if appropriate. Please contact the publisher.

Careful preparation of the manuscripts will help keep production time short besides ensuring satisfactory appearance of the finished book in print and online.

5. The following terms and conditions hold. Categories i), ii) and iii):

Authors receive 50 free copies of their book. No royalty is paid.
Volume editors receive a total of 50 free copies of their volume to be shared with authors, but no royalties.

Authors and volume editors are entitled to a discount of 33.3 % on the price of Springer books purchased for their personal use, if ordering directly from Springer.

6. Springer secures the copyright for each volume.

Addresses:

Timothy J. Barth
NASA Ames Research Center
NAS Division
Moffett Field, CA 94035, USA
barth@nas.nasa.gov

Michael Griebel
Institut für Numerische Simulation
der Universität Bonn
Wegelerstr. 6
53115 Bonn, Germany
griebel@ins.uni-bonn.de

David E. Keyes
Mathematical and Computer Sciences
and Engineering
King Abdullah University of Science
and Technology
P.O. Box 55455
Jeddah 21534, Saudi Arabia
david.keyes@kaust.edu.sa

and

Department of Applied Physics
and Applied Mathematics
Columbia University
500 W. 120 th Street
New York, NY 10027, USA
kd2112@columbia.edu

Risto M. Nieminen
Department of Applied Physics
Aalto University School of Science
and Technology
00076 Aalto, Finland
risto.nieminen@aalto.fi

Dirk Roose
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
3001 Leuven-Heverlee, Belgium
dirk.roose@cs.kuleuven.be

Tamar Schlick
Department of Chemistry
and Courant Institute
of Mathematical Sciences
New York University
251 Mercer Street
New York, NY 10012, USA
schlick@nyu.edu

Editor for Computational Science
and Engineering at Springer:
Martin Peters
Springer-Verlag
Mathematics Editorial IV
Tiergartenstrasse 17
69121 Heidelberg, Germany
martin.peters@springer.com

Lecture Notes in Computational Science and Engineering

1. D. Funaro, *Spectral Elements for Transport-Dominated Equations.*
2. H.P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.
3. W. Hackbusch, G. Wittum (eds.), *Multigrid Methods V.*
4. P. Deuflhard, J. Hermans, B. Leimkuhler, A.E. Mark, S. Reich, R.D. Skeel (eds.), *Computational Molecular Dynamics: Challenges, Methods, Ideas.*
5. D. Kröner, M. Ohlberger, C. Rohde (eds.), *An Introduction to Recent Developments in Theory and Numerics for Conservation Laws.*
6. S. Turek, *Efficient Solvers for Incompressible Flow Problems.* An Algorithmic and Computational Approach.
7. R. von Schwerin, ***M**ulti **B**ody **S**ystem **SIM**ulation.* Numerical Methods, Algorithms, and Software.
8. H.-J. Bungartz, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*
9. T.J. Barth, H. Deconinck (eds.), *High-Order Methods for Computational Physics.*
10. H.P. Langtangen, A.M. Bruaset, E. Quak (eds.), *Advances in Software Tools for Scientific Computing.*
11. B. Cockburn, G.E. Karniadakis, C.-W. Shu (eds.), *Discontinuous Galerkin Methods.* Theory, Computation and Applications.
12. U. van Rienen, *Numerical Methods in Computational Electrodynamics.* Linear Systems in Practical Applications.
13. B. Engquist, L. Johnsson, M. Hammill, F. Short (eds.), *Simulation and Visualization on the Grid.*
14. E. Dick, K. Riemslagh, J. Vierendeels (eds.), *Multigrid Methods VI.*
15. A. Frommer, T. Lippert, B. Medeke, K. Schilling (eds.), *Numerical Challenges in Lattice Quantum Chromodynamics.*
16. J. Lang, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems.* Theory, Algorithm, and Applications.
17. B.I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition.*
18. U. van Rienen, M. Günther, D. Hecht (eds.), *Scientific Computing in Electrical Engineering.*
19. I. Babuška, P.G. Ciarlet, T. Miyoshi (eds.), *Mathematical Modeling and Numerical Simulation in Continuum Mechanics.*
20. T.J. Barth, T. Chan, R. Haimes (eds.), *Multiscale and Multiresolution Methods.* Theory and Applications.
21. M. Breuer, F. Durst, C. Zenger (eds.), *High Performance Scientific and Engineering Computing.*
22. K. Urban, *Wavelets in Numerical Simulation.* Problem Adapted Construction and Applications.
23. L.F. Pavarino, A. Toselli (eds.), *Recent Developments in Domain Decomposition Methods.*

24. T. Schlick, H.H. Gan (eds.), *Computational Methods for Macromolecules: Challenges and Applications.*

25. T.J. Barth, H. Deconinck (eds.), *Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics*.

26. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations*.

27. S. Müller, *Adaptive Multiscale Schemes for Conservation Laws*.

28. C. Carstensen, S. Funken, W. Hackbusch, R.H.W. Hoppe, P. Monk (eds.), *Computational Electromagnetics*.

29. M.A. Schweitzer, *A Parallel Multilevel Partition of Unity Method for Elliptic Partial Differential Equations*.

30. T. Biegler, O. Ghattas, M. Heinkenschloss, B. van Bloemen Waanders (eds.), *Large-Scale PDE-Constrained Optimization*.

31. M. Ainsworth, P. Davies, D. Duncan, P. Martin, B. Rynne (eds.), *Topics in Computational Wave Propagation*. Direct and Inverse Problems.

32. H. Emmerich, B. Nestler, M. Schreckenberg (eds.), *Interface and Transport Dynamics.* Computational Modelling.

33. H.P. Langtangen, A. Tveito (eds.), *Advanced Topics in Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming.

34. V. John, *Large Eddy Simulation of Turbulent Incompressible Flows.* Analytical and Numerical Results for a Class of LES Models.

35. E. Bänsch (ed.), *Challenges in Scientific Computing - CISC 2002.*

36. B.N. Khoromskij, G. Wittum, *Numerical Solution of Elliptic Differential Equations by Reduction to the Interface.*

37. A. Iske, *Multiresolution Methods in Scattered Data Modelling.*

38. S.-I. Niculescu, K. Gu (eds.), *Advances in Time-Delay Systems.*

39. S. Attinger, P. Koumoutsakos (eds.), *Multiscale Modelling and Simulation.*

40. R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Wildlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering.*

41. T. Plewa, T. Linde, V.G. Weirs (eds.), *Adaptive Mesh Refinement – Theory and Applications.*

42. A. Schmidt, K.G. Siebert, *Design of Adaptive Finite Element Software.* The Finite Element Toolbox ALBERTA.

43. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations II.*

44. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Methods in Science and Engineering.*

45. P. Benner, V. Mehrmann, D.C. Sorensen (eds.), *Dimension Reduction of Large-Scale Systems.*

46. D. Kressner, *Numerical Methods for General and Structured Eigenvalue Problems.*

47. A. Boriçi, A. Frommer, B. Joó, A. Kennedy, B. Pendleton (eds.), *QCD and Numerical Analysis III*.

48. F. Graziani (ed.), *Computational Methods in Transport.*

49. B. Leimkuhler, C. Chipot, R. Elber, A. Laaksonen, A. Mark, T. Schlick, C. Schütte, R. Skeel (eds.), *New Algorithms for Macromolecular Simulation.*

50. M. Bücker, G. Corliss, P. Hovland, U. Naumann, B. Norris (eds.), *Automatic Differentiation: Applications, Theory, and Implementations.*

51. A.M. Bruaset, A. Tveito (eds.), *Numerical Solution of Partial Differential Equations on Parallel Computers.*

52. K.H. Hoffmann, A. Meyer (eds.), *Parallel Algorithms and Cluster Computing.*

53. H.-J. Bungartz, M. Schäfer (eds.), *Fluid-Structure Interaction.*

54. J. Behrens, *Adaptive Atmospheric Modeling.*

55. O. Widlund, D. Keyes (eds.), *Domain Decomposition Methods in Science and Engineering XVI.*

56. S. Kassinos, C. Langer, G. Iaccarino, P. Moin (eds.), *Complex Effects in Large Eddy Simulations.*

57. M. Griebel, M.A Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations III.*

58. A.N. Gorban, B. Kégl, D.C. Wunsch, A. Zinovyev (eds.), *Principal Manifolds for Data Visualization and Dimension Reduction.*

59. H. Ammari (ed.), *Modeling and Computations in Electromagnetics: A Volume Dedicated to Jean-Claude Nédélec.*

60. U. Langer, M. Discacciati, D. Keyes, O. Widlund, W. Zulehner (eds.), *Domain Decomposition Methods in Science and Engineering XVII.*

61. T. Mathew, *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations.*

62. F. Graziani (ed.), *Computational Methods in Transport: Verification and Validation.*

63. M. Bebendorf, *Hierarchical Matrices.* A Means to Efficiently Solve Elliptic Boundary Value Problems.

64. C.H. Bischof, H.M. Bücker, P. Hovland, U. Naumann, J. Utke (eds.), *Advances in Automatic Differentiation.*

65. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations IV*.

66. B. Engquist, P. Lötstedt, O. Runborg (eds.), *Multiscale Modeling and Simulation in Science.*

67. I.H. Tuncer, Ü. Gülcat, D.R. Emerson, K. Matsuno (eds.), *Parallel Computational Fluid Dynamics 2007.*

68. S. Yip, T. Diaz de la Rubia (eds.), *Scientific Modeling and Simulations*.

69. A. Hegarty, N. Kopteva, E. O'Riordan, M. Stynes (eds.), *BAIL* 2008 – *Boundary and Interior Layers.*

70. M. Bercovier, M.J. Gander, R. Kornhuber, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XVIII.*

71. B. Koren, C. Vuik (eds.), *Advanced Computational Methods in Science and Engineering.*

72. M. Peters (ed.), *Computational Fluid Dynamics for Sport Simulation.*

73. H.-J. Bungartz, M. Mehl, M. Schäfer (eds.), *Fluid Structure Interaction II - Modelling, Simulation, Optimization.*

74. D. Tromeur-Dervout, G. Brenner, D.R. Emerson, J. Erhel (eds.), *Parallel Computational Fluid Dynamics 2008.*

75. A.N. Gorban, D. Roose (eds.), *Coping with Complexity: Model Reduction and Data Analysis.*

76. J.S. Hesthaven, E.M. Rønquist (eds.), *Spectral and High Order Methods for Partial Differential Equations.*

77. M. Holtz, *Sparse Grid Quadrature in High Dimensions with Applications in Finance and Insurance*.

78. Y. Huang, R. Kornhuber, O.Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XIX*.

79. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations V*.

80. P.H. Lauritzen, C. Jablonowski, M.A. Taylor, R.D. Nair (eds.), *Numerical Techniques for Global Atmospheric Models*.

81. C. Clavero, J.L. Gracia, F.J. Lisbona (eds.), *BAIL 2010 – Boundary and Interior Layers, Computational and Asymptotic Methods*.

82. B. Engquist, O. Runborg, Y.R. Tsai (eds.), *Numerical Analysis and Multiscale Computations*.

83. I.G. Graham, T.Y. Hou, O. Lakkis, R. Scheichl (eds.), *Numerical Analysis of Multiscale Problems*.

84. A. Logg, K.-A. Mardal, G. Wells (eds.), *Automated Solution of Differential Equations by the Finite Element Method*.

85. J. Blowey, M. Jensen (eds.), *Frontiers in Numerical Analysis - Durham 2010*.

86. O. Kolditz, U.-J. Gorke, H. Shao, W. Wang (eds.), *Thermo-Hydro-Mechanical-Chemical Processes in Fractured Porous Media - Benchmarks and Examples*.

87. S. Forth, P. Hovland, E. Phipps, J. Utke, A. Walther (eds.), *Recent Advances in Algorithmic Differentiation*.

88. J. Garcke, M. Griebel (eds.), *Sparse Grids and Applications*.

89. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VI*.

90. C. Pechstein, *Finite and Boundary Element Tearing and Interconnecting Solvers for Multiscale Problems*.

91. R. Bank, M. Holst, O. Widlund, J. Xu (eds.), *Domain Decomposition Methods in Science and Engineering XX*.

92. H. Bijl, D. Lucor, S. Mishra, C. Schwab (eds.), *Uncertainty Quantification in Computational Fluid Dynamics*.

93. M. Bader, H.-J. Bungartz, T. Weinzierl (eds.), *Advanced Computing*.

94. M. Ehrhardt, T. Koprucki (eds.), *Advanced Mathematical Models and Numerical Techniques for Multi-Band Effective Mass Approximations*.

95. M. Azaïez, H. El Fekih, J.S. Hesthaven (eds.), *Spectral and High Order Methods for Partial Differential Equations ICOSAHOM 2012*.

96. F. Graziani, M.P. Desjarlais, R. Redmer, S.B. Trickey (eds.), *Frontiers and Challenges in Warm Dense Matter*.

97. J. Garcke, D. Pflüger (eds.), *Sparse Grids and Applications – Munich 2012*.

98. J. Erhel, M. Gander, L. Halpern, G. Pichot, T. Sassi, O. Widlund (eds.), *Domain Decomposition Methods in Science and Engineering XXI*.

99. R. Abgrall, H. Beaugendre, P.M. Congedo, C. Dobrzynski, V. Perrier, M. Ricchiuto (eds.), *High Order Nonlinear Numerical Methods for Evolutionary PDEs - HONOM 2013*.

100. M. Griebel, M.A. Schweitzer (eds.), *Meshfree Methods for Partial Differential Equations VII*.

101. R. Hoppe (ed.), *Optimization with PDE Constraints - OPTPDE 2014*.

102. S. Dahlke, W. Dahmen, M. Griebel, W. Hackbusch, K. Ritter, R. Schneider, C. Schwab, H. Yserentant (eds.), *Extraction of Quantifiable Information from Complex Systems*.

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/3527

Monographs in Computational Science and Engineering

1. J. Sundnes, G.T. Lines, X. Cai, B.F. Nielsen, K.-A. Mardal, A. Tveito, *Computing the Electrical Activity in the Heart.*

For further information on this book, please have a look at our mathematics catalogue at the following URL: www.springer.com/series/7417

Texts in Computational Science and Engineering

1. H. P. Langtangen, *Computational Partial Differential Equations.* Numerical Methods and Diffpack Programming. 2nd Edition
2. A. Quarteroni, F. Saleri, P. Gervasio, *Scientific Computing with MATLAB and Octave.* 4th Edition
3. H. P. Langtangen, *Python Scripting for Computational Science*. 3rd Edition
4. H. Gardner, G. Manduchi, *Design Patterns for e-Science*.
5. M. Griebel, S. Knapek, G. Zumbusch, *Numerical Simulation in Molecular Dynamics*.
6. H. P. Langtangen, *A Primer on Scientific Programming with Python.* 4th Edition
7. A. Tveito, H. P. Langtangen, B. F. Nielsen, X. Cai, *Elements of Scientific Computing.*
8. B. Gustafsson, *Fundamentals of Scientific Computing.*
9. M. Bader, *Space-Filling Curves.*
10. M. Larson, F. Bengzon, *The Finite Element Method: Theory, Implementation and Applications.*
11. W. Gander, M. Gander, F. Kwok, *Scientific Computing: An Introduction using Maple and MATLAB.*

For further information on these books please have a look at our mathematics catalogue at the following URL: www.springer.com/series/5151